An expert guide to planning experimental tasks for evidence accumulation modelling

Russell J. Boag[1,*,†]; Reilly J. Innes[1,*]; Niek Stevenson[1]; Giwon Bahg[2]; Jerome R. Busemeyer[3]; Gregory E. Cox[4]; Chris Donkin[5]; Michael J. Frank[6]; Guy E. Hawkins[7]; Andrew Heathcote[1,7]; Craig Hedge[8]; Veronika Lerche[9]; Simon D. Lilburn[2]; Gordon D. Logan[2]; Dora Matzke[1]; Steven Miletić[1,10]; Adam F. Osth[11]; Thomas J. Palmeri[2]; Per B. Sederberg[12]; Henrik Singmann[13]; Philip L. Smith[11]; Tom Stafford[14]; Mark Steyvers[15]; Luke Strickland[16]; Jennifer S. Trueblood[3]; Konstantinos Tsetsos[17]; Brandon M. Turner[18]; Marius Usher[19]; Leendert van Maanen[20]; Don van Ravenzwaaij[21]; Joachim Vandekerckhove[15]; Andreas Voss[22]; Emily R. Weichart[23]; Gabriel Weindel[20]; Corey N. White[24]; Nathan J. Evans[5]; Scott D. Brown[7,*]; Birte U. Forstmann[1,*]

1. University of Amsterdam, The Netherlands
2. Vanderbilt University, USA
3. Indiana University, USA
4. SUNY University at Albany, USA
5. Ludwig Maximilian University of Munich, Germany
6. Brown University, USA
7. University of Newcastle, Australia
8. Aston University, UK
9. Kiel University, Germany
10. Leiden University, The Netherlands
11. University of Melbourne, Australia
12. University of Virginia, USA
13. University College London, UK
14. University of Sheffield, UK
15. University of California, Irvine, USA
16. Future of Work Institute, Curtin University, Australia
17. University of Bristol, UK
18. Ohio State University, USA
19. Tel-Aviv University, Israel
20. University of Utrecht, The Netherlands
21. University of Groningen, The Netherlands
22. Heidelberg University, Germany
23. Utah State University, USA
24. Missouri Western State University, USA

*These authors contributed equally (see author note).

†Corresponding author.

Author note

RJB and RJI contributed equally to the research and writing of the manuscript. SDB and BUF contributed equally to the supervision and administration of this project. We thank Joshua I. Gold, Michael D. Lee, and Roger Ratcliff for their helpful guidance on an earlier version of the manuscript. Correspondence concerning this article should be addressed to Russell J. Boag (r.j.boag@uva.nl), Nieuwe Achtergracht 129-B, 1018 WS, Amsterdam, University of Amsterdam, The Netherlands.

Funding information

RJB, RJI, NS, and BUF were supported by European Research Council Consolidator Grant (864750) and NWO Vici (016.Vici.185.052) awarded to BUF. NJE was supported by Australian Research Council Discovery Early Career Researcher Award (DE200101130). SDB was supported by Australian Research Council Discovery Project grants (DP210100313 and DP210103873).

Author contributions (CRediT statement)

RJB and RJI: Conceptualization; Writing - Original Draft; Writing - Review & Editing. **NS:** Conceptualization; Writing - Review & Editing. **NJE, SDB, and BUF:** Conceptualization; Writing - Review & Editing; Supervision; Project administration; Funding acquisition. All other co-authors (appearing in alphabetical order) contributed to Writing - Review & Editing and provided valuable feedback and critical discussion throughout the development of this manuscript.

Declaration of interests

The authors declare that they have no competing financial interests or personal relationships that could have influenced this work.

Abstract

Evidence accumulation models (EAMs) are powerful tools for making sense of human and animal decision-making behaviour. EAMs have generated significant theoretical advances in psychology, behavioural economics, and cognitive neuroscience, and are increasingly used as a measurement tool in clinical research and other applied settings. Obtaining valid and reliable inferences from EAMs depends on knowing how to establish a close match between model assumptions and features of the task/data to which the model is applied. However, this knowledge is rarely articulated in the EAM literature, leaving beginners to rely on the private advice of mentors and colleagues, and on inefficient trial-and-error learning. In this article, we provide practical guidance for designing tasks appropriate for EAMs, for relating experimental manipulations to EAM parameters, for planning appropriate sample sizes, and for preparing data and conducting an EAM analysis. Our advice is based on prior methodological studies and the authors' substantial collective experience with EAMs. By encouraging good task design practices, and warning of potential pitfalls, we hope to improve the quality and trustworthiness of future EAM research and applications.

Keywords: evidence accumulation models; experimental design; decision making; response time; model-based cognitive neuroscience

Introduction

Evidence accumulation models (EAMs) are powerful tools for understanding human and animal decision-making (Donkin & Brown, 2018; Evans & Wagenmakers, 2019; Gold & Shadlen, 2007; Smith & Ratcliff, 2024). They enable quantitative measurement of latent decision processes that are confounded in typical (e.g., linear model) analyses of response time (RT) and error rate (Lerche & Voss, 2020). EAMs explain key benchmark phenomena that arise in decision-making tasks (e.g., speed—accuracy trade-offs, asymmetries in the speed of correct and incorrect responses, and the characteristic positive-skew of RT distributions; Ratcliff & McKoon, 2008). Since their introduction in the 1960s and 1970s (Audley & Pike, 1965; Laming, 1968; Link & Heath, 1975; Stone, 1960; Vickers, 1970), EAMs have become one of the most successful theoretical frameworks in cognitive psychology (Evans & Wagenmakers, 2019; Ratcliff et al., 2016; Ratcliff & McKoon, 2008; Smith & Ratcliff, 2024) and cognitive neuroscience (Forstmann, Wagenmakers, et al., 2011; Forstmann et al., 2016; Gold & Shadlen, 2007; Mulder et al., 2014; Schall, 2019; Smith & Ratcliff, 2004). Further, they are increasingly being used to answer questions in domains such as behavioural economics (Busemeyer et al., 2019; Krajbich et al., 2014; Krajbich & Rangel, 2011), Human Factors/ergonomics (Boag et al., 2023) and in clinical/healthcare settings (Copeland et al., 2023; Ratcliff et al., 2022; White et al., 2010).

Obtaining valid inferences from EAMs relies on achieving a close match between model assumptions and features of the task and data to which the model is applied. Failing to achieve an appropriate task—model match can lead to misleading or spurious conclusions (e.g., Cassey et al., 2014; Ratcliff & Kang, 2021). However, the EAM literature lacks a comprehensive articulation of how to achieve a good task—model match. In this article, we provide practical guidance for designing tasks appropriate for EAMs, for relating experimental manipulations to EAM parameters, for sample size planning, for collecting and preparing data, and for conducting and reporting an EAM analysis. We point out problems that can arise if the models are used without sufficient regard for the factors that determine their validity. Sometimes there is no one-size-fits-all answer and finding an appropriate design may require careful judgment and consideration of trade-offs (e.g., collecting more trials versus maintaining participant engagement). To aid this process, we highlight the key issues and potential pitfalls affecting EAM analyses, so that readers can better plan

experiments for reliable EAM analysis. Our advice is grounded in prior methodological studies and in the authors' years of collective experience using EAMs to understand human and animal decision making.

By encouraging good task design practices, we hope to improve the quality and trustworthiness of future EAM research and applications. To make our advice as broadly applicable as possible, we do not focus on the details of specific EAMs. Instead, we focus on the common properties and design considerations shared by the most prominent basic EAM architectures (i.e., *relative evidence models*, e.g., Ratcliff, 1978; Wagenmakers et al., 2007, 2008; and *racing accumulator models*, e.g., Brown & Heathcote, 2008; Tillman et al., 2020; Usher & McClelland, 2001; see Figure 1). Our advice is intended for researchers and students who wish to apply an existing 'off-the-shelf' EAM to an experimental task in order to measure the cognitive processes driving decision-making behaviour. While our recommendations are intended for EAMs, many also apply more broadly to other cognitive modelling approaches (e.g., reinforcement learning, Wilson & Collins, 2019).

In the next section, we outline the general features and assumptions of EAMs. The remainder of the article is structured according to a typical EAM study workflow: We first consider whether an EAM is the appropriate tool for our research question. Next, we look at how to design EAM-appropriate experimental tasks, and strategies for collecting suitable data. We cover sample size planning and discuss best practices for experimental procedure, for assessing the quality of collected data, and for obtaining valid and reliable inferences from EAM analyses. We discuss interpreting and reporting the results of an EAM analysis and close with advice on what to do when the standard models fail.

The architecture of standard EAMs

EAMs assume that when presented with a stimulus (e.g., a left- or right-facing arrow), the decision maker samples evidence for the available actions or choice options (e.g., "Should I press the left or right arrow key?") until a threshold amount of evidence is reached. Many prominent models assume within-trial noise in this accumulation process (Ratcliff, 1978; Tillman et al., 2020; Usher & McClelland, 2001), although it is possible to capture key RT phenomena assuming only between-trial noise (Brown & Heathcote, 2008). Reaching a threshold immediately triggers the motor movement for the overt response (e.g., pressing

the left arrow key). Total RT is assumed to be the sum of three strictly sequential processing stages: 1) stimulus encoding, 2) decision making (evidence accumulation), and 3) motor response execution[1] (Bompas et al., 2023; Kelly et al., 2021; Servant et al., 2021; Weindel, Gajdos, et al., 2021). As we will see, this places constraints on the timing and structure of decision-making tasks appropriate for use with EAMs.

Figure 1 depicts the two prominent classes of EAM architectures. In *relative evidence models*, decisions are based on accumulating the difference in evidence between response options (e.g., Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998; van Ravenzwaaij et al., 2017; Wagenmakers et al., 2007, 2008). Relative evidence models have historically been limited to decisions involving two choice options (but see, Churchland et al., 2008; Ditterich, 2010; Kvam, 2019; Niwa & Ditterich, 2008; Smith et al., 2020). By contrast, in *racing accumulator models*, decisions are based on accumulating the absolute evidence for response options in separate modular accumulators (e.g., Bogacz et al., 2007; Brown & Heathcote, 2008; Heathcote & Love, 2012; Kirkpatrick et al., 2021; Rouder et al., 2015; Teodorescu & Usher, 2013; Tillman et al., 2020; Tsetsos et al., 2011; Usher et al., 2002; Usher & McClelland, 2001). Racing accumulator models can accommodate any number of choice options, typically with an accumulator per choice. Although relative and absolute evidence models differ regarding how they conceptualize evidence, they have similar requirements for achieving a good task—model match and often arrive at the same substantive conclusions (Donkin, Brown, Heathcote, et al., 2011). In both architectures, decision making is governed by the same 3 or 4 parameters, which are interpreted similarly across models (Voss et al., 2004). Moreover, both architectures have similar data quality requirements and often give convergent results when applied to the same data (Donkin, Brown, Heathcote, et al., 2011; Dutilh et al., 2019).

[1] In most EAMs, the time taken for stimulus encoding and motor responding are not separately identifiable. Instead, only their sum (referred to as "nondecision time") is estimated (i.e., total RT = decision time + nondecision time).

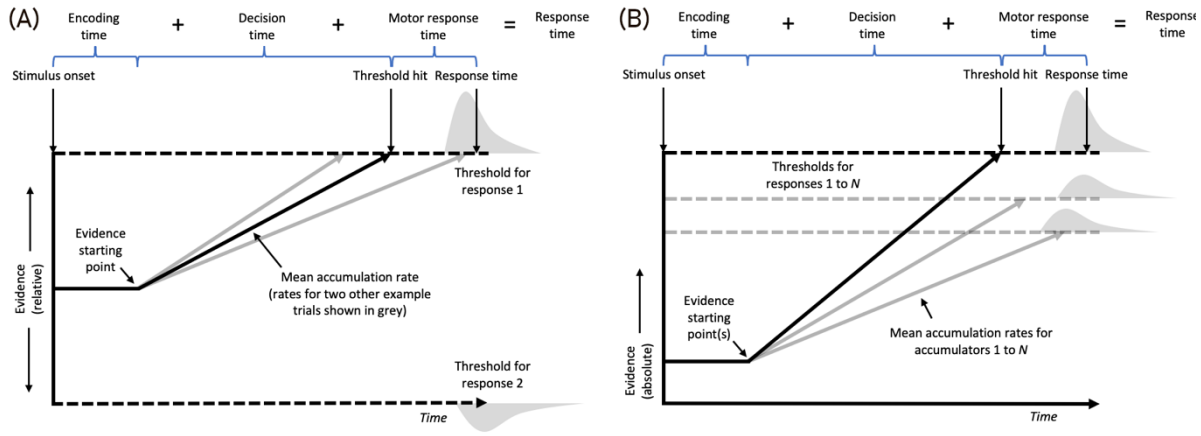

Figure 1. Illustration of two standard EAM architectures. In relative evidence models (Panel A), decisions are based on accumulating the difference in evidence between response options. The first threshold to be reached determines the overt response and RT. In racing accumulator models (Panel B), decisions are based on accumulating the absolute evidence for *N* response options in separate modular accumulators. In these models, the first accumulator to reach threshold determines the overt response and RT. In both architectures, RT is the sum of decision time plus the time taken for nondecision processes such as sensory encoding and production of the motor response. Both architectures share common processing assumptions and interpretation of core parameters (see text for details). Note that only the noiseless mean accumulation rate is depicted. For models with within-trial noise, each accumulation process traces a noisy trajectory around this mean rate (e.g., Ratcliff, 1978).

A comprehensive overview of key model parameters and their uses is given in the section 'Mapping EAM Parameters to Experimental Manipulations'. However, briefly, the models contain parameters controlling the evidence starting point (allowing for a priori biases), accumulation rate (controlling the speed of processing), threshold/boundary separation (controlling the amount of evidence required to make a response), and nondecision time (the sum of time taken for stimulus encoding and motor response production). The basic frameworks also allow across-trial variability in accumulation rate, starting point, and nondecision time, which account for commonly observed differences in the speed of correct and incorrect responses (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002).

As will be discussed (see section 'Going Beyond the Standard Models'), the basic architecture has been extended to include additional mechanisms (e.g., Fific et al., 2010; McDougle & Collins, 2021; Miletić et al., 2021; Nosofsky & Palmeri, 1997, 2015; Pedersen et al., 2017) and to account for tasks/situations that violate various processing assumptions of

the standard models (e.g., Diederich, 2024; Diederich & Trueblood, 2018; Hawkins et al., 2015; Holmes et al., 2016; Holmes & Trueblood, 2018; Lee & Sewell, 2024; Little et al., 2018; Smith & Ratcliff, 2022; Ulrich et al., 2015; Voss et al., 2019; White et al., 2011; Zhang et al., 2014; for a review, Evans & Wagenmakers, 2019). Most of the advice in this article will apply when working with these models. However, researchers should be aware that extended models may operate on a different set of processing assumptions and thus have idiosyncratic (mechanism-specific) design constraints.

Processing assumptions of standard EAMs

Here, we outline the core assumptions of the basic EAM framework that have implications for the design of tasks suitable for EAMs (summarised in Table 1). For data from an experimental task to be suitable, the task must satisfy the assumptions of the model. The core structural assumption of the models is that *each decision is the result of a single, continuous (uninterrupted) evidence accumulation process and culminates in a single discrete response*. In short, the models apply to tasks in which one decision is followed by one response (Brown & Heathcote, 2008; Busemeyer & Townsend, 1993; Ratcliff, 1978; Usher & McClelland, 2001).

During a trial/decision, the models assume *within-trial stationarity*, which refers to the assumption that model parameters (e.g., accumulation rates and thresholds) do not change in value while a decision is in progress (Ratcliff, 1978). For accumulation rates, this means that evidence accumulates at a constant average rate[2] (although potentially with substantial noise) for the duration of the trial (i.e., from stimulus onset to response onset) (Brown & Heathcote, 2008; Ratcliff, 1978; for alternatives, Stine et al., 2020). In practice, this means that stimuli should provide a constant input to the evidence accumulation process (i.e., stimulus evidence should not change in strength or sign over the course of a trial, Lee & Sewell, 2024; Smith & Lilburn, 2020). For thresholds, within-trial stationarity means that thresholds are set prior to stimulus onset and do not change in value during a trial. This means that individuals are assumed to keep the same cognitive control/speed—accuracy

[2] Some models, most prominently, the leaky competing accumulator model (Usher & McClelland, 2001), relax this assumption in that the dominant response accumulator may provide increasingly strong inhibitory input to its competitors over time, which would reduce the mean accumulation rate for competing responses throughout a trial.

trade-off settings throughout a decision and to not increase or decrease in caution during a trial (for models that allow dynamic thresholds, Hawkins et al., 2015; Smith & Ratcliff, 2022; Voskuilen et al., 2016).

Across trials, the standard application of EAMs assumes *within-condition stationarity*, which refers to the assumption that model parameters do not change in value across trials (of the same type) within a condition. This assumption is important for model fitting, which relies on pooling information across trials of the same type. Theoretically, the assumption is that trials of the same type are independent measurements of the same underlying process (generated from the same cognitive settings). Empirically, the expectation is that participant performance is stable for the duration of the experiment[3] (e.g., RT distributions do not change in shape or scale over time).

The reviewed EAM assumptions have implications for the (choice-RT) data to which they are applied. For one, EAMs can only predict *positively skewed RT distributions*. This owes to the geometry of the models, whereby equal differences in accumulation rate are projected as unequal differences in decision time (see Figure 1) (Ratcliff & McKoon, 2008). In practice, this means that the models can only fit empirical RT distributions with characteristic positive skewness, and fail to fit RT distributions that are normal or negatively skewed in shape (Evans, Hawkins, et al., 2020). Ignoring skew can lead to biases in parameter estimation (Verdonck & Tuerlinckx, 2016). The section 'Planning Tasks that Meet EAM Assumptions' contains advice on ensuring data satisfy this assumption.

Finally, EAMs assume the data are *free of contaminant processes*. That is, data come from an evidence accumulation process and not some other process such as random guessing or nonresponding (Ratcliff, 1993; Ratcliff & Tuerlinckx, 2002). Strategies for identifying and accounting for contaminants are discussed throughout the article.

With this background in place, the remainder of this article steps through the components of a typical EAM study workflow, giving advice on how to plan and conduct a robust study. In doing so, we regularly refer back to the model assumptions outlined in this section.

[3] However, when there are a sufficient number of trials in an experiment, blocks or sessions of trials may be treated as another condition, thus allowing for estimation of block-to-block changes in model parameters (e.g., Dutilh et al., 2009).

Table 1. Standard EAM assumptions and implications for task design.

EAM assumption	**Explanation**	**Design implications**
Decisions well described by a single, continuous accumulation process resulting in a discrete response	The outcome of each decision (trial) is a discrete response resulting from an uninterrupted evidence accumulation process running from stimulus to response onset (i.e., one decision ↦ one response).	Trials should have a clear stimulus onset. The response modality should allow precise measurement of response onset.
Within-trial stationarity	Model parameters do not change during a decision (trial). Stimulus evidence should not change (e.g., ramp up or change sign) during a trial. Thresholds do not change dynamically within a trial or in response to information unknown before stimulus onset.	Use static stimuli that provide a constant evidentiary input from stimulus onset to the response. Use sufficiently long intertrial intervals to avoid interference from processes that ran on previous trials (e.g., process overlap and proactive interference).
Within-condition stationarity	Model parameters do not change across trials of the same type. Trials of the same type should be independent observations generated by the same latent cognitive settings. Necessary for pooling observations for model fitting.	Minimize learning effects that are not modelled. Minimize fluctuations of attention and potential changes in strategy.
Positively skewed RT distributions	RT distributions for each response should be positively skewed and free from truncation in the tails.	Use a well-calibrated response window (calibrated to the mean RT and variance of a typical participant performing the target task).
Data free of contaminant processes	Data come from an evidence accumulation process (and not some other process such as fast guessing). Participants perform the task as instructed.	Provide clear task instructions. Monitor participant behaviour. Display corrective feedback following undesirable responses (e.g., “Too fast!”). Allow participants sufficient breaks.

Planning research questions for EAM analysis

Before the task design and modelling process can begin, the researcher must first decide whether an EAM analysis is the appropriate tool to answer the research question. While EAMs have many uses (Crüwell, Stefan, et al., 2019), our present focus is on using EAMs as a cognitive measurement model (Donkin & Brown, 2018; Lee et al., 2019; see also, Batchelder, 2010, 2016; Batchelder & Riefer, 1999; Smith & Batchelder, 2010). Measurement studies typically focus on interpreting the parameters of an existing 'off-the-shelf' EAM that is taken a priori to adequately characterise the processes individuals use to perform the target task (e.g., Huang-Pollock et al., 2017; Janczyk & Lerche, 2019; Klauer et al., 2007; Ratcliff et al., 2004; Ratcliff & Rouder, 2000). To understand what kinds of research questions are suitable for EAM analysis, it is helpful to consider the output of an EAM that has been fit to participant data. For each participant, the model provides parameters that represent measurements of that individual's latent cognitive settings (e.g., accumulation rate, threshold, bias, and nondecision time). Additional population-level parameters characterising group differences can be obtained using hierarchical modelling approaches (e.g., Chávez De la Peña & Vandekerckhove, 2023; Gunawan et al., 2020; Heathcote et al., 2019; Stevenson, Innes, et al., 2024; Wiecki et al., 2013). Changes in cognitive processes are quantified by changes in the values of this set of model parameters. Therefore, suitable research questions involve assessing how model parameters differ within or between groups (e.g., Ratcliff et al., 2003; Steyvers et al., 2019), individuals (e.g., Evans et al., 2018), or experimental conditions/treatments (e.g., Heathcote et al., 2015; Ratcliff et al., 2003; Strickland et al., 2023), as well as assessing how parameters relate to other individual-level covariates (e.g., eye-tracking, Cavanagh et al., 2014; Fiedler & Glöckner, 2012; Krajbich & Rangel, 2011; and neurophysiological measures, such as EEG, MEG, and fMRI, Forstmann et al., 2011; Harris & Hutcherson, 2022; Nunez et al., 2023, 2024; Turner et al., 2013; Turner, Forstmann, et al., 2019; Turner, Palestro, et al., 2019). EAMs allow multiple data sources to be analysed under a common model and results interpreted in terms of well-supported cognitive theory (Forstmann, Wagenmakers, et al., 2011).

For an EAM analysis to be useful, questions must map to the cognitive processes represented by EAM parameters (i.e., accumulation rate, threshold, bias, and nondecision time). Questions are typically posed in a similar manner to traditional confirmatory

experimental research, where the goal is to understand the effect of particular experimental manipulations, treatments/interventions, or clinical disorders on some measured outcome variable (Donkin & Brown, 2018). For example, in a series of studies, Ratcliff and collaborators asked whether age-related slowing is due to slower evidence accumulation (cognitive impairment hypothesis), higher thresholds (conservative responding hypothesis), or longer nondecision time (physical slowing hypothesis) (Ratcliff et al., 2003, 2006; Ratcliff, Thapar, Gomez, et al., 2004; Ratcliff, Thapar, & McKoon, 2004; Thapar et al., 2003). This question presents a clear test of three competing hypotheses that can be instantiated in EAMs and evaluated. To give an example involving a subject-level covariate, Forstmann et al. (2008) asked whether cue-induced threshold adjustments (a measure of top-down cognitive control) are correlated with fMRI BOLD signal in the striatum and pre-supplementary motor area (two structures hypothesized to be involved in such adaptive control). This question, posed in terms of individual-differences correlations, presents a clear test of the relationship between the model-based measure (magnitude of threshold adjustment) and the hypothesized neural covariates (striatal and pre-SMA BOLD signal). Operationalising questions in this way is necessary to develop clear, testable hypotheses. That is, hypotheses that can be instantiated in an EAM and subjected to model comparison and evaluation. We explore this topic further in the section 'Mapping Experimental Manipulations to EAM Parameters'.

Unsuitable questions for standard EAMs are those that involve violations of their assumptions. For example, asking questions about how parameters change from trial-to-trial (violating within-condition stationarity) require extended models/methods that allow trial-wise parameter estimation (Boehm et al., 2014; Ho et al., 2012; Van Maanen et al., 2011). Formulating good research questions requires a sound understanding of theory of both EAMs and the target domain. The EAM literature, especially measurement studies where the focus is on interpreting parameter effects (e.g., Boag et al., 2023; Evans et al., 2018; Huang-Pollock et al., 2017; Ratcliff & Rouder, 2000; Weigard et al., 2018) can be a rich source of ideas and help build intuition for developing suitable research questions. Getting the research question right is important because it ultimately dictates many experimental design and analysis choices (e.g., sample size planning and whether to use hierarchical or independent-subjects approaches).

Planning tasks that meet EAM assumptions

Having formulated a research question, focus turns to designing an experimental task that will be informative for the research question and that meets the processing assumptions of EAMs. In this section, we discuss EAM-specific constraints on task design, relating each back to the relevant EAM assumptions. Our advice is intended to assist researchers in designing tasks that satisfy the assumptions of the basic EAM framework but allows for judicious deviations such as when the focus is on developing a new model (Crüwell, Stefan, et al., 2019).

One decision, one response. As noted earlier, EAMs assume decisions involve a single, uninterrupted evidence accumulation stage, culminating in a discrete response. Evidence is assumed to accumulate continuously from stimulus onset to the response. EAM-appropriate tasks need clearly defined stimulus and response onsets that do not overlap with processes outside of the response window. Stimulus evidence should be static (e.g., of fixed strength within a trial) and presented for the entire duration of the response window (from stimulus onset to response initiation). This will ensure that stimuli provide a constant input to the evidence accumulation process until a response is initiated, as is assumed in the standard models.

Further, each decision should culminate in a single, discrete response, chosen from a set of two or more choice options. This is because, in standard EAMs, evidence always terminates at a single, discrete response threshold. Consequently, tasks that involve open-ended response options (e.g., free recall tasks) or the possibility of submitting more than one response during a single trial (e.g., change of mind tasks, Stone et al., 2022; double response paradigms, Evans et al., 2020) require extensions beyond standard EAMs.

Within-trial stationarity. EAMs assume that the parameter settings of the model do not change part way through a decision. Specifically, EAMs assume that threshold and bias settings are unaltered in response to the stimulus, and most assume that evidence accumulates at a constant average rate from stimulus onset to response onset. When designing an experiment, any information intended to affect threshold or bias settings must be presented before the onset of the stimulus. Likewise, any information not intended to affect decision-making and cognitive control settings should be kept outside of the response window. With regard to experimental design, this means that the evidence input to the

decision process should not change during a trial, meaning that decision-relevant stimulus features should be constant throughout a trial (Smith & Lilburn, 2020). For example, stimuli in a perceptual decision-making task should not change in brightness or contrast part-way through a trial, since this would require a corresponding change in accumulation rate. Tasks involving dynamic evidence can be modelled using (computationally expensive) extensions to the basic EAMs (e.g., Diederich, 2024; Diederich & Trueblood, 2018; Holmes et al., 2016; Holmes & Trueblood, 2018).

Within-condition stationarity. EAMs also assume stationarity across trials of the same type within a condition. This is because model fitting requires trials of the same type to be treated as independent observations of the same latent cognitive settings. Aside from non-systematic trial-to-trial variation accounted for in the model's across-trial variability parameters, there should be no systematic changes in threshold or mean accumulation rate across trials of the same type. This assumption is important for statistical power and measurement precision, which relies on information pooled across many observations (trials) (Smith & Little, 2018). When designing experiments, researchers should attempt to minimize factors that could cause parameters to change systematically across trials. For example, accumulation rates are known to increase with learning, initially rising steeply before tapering off to a stable asymptotic level (e.g., Fontanesi et al., 2019; Miletić et al., 2021; Pedersen et al., 2017; Sewell et al., 2019). Rates can also decrease with fatigue or inattention/task disengagement (Huang-Pollock et al., 2020; Ratcliff & Van Dongen, 2011; Walsh et al., 2017). Thresholds may also decrease over the course of an experiment due to participants becoming impatient and trading accuracy for speed in an effort to complete the experiment sooner (Hawkins et al., 2012; Larson & Hawkins, 2023).

Trial-to-trial variability is unavoidable (Aschenbrenner et al., 2018; Rouder et al., 2023) due to noise at many levels, including the noise inherent in neural systems (Faisal et al., 2008; Smith, 2010, 2023) and dynamic fluctuations in cognitive and affective state (Schurr et al., 2024). Standard EAMs account for such noise sources via their across-trial variability parameters. Nevertheless, researchers should take reasonable measures to ensure such variability is kept as non-systematic as possible.

Stimuli. Stimuli provide the critical input to the decision-making process. Stimuli supply the evidence upon which decisions are based and largely determine the cognitive domain engaged by a task. For example, in a psychophysics task, evidence might be based on the

objective luminance values of stimuli (e.g., Sewell & Smith, 2012; van Ravenzwaaij et al., 2020). By contrast, evidence in a preferential choice task could be subjective value elicited by viewing images of food items (e.g., Huseynov & Palma, 2021; Milosavljevic et al., 2010). In working memory and categorization tasks, evidence may derive from the strength with which items are activated in memory (Ratcliff, 1978; Shadlen & Shohamy, 2016) or the strength of learned associations between stimuli and expected response outcomes (Dutilh et al., 2009; Dutilh, Krypotos, et al., 2011; Miletić et al., 2021; Sewell et al., 2019). As noted, the evidence supplied by stimuli should be fixed within a trial (i.e., unchanging in strength for the duration of the trial) to provide a constant input to the decision process.

Across trials or blocks, stimuli are often the target of manipulations designed to affect the signal-to-noise ratio of the evidence entering the decision process (e.g., discriminability, difficulty, etc.). When designing experiments, it is important to calibrate stimuli to be of an appropriate difficulty level. This is because EAMs can struggle to fit floor effects (chance-level accuracy) and ceiling effects (e.g., near-perfect accuracy with too few errors) (Dutilh, Wagenmakers, et al., 2011). Floor effects occur when a task is too difficult, and usually mean that participants cannot discriminate between choice options. Consequently, participants may be using a guessing strategy rather than sampling evidence as assumed in EAMs. By contrast, ceiling effects occur when a task is too easy, causing very few incorrect responses to be observed. As we discuss in the section on sample size planning, it is important to elicit enough error observations for reliable model estimation (Lüken et al., 2023). We recommend calibrating stimuli to produce error rates of 5-35% (Dutilh, Wagenmakers, et al., 2011; Lüken et al., 2023; Ratcliff & Childers, 2015). Calibration can be achieved through pilot testing or via more advanced methods that perform individualised calibration based on task performance (e.g., Myung et al., 2009; Myung et al., 2013; Yang et al., 2021).

Response modality. Standard EAMs assume that the onset of the response coincides with termination of the evidence accumulation process (Figure 2). That is, the decision and motor response processes occur sequentially (i.e., a motor response is only initiated once a decision has been reached). As such, we recommend using response modalities with a sharp, clearly defined response onset and short execution times, such as manual key presses (*Mean* = 160 ms [120—230 ms]) or saccades (*Mean* = 60 ms [30—100 ms]) (Bompas et al., 2023). Other response modalities such as computer mouse or foot pedal are also possible (e.g., Leontyev & Yamauchi, 2021; Michmizos & Krebs, 2014). However, responses using such

modalities may produce relatively variable response onsets (and consequently less precise estimates of nondecision time).

The most critical consideration is that the chosen modality should enable the precise measurement of RT. For most purposes, a standard computer keyboard provides sufficiently precise RT measurements (up to the limit of the internal refresh rate). However, highly precise (i.e., to the millisecond) timing can be obtained with specialized computer systems and the use of precision-timing software/apparatus (Bridges et al., 2020; Plant et al., 2002). We recommend using left—right symmetric key arrangements [with the layout counterbalanced across participants, if appropriate to do so; see section 'Stimulus—Response (Decision Outcome) Mapping']. This allows for the highly desirable assumption of a common nondecision time across responses (and potentially across subjects).

Mapping experimental manipulations to EAM parameters

It is important to establish clear theoretical links between experimental manipulations (e.g., speed vs. accuracy instructions, task difficulty, or working memory load) and their expected effects on EAM parameters and data. Understanding the behavioural signatures of experimental manipulations can give confidence that a manipulation is working as intended. Becoming familiar with EAM theory and reading published EAM studies can help build intuition for which model parameters are likely to be affected by a given manipulation. Much of the key theoretical EAM literature and a variety of application studies are cited in this article.

Not all EAM parameters will be relevant to every analysis. For example, a researcher studying consumer choice preferences (e.g., preference for one product over another) may be uninterested in nondecision time but be highly interested in using accumulation rates to measure preference strength and starting point (or thresholds) to measure choice biases (Busemeyer & Townsend, 1993; Cerracchio et al., 2023; Krajbich et al., 2012, 2015). Additionally, it is common practice to not estimate variability parameters (e.g., by fixing them to zero) unless they are needed to account for certain data features (e.g., fast guesses) (Lerche & Voss, 2016; Ratcliff & Rouder, 1998).

Below, we briefly review common manipulations that have been used to selectively influence each standard EAM parameter (see Box 1). The primary uses of each model

parameter, common mappings to experimental manipulations, and expected effects on behaviour are summarized in Table 2.

Stimulus—response (decision outcome) mapping. Some tasks will have stimulus—response mappings that naturally correspond to objectively correct or incorrect decision outcomes (e.g., pressing the left arrow key in response to a predominantly left-moving stimulus). However, standard EAMs can easily accommodate tasks with subjective or probabilistic stimulus—response mappings (e.g., preferential choice tasks, probabilistic categorization tasks, and tasks with probabilistic rewards/payoffs; Lee & Usher, 2023; Milosavljevic et al., 2010; Sewell & Stallman, 2020). In relative evidence models (e.g., Ratcliff, 1978; Wagenmakers et al., 2007), which are limited to two-choice tasks, each threshold is mapped to one of the possible response options, and a single accumulation rate measures the difference in evidence between options. However, in race models (e.g., Brown & Heathcote, 2008; Tillman et al., 2020), which can accommodate an arbitrary number of response options, each latent response is assigned an accumulator with its own threshold and an accumulation rate representing the *absolute* evidence for that response. Race models can also instantiate more complex decision rules (e.g., AND and OR rules) used for combining multiple stimulus attributes into a final decision (e.g., Fific et al., 2010; Little et al., 2018; van Ravenzwaaij et al., 2020).

Accumulation rate. Accumulation rates measure the strength (signal-to-noise ratio) of evidence extracted from the stimulus (e.g., salience, preference strength, or discriminability relative to other choice options; Gold & Shadlen, 2007; Palmer et al., 2005; Ratcliff & McKoon, 2008). Rates are sensitive to the processing abilities of the decision maker (Schmiedek et al., 2007) and the amount of attention or cognitive resources deployed to the task (i.e., the degree to which the participant is paying attention; Boag et al., 2023; Castro et al., 2019; Eidels et al., 2010). Holding one constant allows measurement of the other (e.g., for equivalent stimuli, different rates reflect differences in attention/capacity).

In a typical experiment, rates are used to account for manipulations of evidence strength (e.g., low- versus high-discriminability stimuli), attention or processing capacity, and task difficulty. That is, manipulations affecting how easily stimuli are perceived and/or processed (Mulder et al., 2014; Palmer et al., 2005; Ratcliff & McKoon, 2008; Smith et al., 2015; Smith & Sewell, 2013). This is accomplished by estimating a different accumulation rate for each difficulty level (Ratcliff & Rouder, 1998). Behaviourally, a faster accumulation rate predicts

faster responses and fewer errors, while a slower rate predicts the converse (Ratcliff & McKoon, 2008). Accumulation is typically faster for easier decisions (Ratcliff & Rouder, 1998) and faster for responses associated with higher reward or subjective value (Busemeyer & Townsend, 1993; Krajbich et al., 2012, 2015). Rates track the strength of associative relationships learned via feedback (e.g., Fontanesi et al., 2019; Miletić et al., 2021; Pedersen et al., 2017; Sewell et al., 2019) and the activation strength of items retrieved from memory (Ratcliff, 1978; Ratcliff & McKoon, 1988). Further, these mappings hold in more complex naturalistic tasks (for a review, Boag et al., 2023).

Threshold. Thresholds are a locus of proactive cognitive control (Strickland et al., 2018). Thresholds control the amount of evidence needed to trigger a response and thus measure response caution or speed—accuracy settings. As noted earlier, EAMs assume thresholds are set in advance of stimulus onset (i.e., not adjusted based on features of the current stimulus, since it would be circular for the threshold used to identify a stimulus to depend on knowing the identity of that stimulus). In other words, thresholds cannot be altered based on information that was unknown before the trial began (Donkin, Averell, et al., 2009). Consequently, manipulations intended to affect threshold or bias settings must be presented before the onset of a trial/stimulus. This is typically achieved using pre-trial cues or blocked instructions (e.g., Forstmann et al., 2008; Katsimpokis et al., 2020), the aim of which is to allow participants to make strategic adjustments (e.g., adopt different threshold/bias settings) before encountering the upcoming stimulus.

In a typical experiment, thresholds are used to explain speed—accuracy trade-off effects whereby individuals set lower thresholds when less time is available, and higher thresholds when more time is available (Bogacz et al., 2010; Evans, Hawkins, et al., 2020; Forstmann et al., 2008; Frazier & Yu, 2007; Heitz & Schall, 2012; Katsimpokis et al., 2020; Rae et al., 2014; Ratcliff & McKoon, 2008). Behaviourally, higher thresholds predict slower, more accurate decisions, while lower thresholds predict faster, less accurate decisions (Ratcliff & Rouder, 1998). In practice, thresholds are sensitive to task importance and response bias manipulations, in which participants set lower thresholds for prioritized/more rewarding/higher frequency responses and higher thresholds for non-prioritized/less rewarding/lower frequency responses (Boag et al., 2019; Mulder et al., 2012; Strickland et al., 2018; Trueblood et al., 2021; for a review, Cerracchio et al., 2023). Thresholds are

further implicated in post-error slowing (Damaso, Castro, et al., 2022; Damaso, Williams, et al., 2022), a kind of trial-to-trial speed—accuracy trade-off (Larson & Hawkins, 2023).

Starting point. As noted in the preceding paragraph, racing accumulator models measure biases for one response over another by allowing competing response options to have different thresholds. By contrast, relative evidence models measure response biases by assessing how the starting point of the evidence accumulation process deviates from the neutral midpoint between the two response boundaries (Leite & Ratcliff, 2011; Ratcliff & McKoon, 2008; see also, Edwards, 1965). Like thresholds, the evidence starting point is assumed to be under the control of the decision maker and manipulations intended to affect starting point must be presented before stimulus onset. Behaviourally, deviating from the neutral midpoint makes responses for the favoured (closer) threshold faster and more accurate while making responses for the non-favoured (further) threshold slower and less accurate (Ratcliff & McKoon, 2008; for a review, Cerracchio et al., 2023). In experiments, starting point biases have been used to measure biases in police officer's decisions to shoot lighter- versus darker-skinned suspects (Johnson et al., 2018, 2021; Pleskac et al., 2018) and to quantify individuals' tendency to identify items as weapons versus non-weapons (Todd et al., 2021). Starting point has also been used to understand how various response biases are affected by factors such as heightened time pressure (Chen & Krajbich, 2018), changes in stimulus prevalence (Trueblood et al., 2021; see also, Leite & Ratcliff, 2011), and payoff structure (Leite & Ratcliff, 2011).

Nondecision time. Nondecision time measures the sum of the time taken to encode the stimulus (at stimulus onset) and time to produce the motor response (at response onset) (Bompas et al., 2023). Nondecision time is sensitive to the difficulty of both the encoding and motor responding stages. For example, it is sensitive to changes in low-level visual features of stimuli and the complexity or force required to produce the motor response (Bompas et al., 2023; Gomez et al., 2015; Ho et al., 2009; Sandry & Ricker, 2022; Servant et al., 2016; Voss et al., 2004; Weindel, Gajdos, et al., 2021). Although encoding and motor RT cannot be separately identified in standard EAMs, they may be disentangled experimentally (e.g., by holding stimulus properties constant while manipulating response modality or vice-versa). Empirically, nondecision time shifts RT distributions in time without affecting accuracy or the shape or scale of the distribution (Ratcliff & McKoon, 2008).

In experimental settings, nondecision time has been used to measure potential differences in encoding or motor response production (Ratcliff, Thapar, Gomez, et al., 2004; Van Maanen et al., 2016). For example, Ratcliff et al. (2004) found that older participants produced reliably slower nondecision times than did younger participants (see also, Van Maanen et al., 2016). Saccadic eye movements have been found to elicit reliably shorter nondecision times than manual key-press responses (Bompas et al., 2023; Ho et al., 2009). Nondecision time has also been found to be shorter under conditions of greater urgency (e.g., Rae et al., 2014; Ratcliff, 2006), potentially reflecting a tendency to encode stimuli less deeply when under time pressure (e.g., Palada et al., 2018, 2019). However, we caution that nondecision time is sometimes estimated less reliably than other EAM parameters (Lerche & Voss, 2018), and can be highly variable across individuals, conditions, and tasks (Bompas et al., 2023). Refining EAMs account of nondecision time is a topic of ongoing model development work (Bompas et al., 2023; Kelly et al., 2021; Servant et al., 2021).

Variability parameters. The across-trial variability parameters (i.e., in accumulation rate, starting point, and nondecision time) are less frequently used for measurement or inference. Rather, they allow the model to account for a number of commonly observed features of behavioural data, such as crossovers in the speed of correct and incorrect responses (Ratcliff, 2013; Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004).

Across-trial variability in accumulation rate can account for slow errors (Ratcliff, 1978). This is because trials with faster-than-average accumulation produce fast responses with very few errors. By contrast, trials with slower-than-average accumulation produce slow, error-prone responses, which together results in disproportionately many slow errors (Lerche & Voss, 2016). In experiments, across-trial rate variability can be used to account for manipulations affecting variability in evidence extracted from the stimulus (Starns, 2014; Yap et al., 2012), and to identify factors that lead to increased uncertainty (greater variability) in decision making (Palada et al., 2020; Starns, 2014).

Across-trial variability in starting point can account for fast errors (Laming, 1968). This is because when the accumulation process starts closer to the threshold for the incorrect latent response, errors become both faster and more frequent. By contrast, when accumulation starts closer to the threshold for the correct latent response, errors become slower and less frequent, resulting in disproportionately many fast errors (Lerche & Voss, 2016). Including starting point variability alongside rate variability allows the model to

account for interactions (crossovers or reversals) between correct and incorrect RT (e.g., fast errors in some cells and slow errors in others; Ratcliff et al., 1999; Ratcliff & Rouder, 1998; Wagenmakers, Ratcliff, et al., 2008). Starting point variability may be used to account for factors affecting uncertainty (variability) in prior beliefs or expectations (Mulder et al., 2012).

Across-trial variability in nondecision time can account for changes in the leading edge (e.g., the 0.1 quantile) of RT distributions (e.g., Ratcliff et al., 2004; Ratcliff & Tuerlinckx, 2002), including those caused by contaminant processes such as fast guesses (Ratcliff & Tuerlinckx, 2002). This is because nondecision time variability fattens the tails (i.e., decreases skew) of RT distributions (Lerche & Voss, 2016), making the model more robust to fast contaminants. Models with nondecision time variability predict a shallower onset of responding than models without. Empirically, nondecision time variability accounts for variability in encoding and motor response production (Bompas et al., 2023).

We reiterate that across-trial variability parameters tend to be estimated less reliably than other parameters (Boehm et al., 2018; Vandekerckhove & Tuerlinckx, 2007; van Ravenzwaaij & Oberauer, 2009; Lerche & Voss, 2016; Lerche, Voss, & Nagler, 2017; Yap, Balota, et al., 2012). Moreover, at least one rate variability parameter is typically held fixed in at least one design cell in order to satisfy the scaling property of EAMs (Donkin, Brown, et al., 2009). In racing accumulator models, a common choice is to set across-trial rate variability to 0.1 or 1. Although some work suggests that differences in across-trial variability in accumulation rate and/or nondecision time can be recovered reasonably reliably in some cases (e.g., Boehm et al., 2018; Starns & Ratcliff, 2014), there is evidence suggesting variability parameters trade-off with other model parameters and can exhibit non-stationarity over the course of an experiment (e.g., Dutilh et al., 2011; Evans et al., 2018; Evans & Hawkins, 2019). Estimation and reliability issues with variability parameters can be improved by fixing parameters (e.g., by constraining variability parameters to a single estimated value or removing them entirely by setting variability to zero, Boehm et al., 2018; Lerche & Voss, 2016; van Ravenzwaaij et al., 2017). Moreover, some EAM software simply does not allow for the estimation of across-trial variability (e.g., EZ-diffusion, Dutilh et al., 2013; Grasman et al., 2009; Schmiedek et al., 2007; Souza & Frischkorn, 2023; van Ravenzwaaij et al., 2012, 2017; Wagenmakers et al., 2007, 2008), or requires variability to be fixed across participants (e.g., HDDM, Wiecki et al., 2013). Overall, we recommend

exercising caution if answering the research question relies on inferences based on potentially unreliable variability parameters.

In the next section, we outline the elements of a single trial in a typical EAM experiment and considerations for task design.

Table 2. Mapping experimental manipulations to EAM parameters.

Parameter	**Common manipulations**	**Data effect**
Accumulation rate	Stimulus discriminability; subjective task difficulty; strength of preference; strength of memory trace; attention; effort	Increasing accumulation rate produces faster, more accurate decisions and reduces RT variability.
Threshold	Speed—accuracy trade-off; instructions; cognitive control; urgency; choice biases (in racing accumulator models)	Increasing threshold/boundary separation produces slower, more accurate decisions, and increases RT variability.
Starting point	Response biases; stimulus prevalence/base rate; reward/payoff structure; prior knowledge and expectations	Starting closer to a boundary makes that response occur more quickly and frequently than the non-favoured response.
Nondecision time	Accounts for complexity of encoding and the complexity or difficulty of producing the motor response.	Shifts RT distributions by a constant amount without affecting accuracy or the shape and skewness of the distribution.
Rate variability	Accounts for decision uncertainty/evidence variability and slower-than-average errors.	Greater across-trial rate variability increases the proportion of slow errors.
Starting point variability	Accounts for variability in prior beliefs or expectations, and faster-than-average errors.	Greater starting point variability increases the proportion of fast errors.
Nondecision time variability	Account for variability in motor responding and RT distributions with reduced skewness (e.g., a shallower onset of responding due to fast contaminants).	Greater nondecision time variability ‘smears’ the RT distribution along the time axis, creating fatter tails (i.e., greater probability of both faster and slower responses) and shallower onset of responding.

Box 1. Selective influence.

Selective influence refers to the idea that an experimental manipulation should directly and selectively engage the target cognitive process. That is, a manipulation should affect only the EAM parameter it is theoretically expected to affect, and it should not affect other parameters (Jones & Dzhafarov, 2014). Selective influence was neatly demonstrated by Ratcliff and Rouder (1998), who orthogonally manipulated decision difficulty and speed/accuracy instructions. Decision difficulty was found to selectively influence diffusion model accumulation rates, whereas speed/accuracy instructions selectively influenced thresholds (see also, Forstmann et al., 2011; Hawkins et al., 2012; Starns & Ratcliff, 2010; Usher & McClelland, 2001; Wagenmakers et al., 2008; but see, Katsimpokis et al., 2020). Subsequent work demonstrated selective influence for other parameters. Changing the rewards/payoffs associated with different responses selectively influenced starting point bias (Voss et al., 2004) while changing response modality (e.g., saccades vs. manual key presses) selectively affected nondecision time (Ho et al., 2009), consistent with the theorized role of those parameters.

Selective influence is desirable because it greatly simplifies interpreting the results of an EAM analysis. However, it is not strictly necessary. Many theoretically interesting violations of selective influence have been reported. In one prominent example, Rae et al. (2014) demonstrated that an urgency manipulation affected both accumulation rate and thresholds (a finding that has since been well replicated, e.g., Boag et al., 2019; Heathcote & Love, 2012; Palada et al., 2020; Starns et al., 2012; see also, Vandekerckhove et al., 2008). Yet other work has shown that speed—accuracy instructions can additionally affect nondecision time (Arnold et al., 2015; de Hollander et al., 2016; Donkin, Brown, Heathcote, et al., 2011; Dutilh et al., 2019; Heathcote & Love, 2012; Ho et al., 2012; Huang et al., 2015; Kelly et al., 2021; Palmer et al., 2005; Ratcliff, 2006; Rinkenauer et al., 2004; Servant et al., 2018, 2021; Voss et al., 2004; Weindel, Anders, et al., 2021; Weindel, Gajdos, et al., 2021).

Overall, this work suggests that inappropriately assuming selective influence may lead to misleading conclusions or to real effects being missed. We recommend comparing models that do and do not assume selective influence to ensure the extra complexity of more flexible models is warranted (see section ‘Comparing and Evaluating EAMs’).

Trial structure and event timing

One of the most important design considerations for model plausibility is how trials are structured in terms of the timing of events within a trial (e.g., cue and stimulus presentation). For an EAM to be a plausible model of the true decision process, the sequence and timing of events within a trial must match the processing assumptions of the model. A typical trial structure/sequence of a standard EAM is illustrated in Figure 2. In the following subsections, we discuss the components that make up a typical trial, their purpose, and common pitfalls surrounding their implementation. Note that the advice presented here allows for judicious deviations, such as when developing a model or using an extended EAM with different processing assumptions.

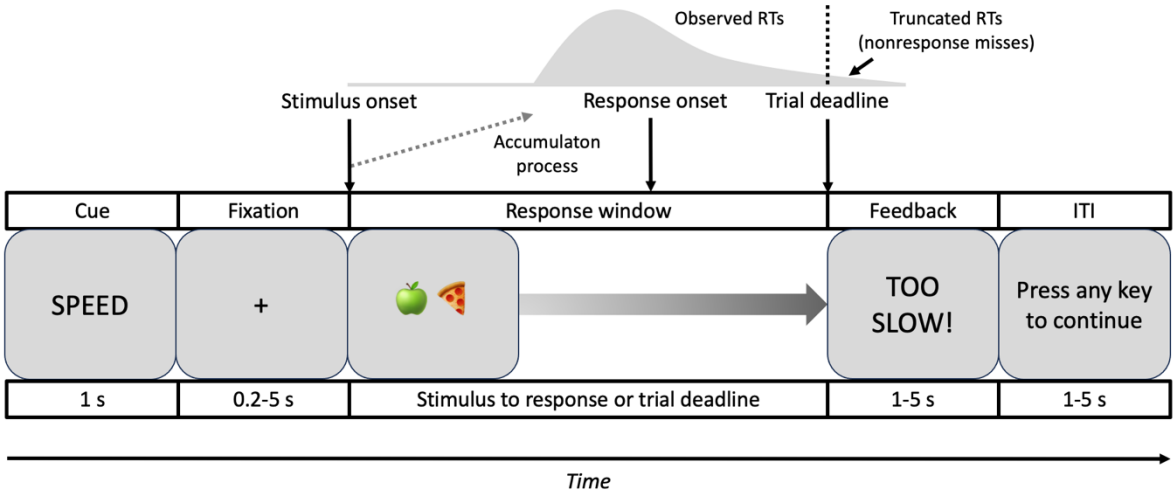

Figure 2. Structure of a typical decision trial for an EAM-appropriate task. The trial begins with a cue (e.g., instructing the participant to emphasize response speed or accuracy), followed by a fixation interval of variable (unpredictable) duration. Next, a stimulus is presented (stimulus onset) continuously until either the participant makes a response (response onset), or the trial time limit expires (which produces a nonresponse that is truncated from the RT distribution). Feedback indicating that the participant responded too slowly is then displayed. Finally, an intertrial interval gives the participant time to prepare for the next trial. The theoretical accumulation process is illustrated by the dotted arrow. Observing the outcome of many such decision trials produces a distribution of RTs with a characteristic positive skew (the density of which is illustrated in grey at the top of the figure). The presentation durations shown are suggestions only and should be calibrated to the specific task.

Cue. In some studies, trials begin with a cue that indicates how participants should perform the upcoming trial (Figure 2). The cue interval is an opportunity to present information intended to affect the decision maker's processing and cognitive control settings (e.g., thresholds and biases) prior to the decision. For example, presenting the text "Fast!" or "Accurate!" may signal that participants should respond either quickly or accurately, respectively (e.g., Forstmann et al., 2008; Katsimpokis et al., 2020). Other kinds of cues may direct participants' gaze to a particular item or spatial location (allowing comparison of attended versus unattended performance, e.g., Liu et al., 2009; Logan et al., 2023; Smith et al., 2015) or provide prior information intended to set up biases in the decision maker before encountering the stimulus (Karayanidis et al., 2009; Mulder et al., 2012; Trueblood et al., 2021).

Fixation. Fixation intervals serve the twofold purpose of concentrating participants' eye gaze/attention on the location of the upcoming stimulus (usually at the centre of the display) and of allowing time for residual processes (such as those stemming from the preceding cue or trial) to complete and return to baseline to avoid process overlap (Pashler, 1994). In a typical fixation interval, participants fixate their gaze on a centrally presented fixation cross while awaiting the stimulus. One issue that can arise with fixed-duration fixation intervals is that participants learn to anticipate the onset of the upcoming stimulus. Participants' expectation of the onset of the next trial increases over time according to a hazard function (Luce, 1991). This can lead some participants to prematurely sample evidence in anticipation of the stimulus, resulting in disproportionate anticipatory responses for longer intervals (Oswal et al., 2007), which produces biased estimates of nondecision time (Jepma et al., 2012). To avoid this problem, we recommend sampling the duration of fixation intervals from an exponential (or pseudo-exponential) distribution (e.g., with mean around 0.7 s and range of about 0.2—5 s) to avoid implausibly short intervals and excessively long waiting times (e.g., Evans & Hawkins, 2019).

Stimulus onset. Following the fixation interval, the stimulus is presented. EAMs assume that stimulus onset represents the beginning of the evidence accumulation process (plus the time taken to encode the stimulus; Bompas et al., 2023). This structural constraint makes certain tasks unsuitable for EAMs. For example, interrogation paradigms are inappropriate for standard EAMs because the decision maker first views (and presumably accumulates evidence about) the stimulus but must wait until prompted to give a response (Bogacz et al.,

2006; Ratcliff, 2006). One reason this is problematic is because the evidence accumulation process may terminate before the response prompt is presented, making it unclear what cognitive processes might have occurred in the intervening time (or what the observed RT is measuring). In sum, for the standard framework, it is crucial that the evidence accumulation process runs uninterrupted from the onset of the stimulus until the response.

Response window. The onset of the stimulus marks the beginning of the response window, which ends either when a response is submitted or upon expiry of a predefined deadline. The response window should allow enough time for participants to process and respond to the stimuli, and thus should be calibrated to the RT (and RT variability) of actual participants performing the proposed task. An inappropriately calibrated response window can lead participants to adopt undesirable/contaminant response strategies that are not accounted for in standard EAMs. For example, an excessively short response window may lead to a high proportion of fast guesses or cause slower responses to be truncated from the tail of RT distributions (i.e., responses falling outside of the response window, as illustrated in Figure 2). These processes can produce RT distributions that lack the characteristic positive skew and thus cannot be fit by standard EAMs (Evans, Hawkins, et al., 2020). Ignoring these issues can compromise parameter estimation (Verdonck & Tuerlinckx, 2016). We recommend pilot testing novel tasks to find an appropriate response window, since the optimal window will depend upon the task.

Another consideration is whether the average duration of decisions in the experimental task is appropriate for EAMs. Participants making perceptual decisions about simple psychophysical stimuli can usually respond within a 1.5 second response window. By contrast, tasks typical of cognitive psychology (e.g., lexical decision, preferential choice) may require up to 4 seconds to respond (Glickman & Usher, 2019), and more complex naturalistic tasks can take even longer (e.g., up to 10 seconds, Boag et al., 2023; Boehm et al., 2021). It is sometimes advised that standard EAMs be applied only to relatively rapid choice tasks (e.g., mean RT < 1.5 s, Ratcliff et al., 2004; Ratcliff & McKoon, 2008). This is intended to ensure that the assumption of a single continuous evidence accumulation process is upheld, since violations of the single stage assumption become increasingly plausible for decisions that unfold over longer timescales. If longer decisions do in fact involve different underlying processes, such as multiple processing stages, then they may not be accurately represented

by a standard single-stage EAM, rendering the model difficult to interpret (Heathcote, Brown, et al., 2015).

Some work suggests that standard EAMs may be a valid measurement model of more complex naturalistic decisions that unfold over longer timescales (Boehm et al., 2021), even when the assumption of a single accumulation process is explicitly violated (Boag et al., 2023; Lerche & Voss, 2019). For example, Lerche and Voss (2019) found good fits to simulated longer-RT data (mean RT ≈ 7.4 seconds) in which the single accumulation process assumption was violated. Empirical work has also found good fits to decisions with relatively long RTs (e.g., Aschenbrenner et al., 2016, Experiment 2; Glickman & Usher, 2019). Measurement-focussed application studies have also reported good fits of standard EAMs to more complex naturalistic tasks (e.g., air-traffic control, maritime surveillance, and forensic decision making) with longer RTs (2 s < mean RT < 10 s; for a review, Boag et al., 2023). In this work, experimental manipulations were found to affect model parameters in the same way in longer- and shorter-RT studies (i.e., task difficulty and stimulus discriminability effects mapped to accumulation rates; speed—accuracy trade-off, cognitive control, and bias effects mapped to thresholds and starting point).

Overall, when designing a novel task, researchers should consider whether the assumption of a single uninterrupted accumulation process is appropriate, especially in longer-RT tasks. If not, the researcher may turn to extended EAMs designed to account for phenomena associated with longer RTs, such as models that allow for slow contaminant processes (e.g., Dolan et al., 2002; Ratcliff & Tuerlinckx, 2002), randomly slow or non-terminating accumulation processes (Damaso, Castro, et al., 2022; Howard et al., 2020; Tillman et al., 2017), off-task mind wandering (Hawkins et al., 2019; Hawkins, Mittner, et al., 2015), and multiple processing stages (Little, 2012; Provost & Heathcote, 2015; Shahar et al., 2019).

Post-response interval. The post-response interval signals that the trial has ended, and a response recorded. The post-response interval provides an opportunity to display corrective feedback. For example, excessively fast or slow responding can be discouraged by displaying a warning message (e.g., “Too fast/slow!”) following such responses. Warning messages can be accompanied by a timeout interval that delays the onset of the next trial (e.g., by 1—5 s) to further encourage compliance (e.g., Evans & Hawkins, 2019). Such feedback can help to keep mean RT within the response window.

Providing feedback on performance (e.g., accuracy or points/rewards for correct responses) on experimental trials may introduce non-stationarities (e.g., post-error speeding/slowing and learning effects) that are not accounted for in the standard EAM framework (Miletić et al., 2020, 2021). Aside from during training (see section 'Task Training'), we advise against providing performance feedback for experimental trials, unless explicitly modelling learning with an extended EAM (e.g., Fontanesi et al., 2019; Miletić et al., 2021; Pedersen et al., 2017). However, since providing no feedback at all may cause participants to become disengaged from the task, it is possible to give summarized performance feedback (e.g., mean accuracy or overall points scored) following each block of trials. 'Gamifying' experiments in this way can increase participant engagement (Lumsden et al., 2016) while avoiding undesirable non-stationarities associated with trial-to-trial feedback (e.g., learning and adaptation). Moreover, such performance summaries can double as an intermittent check that participants are paying attention and performing the task as instructed.

Intertrial interval. Intertrial intervals refer to the time between trials. The intertrial interval gives participants time to 'reset' and to concentrate their attention on the upcoming trial. The intertrial interval is designed to prevent process overlap (Pashler, 1994) and to minimize other potential sources of proactive interference, such as sequential or carry-over effects stemming from events that occurred on previous trials (e.g., Aschenbrenner et al., 2018; Balota et al., 2018; Jones et al., 2013). Avoiding such interference is important for preserving stationarity, both within and across trials (i.e., for treating all trials within a condition as independent observations of the same underlying process). Intertrial intervals can be open-ended (e.g., where the participant must press a key to initiate the next trial), allowing for self-paced breaks, or can automatically progress to the next trial after some delay.

Sample size planning

Trial numbers. Researchers should plan to collect enough observations (trials) per participant in each experimental condition for reliable modelling. Doing so is important because sufficient data are required to obtain precise and unbiased individual measurement

of the EAM parameters representing each participant's latent decision processes (Smith & Little, 2018).

Much methodological work has explored how the number of trials used in fitting affects the reliability (e.g., bias, variability, and recoverability) of EAM parameters (Alexandrowicz & Gula, 2020; Lerche et al., 2017; Lerche & Voss, 2016; Lüken et al., 2023; Ratcliff & Childers, 2015; Ratcliff & Tuerlinckx, 2002; van Ravenzwaaij & Oberauer, 2009; Vandekerckhove & Tuerlinckx, 2007; Visser & Poessé, 2017; Wagenmakers et al., 2007; Wiecki et al., 2013). These studies broadly agree that around 200 trials per condition is sufficient to achieve reasonably precise and unbiased individual-level measurement. In general, more trials afford greater measurement precision and thus greater power to detect effects, since (Gaussian) measurement variance decreases with the square root of the number of measurements (trials) (Ratcliff & Tuerlinckx, 2002). However, they are diminishing returns, with simulations suggesting there is little to gain from collecting more than about 500 trials per condition (Lerche et al., 2017).

When determining the number of trials to collect, a critical question is whether there will be sufficient observations of the least frequently occurring trial type in the data (Donkin, Brown, & Heathcote, 2011). In most designs, the rarest kind of trials are incorrect responses to the most easily discriminable stimuli (i.e., incorrect responses to decisions typically made with high accuracy). However, other infrequent stimulus-response combinations are possible, such as those that arise in paradigms involving the presentation of a rare stimulus or event on a small subset of trials (e.g., Einstein & McDaniel, 1990; Loughnane et al., 2019; Strickland et al., 2018). Lüken et al. (2023) recommend obtaining error rates of at least 5% to ensure reliable parameter estimation with the standard diffusion (Ratcliff, 1978) and linear ballistic accumulator models (Brown & Heathcote, 2008). With 200 trials, a 5% error rate corresponds to 10 observations of incorrect responses. This number should be taken as a minimum: Ten error observations provided just enough information about the shape of the error RT distribution to identify the model. Fitting to data with smaller error rates is risky because the greater estimation uncertainty can make some parameters (e.g., rates and thresholds) unidentifiable (Lüken et al., 2023).

We caution that although 10 error observations may provide the bare minimum constraint needed to identify the models (e.g., by locating the mean of the incorrect RT distribution), many more observations are needed to make reliable inferences about

parameters that rely on information about the variance and skewness of the error RT distribution (e.g., the starting point and rate variability parameters for the incorrect latent response). Parameter recovery simulations can help determine how many trials (and participants) are needed to reliably measure a given effect (Heathcote, Brown, et al., 2015; White et al., 2018; Wilson & Collins, 2019). The simulation procedure is as follows: 1) set model parameters to values representative of the effect of interest, 2) simulate many synthetic participants (datasets), 3) fit the model to the synthetic data, and 4) assess how well the recovered parameters match the known data-generating values. Doing this for a range of effect sizes and for different numbers of trials and participants can help determine the most appropriate design for achieving a desired level of measurement precision (see section 'Parameter Recovery').

Clearly, there is no one-size-fits-all solution to trial number planning, since it depends upon the goals of the researcher, the size of the target effect, and properties of the model. Several thousand observations may be needed to make reliable inferences about across-trial variability parameters, or about parameters associated with rare responses (e.g., the accumulation rate of the incorrect latent response). In general, we recommend researchers use parameter recovery simulations to guide trial number planning (Heathcote, Brown, et al., 2015).

When thousands of trials are required, the experiment may need to be spread across multiple testing sessions. Long-duration experiments have several pitfalls that, if ignored, can compromise an EAM analysis. For example, participants tend to become less engaged (e.g., due to fatigue or boredom) the longer a task goes on (Cunningham et al., 2000; Krimsky et al., 2017). Disengaged or impatient participants may 'satisfice' by processing stimuli less deeply or by lowering their response criteria over time to get through an experiment more quickly (Boehm et al., 2016; Evans et al., 2019; Hawkins et al., 2012). Disengagement can introduce speeding trends and other autocorrelation effects in the data (Gong & Huskey, 2023). Additionally, longer experiments that span multiple days tend to have higher rates of participant attrition and may exacerbate already high day-to-day variability in individuals' cognitive and affective state (Schurr et al., 2024; Stevenson, Innes, et al., 2024). Such effects are problematic because standard EAMs assume data are free of such non-stationarities. These issues can be mitigated by giving participants frequent breaks,

and by using appropriate counterbalancing and trial-randomization schemes to experimentally control for time-on-task effects such as learning and fatigue.

Finally, we note that collecting a large number of trials is not always feasible. This is true for fMRI research (in which scanner time is costly and scarce; Basten et al., 2010; Forstmann et al., 2008), when studying certain clinical populations (Matzke, Hughes, et al., 2017), or when reanalysing existing data. If the use of sparse data is unavoidable, there are several techniques that can improve EAM estimation properties. These include using hierarchical models (e.g., Stevenson et al., 2024), using more informative priors (i.e., for Bayesian analyses, Lee & Vanpaemel, 2018; Matzke et al., 2020; Tran et al., 2021), constructing simpler models (e.g., by not estimating across-trial variability parameters, Boehm et al., 2018; Lerche & Voss, 2016; Ratcliff & Childers, 2015), holding some parameters constant over conditions (Donkin, Brown, & Heathcote, 2011), and by using alternative (simpler) model formulations that require only information about error proportions rather than error RT (e.g., Ludwig et al., 2009). We recommend checking the results obtained from simpler models against those obtained from a model in which the constraints are not applied (Vandekerckhove & Tuerlinckx, 2007). If both approaches arrive at the same conclusions, this provides evidence it is safe to interpret the simpler model. If not, one may need to adjust the experimental design and sampling plan until reliable model estimation is achieved.

Participant numbers. A further consideration concerning data suitability is how many participants to include in the sample. The number of participants determines how well findings generalize to the wider population and contributes to power and measurement precision in certain analyses (e.g., individual-differences correlations; Button et al., 2013; Rouder & Haaf, 2019). Studies investigating individual differences (e.g., examining correlations between EAM parameters and individual-level covariates) typically need many participants (e.g., 80 or more), each performing at least a moderate number of trials (e.g., around 200), to obtain sufficiently low measurement noise to reliably characterise potentially subtle individual differences (Rouder et al., 2023; Rouder & Haaf, 2018). Between-subjects and mixed designs also typically require many participants for sufficiently powered between-group contrasts (e.g., Boag, Strickland, Loft, et al., 2019; Steyvers et al., 2019) and to precisely characterise the distribution of population-level parameters in hierarchical Bayesian analyses (Lee, 2011).

By contrast, studies seeking to reliably measure within-subjects effects without assessing individual differences (e.g., comparing parameters for the same individual between different conditions) typically use fewer participants (e.g., as few as 3; Ratcliff & Rouder, 1998) who each perform a large number (typically thousands) of trials to ensure high individual measurement precision (Smith & Little, 2018). An advantage of fully within-subjects designs is that the unit of replication is the individual participant rather than the whole study, meaning that each participant serves as an independent replication (validation) of the target effects (Smith & Little, 2018). Replication increases confidence that obtained effects are real and meaningful.

As with trial number planning, we recommend conducting parameter recovery simulations (based on different numbers of synthetic participants) to understand how many participants are needed to obtain a desired level of power or measurement precision for a proposed analysis (White et al., 2018).

Procedural considerations

In this section, we discuss procedural considerations that can help bring participants (and the data they produce) in line with EAM assumptions. We consider task instructions, task training, and the testing environment.

Task instructions. Task instructions should be designed to maximize participant compliance with the task and to minimize undesirable behaviours that may produce data unsuitable for EAMs. Undesirable behaviours may include fast guessing, mind-wandering and inattention, waiting/delayed start-ups, random responding, and nonresponding (e.g., Cassey et al., 2014; Hawkins et al., 2019; Ratcliff & Kang, 2021). The foremost goal of instructions is to ensure that participants understand how to perform the task as intended by the researcher. This may involve explaining how a typical trial is structured and showing examples of different possible decision outcomes. Instructions should also explain key features of the task display, experiment presentation software, and response apparatus.

It is good practice to confirm that participants understand the task instructions and to provide reminders of key instructions before each testing block and following breaks or interruptions. Participant compliance/understanding can be assessed through verbal confirmation or by having participants demonstrate that they meet some performance

criterion. As a generic strategy, we recommend instructing participants to respond to each trial as quickly and accurately as possible. This instruction is designed to ensure that decisions stem from a pure (uninterrupted) evidence accumulation process, as assumed in the models. If using a manual response modality such as a computer keyboard, we suggest instructing participants to keep their fingers positioned directly above the response keys. This serves to reduce across-trial variability in nondecision time (potentially justifying its removal from the model) and ensures motor RT is as similar as possible for all participants (potentially justifying estimating a common nondecision time across participants). We recommend inviting participants to clarify any outstanding questions before commencing the experiment. Doing so may reduce the amount of data lost due to misunderstanding or noncompliance.

Task training. It is good practice to have participants perform practice/training trials before starting the experiment. Practice serves the two-fold purpose of helping participants understand the task and of stabilising performance prior to the experimental trials. Reaching a stable level of performance is important for preserving within-condition stationarity (i.e., that latent decision settings do not change within a condition). Identifying the point of stable performance is difficult since learning and adaptation may continue indefinitely for some tasks. Nevertheless, common practices include having participants practice until reaching some performance criterion (e.g., > 80% accuracy). Providing performance feedback following training trials (e.g., indicating whether the response was correct or incorrect) can help to speed up the learning/performance stabilisation process. Non-stationarities and carry-over effects (e.g., across trials and conditions) can be further minimized using appropriate randomization (e.g., randomizing the presentation of trials within a condition) and counterbalancing regimes (e.g., balancing the order of conditions within an experiment; Brooks, 2012; Lewis, 1989; Zeelenberg & Pecher, 2015).

The testing environment. The testing environment should encourage participants to perform the experimental task in the manner intended by the researcher. For most purposes, this means that participants are seated at a desk with a computer and a keyboard or other response apparatus, and a display monitor positioned at a comfortable viewing distance. In application studies, which use various high-fidelity simulated and virtual-reality environments (e.g., Castro et al., 2022; Ratcliff & Strayer, 2014; Tillman et al., 2017; Vanunu & Ratcliff, 2023), participants should be positioned appropriately for the simulator

environment. To facilitate engaged and attentive task performance, testing should be conducted in a quiet, comfortable space, free from distractions and interruptions. This is important for the EAM assumptions of model plausibility (i.e., that responses are generated by a single continuous evidence accumulation process) and of stationarity (i.e., that latent cognitive settings are stable over time).

Ideally, all participants would be tested in a single in-person session under identical conditions. However, if testing must be conducted across multiple sessions or in different locations, then conditions should be kept as consistent as possible between each session and testing location. Consistency of context is important because individuals are known to use different decision-making strategies in different contexts, such as when performing a task inside versus outside of an fMRI scanner (Forstmann et al., 2008; Van Maanen et al., 2016). Inside the scanner, participants adopted more conservative (higher) response thresholds and had longer nondecision times than they did in the out-of-scanner testing context (Van Maanen et al., 2016; see also, Forstmann et al., 2008; Gunawan et al., 2020). Ignoring or aggregating over such context effects may introduce undesirable data features (e.g., bimodal RT distributions) that may cause failures to fit and produce misleading or meaningless parameter estimates.

Online testing platforms (e.g., Mechanical Turk, Prolific, CloudResearch) give researchers the potential to collect data more quickly and affordably than is possible offline (Barbosa et al., 2023; Birnbaum, 2004). However, there are concerns that unsupervised online participants may generate poor quality data (e.g., data that are noisy, non-stationary, or generated by contaminant processes; Douglas et al., 2023; Peer et al., 2021). These concerns arise because, lacking supervision, online participants may misunderstand task instructions or be inattentive/careless (Albert & Smilek, 2023; Aruguete et al., 2019) and the remote online context makes it difficult for experimenters to identify and correct such problems (Reips, 2002). Ratcliff and Hendrickson (2021) conducted an online replication of several classic EAM studies and found that almost half of the participants in one experiment made a significant number of fast guesses (i.e., premature responses with chance accuracy) and/or produced RTs that were unstable (non-stationary) across the testing session. Nevertheless, inferences based on diffusion model parameters were largely consistent with the prior in-person studies (Ratcliff & Hendrickson, 2021). We recommend approaching online testing with appropriate caution and to avoid collecting mixed samples of online and in-person

participants. We refer readers to Gong and Huskey (2023) for more detailed advice about constructing an online testing pipeline for EAM analyses.

If context effects are suspected, we recommend accounting for these effects in the EAM analyses. This can be done in most EAM software by including a 'session' or 'testing context' factor allowing parameters to vary by context, by fitting the model to data from each context separately, or by building the additional contextual structure into a hierarchical model (e.g., Schurr et al., 2024; Stevenson, Innes, et al., 2024; Wall et al., 2021). Finding a close agreement across contexts may justify pooling data.

Collecting and recording data

EAM analysis requires certain information about each trial to be recorded. Such information is typically recorded by the software used to present the experiment and is saved in the form of a data table or comma-separated values file, in which each row represents a trial and each column an experimental or measured variable. At minimum, each row of the data should record the participant identifier, experimental condition, the presented stimulus, the submitted response and RT.

Data should include the testing session (if more than one) and trial number, and it is good practice to record the timing of events, including stimulus and response onsets, and events such as cues, feedback/reward screens, and intertrial intervals. While not everything will be used in modelling, the raw data should ideally allow one to reconstruct the trial composition and timing of the original experiment. Most EAM software will require as input a data frame of this approximate form (e.g., Heathcote et al., 2019; Stevenson et al., 2024). However, specific data and file formatting requirements will differ depending on the software/fitting routine used.

Screening data prior to EAM analysis

Prior to EAM analysis, it is important to screen data for potentially undesirable features or distributional properties that may violate EAM assumptions. Undesirable data features can include outliers (excessively fast or slow RTs), nonresponses, truncated or misshapen RT distributions, and data from participants who did not comply with task instructions. These

contaminant processes can compromise the validity of an EAM analysis. Specifically, failure to ensure data fidelity can introduce bias and uncertainty into parameter estimates (Ratcliff, 1993; Ratcliff & Tuerlinckx, 2002; Vandekerckhove & Tuerlinckx, 2007).

Outliers. Outliers are contaminant RTs that are generated by processes other than those that the researcher is interested in, and which often lie outside the range of normal observations (Berger & Kiefer, 2021; Miller, 2023). Outliers can be the result of fast guesses (e.g., guesses made without properly inspecting the stimulus), slow guesses (e.g., guesses based on a failure to reach a decision), delayed or failed start-ups (e.g., due to attentional lapses or 'trigger failures', Matzke et al., 2017; Vandekerckhove et al., 2008), or from the participant executing multiple runs of the process of interest (e.g., making multiple assessments before committing to a final response; Ratcliff, 1993; Vandekerckhove & Tuerlinckx, 2007).

The simplest and most common method for removing outliers is to define a range of acceptable RTs and to remove any observations outside of this range. For fast outliers, it is common practice to remove RTs faster than about 150—300 ms (e.g., McVay & Kane, 2012; Rae et al., 2014; White et al., 2010). This practice is motivated by the argument that, since nondecision time (for manual key presses) is typically on the order of 150—250 ms (Bompas et al., 2023), responses executed sooner than this are psychologically implausible because they allow too little time for the accumulation of evidence. A more principled method for removing fast guesses is motivated by the fact that fast guesses tend to have very short RTs and chance-level accuracy (Ratcliff & Kang, 2021; Ratcliff & Tuerlinckx, 2002; Vandekerckhove & Tuerlinckx, 2007). Consequently, one can sort RTs from fastest to slowest, find the RT at which accuracy rises above chance, and discard all RTs below the chance-performance point (Vandekerckhove & Tuerlinckx, 2007). The latter method is preferrable, although differences between approaches will likely be small unless there is a significant proportion (e.g., > 5%) of fast contaminants distorting the leading edges of the RT distributions (Ratcliff, 1993, 2013; Ratcliff & Tuerlinckx, 2002).

For slow outliers, it is more common to define an upper cut-off based on some measure of observed RT variability or to simply not censor slow outliers unless there is clear evidence of their presence. For example, some researchers censor RTs beyond 3 times the interquartile range/1.349 above the mean (a measure of standard deviation that is robust to skew; e.g., Strickland et al., 2018). Since RT variability differs between individuals, the

process of defining and removing slow outliers should be conducted separately for each participant (Miller, 2023). We caution that slow contaminants can be more difficult to detect than fast guesses, or even impossible, as although they are generated by a contaminant process, they may be hidden within the range of normal RTs (Ratcliff, 1993; Ulrich & Miller, 1994; see also, Berger & Kiefer, 2021).

Nonresponses. Nonresponses occur when a participant fails to submit a response (e.g., due to missing the response deadline). Since nonresponses result in missing values for choice and RT, standard EAM likelihood functions cannot be evaluated for nonresponses. Nonresponses are thus uninformative in fitting standard EAMs and should be excluded prior to fitting the model. Some kinds of nonresponses, such as 'trigger failures' (i.e., failures to run the evidence accumulation process, Matzke et al., 2017) can be incorporated into standard EAMs via mixture modelling (Heathcote et al., 2019) or with the aid of specialized experimental designs (Verbruggen et al., 2019).

Misshapen or non-stationary RT distributions. The geometry of standard EAMs predicts positively skewed, stationary RT distributions free of truncation (i.e., without censorship of the leading or trailing edge of an RT distribution). EAMs struggle to capture the shape of truncated distributions, since the truncation process is not accounted for in the model. Similarly, EAMs cannot predict normally distributed or negatively skewed RT distributions (Evans, Hawkins, et al., 2020), or non-stationary distributions that change in shape or scale over time (Miletić et al., 2021; Walsh et al., 2017). We recommend checking that RT distributions are positively skewed, stationary, and free of truncation. Non-stationarity can be checked by testing the correlation between RT and trial number or by dividing the RTs into sequential bins and testing for changes in mean RT/variance/skewness. Significant correlations or between-bin differences suggest non-stationarity.

Noncompliant participants. In addition to excluding individual contaminant trials, it is prudent to exclude data from participants who failed to comply with task instructions. The reason is that noncompliant participants are unlikely to have used the same cognitive strategies as compliant participants who performed the task as instructed. Consequently, standard EAMs may be a poor model of the unknown processes underlying noncompliant participants' data. One indicator of noncompliance is chance-level performance. It is common practice to exclude data from participants with near-chance performance over all or part of the experiment (e.g., Stevenson, Innes, et al., 2024).

Manipulation check. It is important to check that experimental manipulations produced the expected effects on accuracy and mean RT because it may be not worth modelling data that lack convincing behavioural effects (Palminteri et al., 2017; Wilson & Collins, 2019). Manipulation checks can be conducted by testing for differences in accuracy or mean RT using traditional or Bayesian linear models (e.g., mixed-effects regression models, Rouder et al., 2017). Bayesian approaches further allow for quantifying evidence for null effects using Bayes Factors (Dienes, 2016; Lakens et al., 2020; Morey & Rouder, 2011). A lack of convincing behavioural effects could indicate that the experimental manipulations were weak or ineffective. Nevertheless, it is possible to find theoretically interesting latent effects that are masked in accuracy or RT (Lerche & Voss, 2020). We recommend pilot testing proposed tasks on a small sample of participants to ensure novel designs/manipulations are effective.

When it comes to data exclusions, it is our view that prevention is better than a cure. Good data is a hard-won resource, and researchers should seek to minimize the amount of it lost to exclusions. We encourage researchers to take measures to minimize contaminants such as fast guesses and nonresponses and to ensure participants comply with task instructions (e.g., by providing sufficient task training and penalizing undesirable behaviours). Encouraging compliance will help maximize the data quality and minimize the data lost to exclusions. All data exclusions and exclusion criteria should be reported transparently. Further, it is good practice to check whether results are robust to exclusions (e.g., by conducting the same analysis with and without the exclusions applied).

Fitting EAMs to data

Once satisfied the data are appropriate for EAM modelling, the process of model fitting can begin. There are numerous freely available software packages that enable fitting EAMs to data (e.g., Heathcote et al., 2019; Innes et al., 2022; Stevenson et al., 2024; Vandekerckhove & Tuerlinckx, 2008; Voss et al., 2015; Voss & Voss, 2007; Wagenmakers et al., 2007, 2008; Wiecki et al., 2013). Some fitting software takes a Bayesian approach and some use frequentist methods. Software differs on which models are supported and in how readily the software can be modified or extended (e.g., to support novel models). It is beyond the scope of this article to weigh the merits of various software packages and fitting

methods. We direct interested readers to several detailed comparative studies (e.g., Alexandrowicz & Gula, 2020; Evans, 2019; Lerche et al., 2017; Ratcliff & Childers, 2015; van Ravenzwaaij & Oberauer, 2009) and to existing comprehensive resources on evaluating and troubleshooting the model fitting process (e.g., assessing convergence and diagnosing problems with sampling/fitting algorithms; Baribault & Collins, 2023; Gelman et al., 1995; Kruschke, 2014; McElreath, 2016).

We recommend fitting EAMs to the data of individuals rather than to group-aggregated data (e.g., data that has been collapsed or averaged across participants). This is because nonlinear models (such as EAMs) can produce misleading inferences when fit to aggregated data (Heathcote et al., 2015; see also, Averell & Heathcote, 2011; Brown & Heathcote, 2003; Heathcote et al., 2000). In some cases, one may want to fit just a single model, such as when the researcher has in mind a specific EAM, and clear expectations for how model parameters should change. In this case, the researcher moves on to assessing absolute fit (i.e., how well the chosen model accounts for important data features) and then on to interpreting parameters. An alternative (and more common) situation is to have several plausible models of the data, with the goal of finding the one that gives the best (e.g., most parsimonious) account of the data. Finding a good model involves assessing relative fit (i.e., how well a model accounts for data relative to other models) and absolute fit and evaluating the reliability of parameter effects. These are the topics of the next section.

Comparing and evaluating EAMs

A thorough modelling analysis involves evaluating both relative fit (a model's ability to account for data relative to other models) and absolute fit (a model's absolute ability to capture the data). Model comparison enables researchers to evaluate competing cognitive theories against one another (Pitt et al., 2002), the goal being to find the simplest model that also fits the data well (Myung & Pitt, 1997). Model comparison is important because more flexible models will have an unfair advantage in fitting data more closely than a simpler model but will also tend to predict future data less well than a simpler model that only captures robust/reliable effects (Busemeyer & Wang, 2000; Cutting et al., 1992; Myung, 2000; Myung & Pitt, 1997; Roberts & Pashler, 2000; Yarkoni & Westfall, 2017).

Model comparison requires the researcher to propose a set of candidate models, each of which constitutes a different theory of decision making, as instantiated in an EAM. For example, a researcher might be interested in whether participants' slower RTs in one condition are due to slower accumulation, higher thresholds, or longer nondecision time (or some combination thereof). The researcher would then build models that explain the effect (i.e., slower RTs) using (the appropriate combination of) accumulation rates, thresholds, or nondecision time, while holding the other parameters fixed. The proposed models may vary in complexity (e.g., the number of free parameters and how model parameters are combined in the model equations, Myung & Pitt, 1997) and in which parameters are used to explain the target effects (e.g., whether a manipulation is assumed to affect accumulation rates or thresholds or both). Moreover, researchers may seek converging evidence by fitting the same theory instantiated in different EAM architectures (e.g., using relative evidence and racing accumulator models). Doing so helps to ensure results are not dependent on the specific choice of EAM (Singmann et al., 2018).

Relative fit. Relative fit can be assessed using model comparison metrics (e.g., Akaike, 1974; Ando, 2007; Schwarz, 1978; Spiegelhalter et al., 2002; Watanabe & Opper, 2010) that account for both model fit and model complexity (for a review, Evans, 2019). These metrics can identify the model that, out of the models considered, provides the most parsimonious account of the data (i.e., offers the best trade-off between fit and complexity). Methodological work indicates that even the relatively simple 'parameter counting' metrics (e.g., AIC, Akaike, 1974; BIC, Schwarz, 1978; DIC, Spiegelhalter et al., 2002) give similar results to gold-standard methods such as Bayes Factors (Evans, 2019), which can be difficult to implement for complex cognitive models (Annis et al., 2019; Evans & Brown, 2018; Gronau, Heathcote, et al., 2020), but are argued to give the optimal trade-off between flexibility and goodness-of-fit (Jeffreys, 1998; Kass & Raftery, 1995).

When multiple models are under consideration, we recommend the 'bookending' strategy (Lee et al., 2019) in which the set of candidate models includes a minimally parameterized base model (where all target effects are removed/held fixed) and a fully flexible top model (where all target effects are included). This strategy helps establish upper and lower bounds on model complexity and to the find the model (from the set of candidate models) that provides the most parsimonious account of the data (Heathcote, Brown, et al., 2015; Lee et al., 2019; Shiffrin et al., 2008). Bookending helps to navigate the treacherous

waters between underfitting (i.e., failing to capture important data features) and overfitting (i.e., capturing noise or idiosyncratic data features).

When participants have different preferred models, it can indicate the use of distinct cognitive strategies. For example, in a speed—accuracy trade-off experiment, some participants may be better fit by a model in which urgency selectively influences thresholds, whereas others may prefer a model in which urgency affects both rates and thresholds. In such cases, we recommend reporting the proportion of participants best represented by each model.[4] We further encourage researchers to seek converging evidence (e.g., by comparing multiple complexity metrics) when choosing from among many possible models.

Absolute fit. One limitation of relative fit metrics is that there is no guarantee that a model selected in this manner actually provides a good account of the data (Box, 1976). The winner may be the best of a bad bunch. This limitation makes relative fit metrics inappropriate for falsifying models, since they only consider the relative evidence for the winning model against (an incomplete set of) rival models, while ignoring whether the winner gives an adequate account of the data (Palminteri et al., 2017). The ability to falsify models is important for scientific progress, since it allows researchers to discard bad theories (models) and to propose better ones that become the target of future falsification attempts (Popper, 2005). Falsification requires assessing the absolute fit of a model, that is, its ability to account for all the important trends in the data. A further reason assessing absolute fit is critical is that parameters derived from models that fail to capture important data features may be misleading or uninterpretable (Anscombe, 1973; Heathcote, Brown, et al., 2015).

Absolute fit is commonly assessed via visual inspection (Dutilh et al., 2019). In this method, model predictions are overlaid against empirical data (Heathcote, Brown, et al., 2015). We recommend assessing model fit to both accuracy (response proportion) and RT in each cell of the design. Fit to RT should be assessed across the entire range of RTs (e.g., by plotting fits to the 0.1, 0.5, and 0.9 RT quantiles, which correspond to the leading edge, median, and tail, of an RT distribution, respectively). Some researchers also check whether models capture higher moments (e.g., variance and skewness) of RT distributions (e.g.,

[4] For hierarchical analyses, which assume a common model across participants, one may instead investigate individual differences in the pattern (e.g., size and direction) of parameter effects.

Evans, Hawkins, et al., 2020). Conducting a thorough evaluation of absolute fit can help diagnose potential sources of mis-fit and identify where a model might be mis-specified.

We recommend visually inspecting model fits for each participant individually. Poor individual-level fits can reveal noncompliant participants (e.g., using alternate or contaminant strategies), since the EAM failed to adequately describe the processes at play. We suggest running modelling analyses with and without poorly fit participants and comparing the results of the two analyses. Points of disagreement may reveal findings that are driven by processes other than those the researcher is interested in. We caution that graphical assessment of fit is inherently subjective and thus subject to human error and judgement biases (Browne & Cudeck, 1992; Korteling & Toet, 2022; Kunda, 1990). Confidence can be increased by using multiple independent assessors (D'Agostino, 1986). For reporting purposes, it usually suffices to show the overall fit averaged over participants (although the model was fit individually), since it may be infeasible to display comprehensive model fits for potentially hundreds of individual participants.

Parameter recovery. Having chosen an adequate model, it is good practice to assess parameter recovery (Heathcote, Brown, et al., 2015). Parameter recovery refers to the practice of fitting a model to many synthetic datasets (simulated from known parameter values) and assessing whether the model consistently returns the known data-generating parameters. Recovery can be assessed graphically by plotting the correlation between true and recovered values. Parameter recovery studies have utility for establishing the reliability of model inferences and for identifying potentially unreliable (poorly recovered) parameters/effects. Parameter recovery simulations are also useful for assessing a design's suitability for modelling (in terms of trial and participant numbers) and for verifying the efficacy of experimental manipulations (in terms of expected effect size; Heathcote, Brown, et al., 2015; Miletić et al., 2017; Wilson & Collins, 2019). To generate the synthetic data used to assess recovery, one can simulate from parameter values that have been previously reported for similar tasks (Tran et al., 2021), from values (e.g., posterior means) derived from fitting the target model to prior data, or from values derived from the beliefs of subject matter experts (Gronau, Ly, et al., 2020; Kadane & Wolfson, 1998; Stefan et al., 2022). Parameter recovery should be assessed across a range of 'true' generating values, in case there are biases in specific generating-parameter ranges.

Test and interpret parameter effects. Having established a reliably estimated model that is preferred based on relative and absolute fit, focus turns to testing and interpreting parameter effects (i.e., differences across conditions or correlations) contained within the preferred model. Tests can be conducted using traditional statistical approaches (e.g., ANOVA, Ratcliff et al., 2004; *t*-tests, Voss et al., 2004) or by comparing posterior parameter distributions using Bayesian approaches (e.g., Kruschke, 2010; Meng, 1994). Establishing that there are strong parameter effects can help justify complexity in a model (Heathcote, Brown, et al., 2015). To aid interpretation, it is good practice to visualize parameter effects (e.g., by plotting parameter means and variances or credible intervals across the levels of the relevant manipulation).

Interpreting parameters involves mapping parameter effects back to cognitive theory. For example, in working memory tasks, accumulation rate effects might be interpreted in terms of differences in item activation in memory (e.g., Donkin & Nosofsky, 2012; Ratcliff, 1978; Zhou et al., 2021). By contrast, in preferential choice tasks, rate effects might be interpreted in terms of subjective utility or preference strength (e.g., Busemeyer et al., 2019; Konovalov & Krajbich, 2017). Likewise, in different tasks, threshold effects might be interpreted in terms of urgency settings (e.g., Evans, 2021) or the operation of adaptive cognitive control (e.g., Boag et al., 2019; Strickland et al., 2018). Linking parameters to broader cognitive theory helps readers understand and interpret the results of an EAM analysis.

These evaluation practices constitute a minimal set of checks intended to promote robust cognitive modelling (Lee et al., 2019), rather than an exhaustive list of best practices. A complete tutorial on evaluating EAMs is beyond the scope of this article. We point interested readers to a number of excellent sources on more advanced model evaluation techniques (e.g., Evans, 2019; Heathcote et al., 2015; Shiffrin et al., 2008). These techniques include model recovery and cross-fitting methods to assess mimicry between models (Donkin, Brown, Heathcote, et al., 2011; Evans, 2020; Hawkins, Forstmann, et al., 2015) and generalization tests to assess how well model predictions match new data and experimental contexts (Busemeyer & Wang, 2000; Vehtari et al., 2017).

Reporting an EAM analysis

We encourage researchers to carefully report all stimuli, materials, procedures, and analysis choices. Table 3 lists essential information to include when reporting an EAM analysis. The purpose of including this information is to help readers interpret and assess the quality of the analysis, and to facilitate future follow-up studies, such as replications and meta-analyses of EAM results (Theisen et al., 2021; Tran et al., 2021). Providing contextual information (e.g., justifying research goals and design choices) can help readers interpret findings and determine their scope of applicability. Thoroughly describing the experimental procedure and analysis pipeline can help readers assess the trustworthiness of your results. To promote transparency and openness in science (Hales et al., 2019; Nosek et al., 2016), we encourage researchers to openly report potential flaws of models and methods. To further encourage open and reproducible research (Crüwell, Van Doorn, et al., 2019; Gilmore et al., 2017; Munafò et al., 2017), we recommend researchers share anonymized raw data (Martone et al., 2018; Wilkinson et al., 2016) along with modelling and analysis code (McDougal et al., 2016; Wilson et al., 2019).

Table 3. Essential components to include when reporting an EAM analysis.

Analyses component	Recommended reporting practice
Research context	Provide background/context to the research question and justify all design choices. Interpret findings in relation to the broader research context.
Stimuli and materials	Describe key properties of the stimuli and how they map to the possible response options. Describe any equipment used for testing.
Task and procedure	Describe the task and any training procedures, instructions, or feedback given to participants. Report any trial-randomization or counterbalancing schemes. Report the timing (onset and duration) of all events (e.g., cue, fixation cross, stimulus, trial deadline, feedback, and intertrial interval). Report the number of participants, trials, and testing blocks, and the trial composition of each block.
Data exclusions	Report all exclusions (e.g., outliers, nonresponses, and noncompliant participants) and exclusion criteria.
Response times	Report RT mean and variance (averaged over participants) for correct and incorrect responses in each condition.

Choices	Report accuracy mean and variance (over participants) in each condition.
Measurement scale/units	Report the measurement scale/units (e.g., seconds vs. milliseconds) of behavioural measures and relevant model parameters (e.g., nondecision time).
Model parameters	Report which parameters were included in the model and over which conditions they varied. Report which parameters were not estimated (e.g., fixed as scaling constants).
Parameter coding	Report whether the model was cell coded (e.g., when different parameters are estimated for each design cell) or whether an alternative parameterisation was used.
Parameter estimates	Report descriptive statistics (e.g., means and standard deviations over participants) for all model parameters.
Model fitting method	Report the fitting method (e.g., the optimization or posterior sampling method and criteria used to assess convergence) and software used.
Model fit	Show whether the model captures the target data (e.g., by plotting model predictions against observed effects).
Model comparison	Report model comparison metrics (e.g., AIC, DIC, or Bayes Factors) and explain their interpretation.
Model evaluation	Report the results of any model evaluation procedures (e.g., parameter recovery, model mimicry, and generalisation tests).
Priors	For Bayesian analyses, describe the priors (i.e., distribution type and parameter settings) for individual- or group-level parameters.
Inferential statistics	Describe all statistical tests and inferential procedures.

Going beyond the standard models

Here we raise the issue of what to do when a proposed task violates the processing assumptions of standard EAMs or when the standard framework fails to provide an adequate account of the data. In these situations, it is prudent to first search the EAM literature to find out whether there already exists an extended EAM that may account for your data. The literature is replete with EAM variants that have been adapted to account for tasks and phenomena not accounted for in the basic EAM framework. One class of extended EAMs account for violations of within-condition stationarity due to learning (Fengler et al., 2022; Fontanesi et al., 2019; Mendonça et al., 2020; Miletić et al., 2021; Pedersen et al., 2017; Pedersen & Frank, 2020; Sewell et al., 2019). In these models, a learning rule allows

parameters to be updated from trial-to-trial in response to feedback (for a review, Miletić et al., 2020). Extensions also exist that account for various violations of within-trial stationarity. These include models that allow for within-trial changes in evidence strength (Diederich, 2024; Holmes et al., 2016; Holmes & Trueblood, 2018; Krajbich et al., 2010; Maier et al., 2020; Sepulveda et al., 2020; Sullivan et al., 2015; Weichart et al., 2022; Yang & Krajbich, 2023) or thresholds (Busemeyer & Rapoport, 1988; Evans, Hawkins, et al., 2020; Hawkins, Forstmann, et al., 2015; Smith & Ratcliff, 2022; Voskuilen et al., 2016; Voss et al., 2019; Zhang et al., 2014), and the effects of multiple, potentially conflicting, sources of evidence on the accumulation process (Lee & Sewell, 2024; Little et al., 2018; Ulrich et al., 2015; Weichart et al., 2020; White et al., 2011, 2018). Another highly active area of model development research seeks to refine the standard account of nondecision time by titrating the sensory encoding and motor components (Bompas et al., 2023; Kelly et al., 2021; Servant et al., 2021; Weindel, Gajdos, et al., 2021).

The basic framework has been extended to decisions involving more than one discrete response per trial (e.g., best—worst ranking tasks, Hawkins et al., 2014; and double-response paradigms; Evans, Dutilh, et al., 2020; Taylor et al., 2023; Ulrich & Stapf, 1984), decisions with continuous response spaces (e.g., colour-matching and continuous-scaling tasks; Kvam, 2019b, 2019a; Kvam et al., 2023; Kvam & Turner, 2021; Qarehdaghi & Amani Rad, 2022; Smith, 2016, 2019; Smith et al., 2020; Zhou et al., 2021), and decisions that involve integrating information along multiple attributes or feature dimensions (Busemeyer et al., 2019; Busemeyer & Townsend, 1993; Fific et al., 2010; Krajbich & Rangel, 2011; Nosofsky et al., 2011; Nosofsky & Palmeri, 1997; Roe et al., 2001; Strickland et al., 2023; Trueblood et al., 2014; Tsetsos et al., 2010).

If no appropriate model exists, focus turns to model development. The goal of model development is to construct a new model that accounts for phenomena that existing models do not (Crüwell, Stefan, et al., 2019). This is often accomplished by adapting or extending an existing model (e.g., Brown & Heathcote, 2005; Evans et al., 2018; Hawkins & Heathcote, 2021; Miletić et al., 2021; Ratcliff & Rouder, 1998) but can also involve constructing an entirely new model to explain the target paradigm (e.g., Ratcliff, 1978; Usher & McClelland, 2001). Model development is an iterative and exploratory process (Crüwell, Stefan, et al., 2019) and one may require specialized knowledge of mathematics and computer programming to successfully build and implement a new model. We refer interested readers

to several excellent resources on cognitive model development (Busemeyer & Diederich, 2010; Farrell & Lewandowsky, 2018; Lee & Wagenmakers, 2014).

One focus of model development concerns how to incorporate choice confidence ratings into the standard account of decision making (Lee et al., 2023; Lee & Dry, 2006; Moran et al., 2015; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009, 2013; Van Zandt & Maldonado-Molina, 2004). Confidence ratings offer a third data source (i.e., choice, RT, and confidence) with which to constrain models of decision making (Vickers, 2014). Current models make different assumptions about the how confidence rating decision trials should be structured. For example, Ratcliff and Starns (2009) measure confidence during the initial decision, while Pleskac and Busemeyer (2010) measure confidence during a subsequent additional decision stage (see also, Moran et al., 2015). This difference is critical if confidence ratings are based on different evidence before, during, and after a decision (Lee & Pezzulo, 2022, 2023). Such structural differences make it difficult to compare models (both to other confidence models and to standard EAMs), especially if eliciting the confidence rating changes how individuals perform the task. We echo Evans and Wagenmakers (2019) in expressing that "although choice confidence is an interesting additional source of data that can help us better understand certain aspects of decision-making, inferences made in these paradigms may lose direct applicability to the standard two-alternative forced choice paradigms used in most decision-making tasks" (p. 18).

Concluding remarks

Our aim in this paper was to provide practical guidance on planning experimental tasks for EAMs. To this end, we gave advice on how to design tasks that meet EAM assumptions, on how to relate experimental manipulations to EAM parameters, and on how to collect and prepare task data for EAM analysis. We discussed techniques for evaluating EAMs and warned of common pitfalls that can arise in EAM analyses. Some issues, such as sample size planning, depend upon the goals of the researcher and may require careful judgement. This article is intended as a resource to aid in planning experiments for reliable EAM analysis. By encouraging good task design practices, we hope to improve the quality and trustworthiness of future EAM studies and to help users obtain valid and interpretable results from EAMs.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Albert, D. A., & Smilek, D. (2023). Comparing attentional disengagement between Prolific and MTurk samples. *Scientific Reports*, *13*(1), 20574. https://doi.org/10.1038/s41598-023-46048-5

Alexandrowicz, R. W., & Gula, B. (2020). Comparing Eight Parameter Estimation Methods for the Ratcliff Diffusion Model Using Free Software. *Frontiers in Psychology*, *11*, 484737. https://doi.org/10.3389/fpsyg.2020.484737

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, *94*(2), 443–458. https://doi.org/10.1093/biomet/asm017

Annis, J., Evans, N. J., Miller, B. J., & Palmeri, T. J. (2019). Thermodynamic integration and steppingstone sampling methods for estimating Bayes factors: A tutorial. *Journal of Mathematical Psychology*, *89*, 67–86. https://doi.org/10.1016/j.jmp.2019.01.005

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17-21. https://doi.org/10.1080/00031305.1973.10478966

Arnold, N. R., Bröder, A., & Bayen, U. J. (2015). Empirical validation of the diffusion model for recognition memory and a comparison of parameter-estimation methods. *Psychological Research*, *79*(5), 882–898. https://doi.org/10.1007/s00426-014-0608-y

Aruguete, M. S., Huynh, H., Browne, B. L., Jurs, B., Flint, E., & McCutcheon, L. E. (2019). How serious is the 'carelessness' problem on Mechanical Turk? *International Journal of Social Research Methodology*. https://doi.org/10.1080/13645579.2018.1563966

Aschenbrenner, A. J., Balota, D. A., Gordon, B. A., Ratcliff, R., & Morris, J. C. (2016). A diffusion model analysis of episodic recognition in preclinical individuals with a family history for Alzheimer's disease: The adult children study. *Neuropsychology*, *30*(2), 225. https://psycnet.apa.org/doi/10.1037/neu0000222

Aschenbrenner, A. J., Yap, M. J., & Balota, D. A. (2018). The generality of dynamic adjustments in decision processes across trials and tasks. *Psychonomic Bulletin & Review*, *25*(5), 1917–1924. https://doi.org/10.3758/s13423-017-1359-8

Audley, R. J., & Pike, A. R. (1965). Some alternative stochastic models of choice. *British Journal of Mathematical and Statistical Psychology, 18*(2), 207-225. https://psycnet.apa.org/record/1966-06301-001

Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, *55*(1), 25–35. https://doi.org/10.1016/j.jmp.2010.08.009

Balota, D. A., Aschenbrenner, A. J., & Yap, M. J. (2018). Dynamic adjustment of lexical processing in the lexical decision task: Cross-trial sequence effects. *Quarterly Journal of Experimental Psychology*, *71*(1), 37–45. https://doi.org/10.1080/17470218.2016.1240814

Barbosa, J., Stein, H., Zorowitz, S., Niv, Y., Summerfield, C., Soto-Faraco, S., & Hyafil, A. (2023). A practical guide for studying human behavior in the lab. *Behavior Research Methods*, *55*(1), 58–76. https://doi.org/10.3758/s13428-022-01793-9

Baribault, B., & Collins, A. G. E. (2023). Troubleshooting Bayesian cognitive models. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000554

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences*, *107*(50), 21767–21772. https://doi.org/10.1073/pnas.0908104107

Batchelder, W. H. (2010). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring Psychological Constructs: Advances in Model-based Approaches* (pp. 71–93). American Psychological Association. https://doi.org/10.1037/12074-004

Batchelder, W. H. (2016). Cognitive psychometrics. In J. Houpt & L. Blaha (Eds.), *Mathematical Models of Perception and Cognition Volume I* (pp. 245–266). Psychology Press. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315647272-12/cognitive-psychometrics-william-batchelder

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*(1), 57–86. https://doi.org/10.3758/BF03210812

Berger, A., & Kiefer, M. (2021). Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in Psychology*, *12*, 675558. https://doi.org/10.3389/fpsyg.2021.675558

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology*, *55*(1), 803–832. https://doi.org/10.1146/annurev.psych.55.090902.141601

Boag, R. J., Strickland, L., Heathcote, A., Neal, A., & Loft, S. (2019). Cognitive control and capacity for prospective memory in complex dynamic environments. *Journal of*

Experimental Psychology: General, 148(12), 2181–2206. https://doi.org/10.1037/xge0000599

Boag, R. J., Strickland, L., Heathcote, A., Neal, A., Palada, H., & Loft, S. (2023). Evidence accumulation modelling in the wild: Understanding safety-critical decisions. *Trends in Cognitive Sciences*, *27*(2), 175–188. https://doi.org/10.1016/j.tics.2022.11.009

Boag, R. J., Strickland, L., Loft, S., & Heathcote, A. (2019). Strategic attention and decision control support prospective memory in a complex dual-task environment. *Cognition*, *191*, 103974. https://doi.org/10.1016/j.cognition.2019.05.011

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., Krypotos, A.-M., Lerche, V., Logan, G. D., Palmeri, T. J., van Ravenzwaaij, D., Servant, M., Singmann, H., Starns, J. J., Voss, A., Wiecki, T. V., Matzke, D., & Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46–75. https://doi.org/10.1016/j.jmp.2018.09.004

Boehm, U., Hawkins, G. E., Brown, S., van Rijn, H., & Wagenmakers, E.-J. (2016). Of monkeys and men: Impatience in perceptual decision-making. *Psychonomic Bulletin & Review*, *23*(3), 738–749. https://doi.org/10.3758/s13423-015-0958-5

Boehm, U., Marsman, M., van der Maas, H. L., & Maris, G. (2021). An attention-based diffusion model for psychometric analyses. *Psychometrika*, *86*(4), 938-972. https://doi.org/10.1007/s11336-021-09783-0

Boehm, U., van Maanen, L., Forstmann, B., & van Rijn, H. (2014). Trial-by-trial fluctuations in CNV amplitude reflect anticipatory adjustment of response caution. *NeuroImage*, *96*, 95–105. https://doi.org/10.1016/j.neuroimage.2014.03.063

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700. https://doi.org/10.1037/0033-295x.113.4.700

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. L. (2007). Extending a biologically inspired model of choice: Multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1655–1670. https://doi.org/10.1098/rstb.2007.2059

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed–accuracy tradeoff. *Trends in Neurosciences*, *33*(1), 10–16. https://doi.org/10.1016/j.tins.2009.09.002

Bompas, A., Sumner, P., & Hedge, C. (2023). Non-decision time: The Higg's boson of decision. *BioRxiv*, 2023–02. https://doi.org/10.1101/2023.02.20.529290

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Brooks, J. L. (2012). Counterbalancing for serial order carryover effects in experimental condition orders. *Psychological Methods*, *17*(4), 600–614. https://doi.org/10.1037/a0029310

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 11–21. https://doi.org/10.3758/BF03195493

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*(1), 117. https://psycnet.apa.org/doi/10.1037/0033-295X.112.1.117

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive Modeling*. SAGE Publications.

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in Cognitive Sciences*, *23*(3), 251–263. https://doi.org/10.1016/j.tics.2018.12.003

Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, *32*(2), 91–134. https://doi.org/10.1016/0022-2496(88)90042-9

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432. https://doi.org/10.1037/0033-295X.100.3.432

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*(1), 171–189. https://doi.org/10.1006/jmps.1999.1282

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Confidence and precision increase with high statistical power. *Nature Reviews Neuroscience*, *14*(8), 585–585. https://doi.org/10.1038/nrn3475-c4

Cassey, P., Heathcote, A., & Brown, S. D. (2014). Brain and behavior in decision-making. *PLoS Computational Biology*, *10*(7), e1003700. https://doi.org/10.1371/journal.pcbi.1003700

Castro, S. C., Heathcote, A., Cooper, J. M., & Strayer, D. L. (2022). Dynamic workload measurement and modeling: Driving and conversing. *Journal of Experimental Psychology: Applied, 29*(3), 645–653. https://doi.org/10.1037/xap0000431

Castro, S. C., Strayer, D. L., Matzke, D., & Heathcote, A. (2019). Cognitive workload measurement and modeling under divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(6), 826–839. https://doi.org/10.1037/xhp0000638

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, *143*(4), 1476–1488. https://doi.org/10.1037/a0035813

Cerracchio, E., Miletić, S., & Forstmann, B. U. (2023). Modelling decision-making biases. *Frontiers in Computational Neuroscience*, *17*, 1222924. https://doi.org/10.3389/fncom.2023.1222924

Chávez De la Peña, A. F., & Vandekerckhove, J. (2023). An EZ Bayesian hierarchical drift diffusion model for response time and accuracy. https://doi.org/10.31234/osf.io/yg9b5

Chen, F., & Krajbich, I. (2018). Biased sequential sampling underlies the effects of time pressure and delay in social decision making. *Nature Communications*, *9*(1), 3557. https://doi.org/10.1038/s41467-018-05994-9

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, *11*(6), 693–702. https://doi.org/10.1038/nn.2123

Copeland, A., Stafford, T., & Field, M. (2023). Recovery from nicotine addiction: A diffusion model decomposition of value-based decision-making in current smokers and ex-smokers. *Nicotine and Tobacco Research*, *25*(7), 1269–1276. https://doi.org/10.1093/ntr/ntad040

Crüwell, S., Stefan, A. M., & Evans, N. J. (2019). Robust standards in cognitive science. *Computational Brain & Behavior*, *2*(3-4), 255–265. https://doi.org/10.1007/s42113-019-00049-8

Crüwell, S., Van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., Orben, A., Parsons, S., & Schulte-Mecklenbeck, M. (2019). Seven easy steps to open science: An annotated reading list. *Zeitschrift Für Psychologie*, *227*(4), 237–248. https://doi.org/10.1027/2151-2604/a000387

Cunningham, S., Scerbo, M. W., & Freeman, F. G. (2000). The electrocortical correlates of daydreaming during vigilance tasks. *Journal of Mental Imagery*, 24(1-2), 61–72. https://psycnet.apa.org/record/2000-05564-002

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, *121*(3), 364–381. https://doi.org/10.1037/0096-3445.121.3.364

D'Agostino, R. B. (1986). Graphical analysis. In R. B. D'Agostino & M. A. Stephens (Eds.) *Goodness-of-Fit-Techniques*. Routledge.

Damaso, K. A. M., Castro, S. C., Todd, J., Strayer, D. L., Provost, A., Matzke, D., & Heathcote, A. (2022). A cognitive model of response omissions in distraction paradigms. *Memory & Cognition*, *50*(5), 962–978. https://doi.org/10.3758/s13421-021-01265-z

Damaso, K. A. M., Williams, P. G., & Heathcote, A. (2022). What happens after a fast versus slow error, and how does it relate to evidence accumulation? *Computational Brain & Behavior*, *5*(4), 527–546. https://doi.org/10.1007/s42113-022-00137-2

de Hollander, G., Labruna, L., Sellaro, R., Trutti, A., Colzato, L. S., Ratcliff, R., Ivry, R. B., & Forstmann, B. U. (2016). Transcranial direct current stimulation does not influence the speed–accuracy tradeoff in perceptual decision-making: Evidence from three independent studies. *Journal of Cognitive Neuroscience*, *28*(9), 1283–1294. https://doi.org/10.1162/jocn_a_00967

Diederich, A. (2024). A dynamic dual process model for binary choices: Serial versus parallel architecture. *Computational Brain & Behavior*, *7*(1), 37–64. https://doi.org/10.1007/s42113-023-00186-1

Diederich, A., & Trueblood, J. S. (2018). A dynamic dual process model of risky decision making. *Psychological Review*, *125*(2), 270–292. https://doi.org/10.1037/rev0000087

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Ditterich, J. (2010). A comparison between mechanisms of multi-alternative perceptual decision making: Ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. *Frontiers in Neuroscience*, *4*, 184. https://doi.org/10.3389/fnins.2010.00184

Dolan, C. V., Van Der Maas, H. L. J., & Molenaar, P. C. M. (2002). A framework for ML estimation of parameters of (mixtures of) common reaction time distributions given optional truncation or censoring. *Behavior Research Methods, Instruments, & Computers*, *34*(3), 304–323. https://doi.org/10.3758/BF03195458

Donkin, C., Averell, L., Brown, S., & Heathcote, A. (2009). Getting more from accuracy and response time data: Methods for fitting the linear ballistic accumulator. *Behavior Research Methods*, *41*(4), 1095–1110. https://doi.org/10.3758/BRM.41.4.1095

Donkin, C., & Brown, S. D. (2018). Response times and decision-making. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, *5*, 349–377. https://doi.org/10.1002/9781119170174.epcn509

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review*, *16*(6), 1129–1135. https://doi.org/10.3758/PBR.16.6.1129

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology*, *55*(2), 140–151. https://doi.org/10.1016/j.jmp.2010.10.001

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: Different models but the same conclusions about psychological processes? *Psychonomic Bulletin & Review*, *18*(1), 61–69. https://doi.org/10.3758/s13423-010-0022-4

Donkin, C., & Nosofsky, R. M. (2012). The structure of short-term memory scanning: An investigation using response time distribution models. *Psychonomic Bulletin & Review*, *19*(3), 363–394. https://doi.org/10.3758/s13423-012-0236-8

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos One*, *18*(3), e0279720. https://doi.org/10.1371/journal.pone.0279720

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A.-M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y.-S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., Van Maanen, L., … Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, *26*(4), 1051–1069. https://doi.org/10.3758/s13423-017-1417-2

Dutilh, G., Forstmann, B. U., Vandekerckhove, J., & Wagenmakers, E.-J. (2013). A diffusion model account of age differences in posterror slowing. *Psychology and Aging*, *28*(1), 64–76. https://psycnet.apa.org/doi/10.1037/a0029875

Dutilh, G., Krypotos, A.-M., & Wagenmakers, E.-J. (2011). Task-related versus stimulus-specific practice: A diffusion model account. *Experimental Psychology*, *58*(6), 434–442. https://doi.org/10.1027/1618-3169/a000111

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review*, *16*(6), 1026–1036. https://doi.org/10.3758/16.6.1026

Dutilh, G., Wagenmakers, E.-J., Visser, I., & van der Maas, H. L. J. (2011). A phase transition model for the speed-accuracy trade-off in response time experiments. *Cognitive Science*, *35*(2), 211–250. https://doi.org/10.1111/j.1551-6709.2010.01147.x

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology*, *2*(2), 312–329. https://doi.org/10.1016/0022-2496(65)90007-6

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010). Converging measures of workload capacity. *Psychonomic Bulletin & Review*, *17*(6), 763–771. https://doi.org/10.3758/PBR.17.6.763

Einstein, G. O., & McDaniel, M. A. (1990). Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 717–726. https://doi.org/10.1037/0278-7393.16.4.717

Evans, N. J. (2019). Assessing the practical differences between model selection methods in inferences about choice response time tasks. *Psychonomic Bulletin & Review*, *26*(4), 1070–1098. https://doi.org/10.3758/s13423-018-01563-9

Evans, N. J. (2020). Same model, different conclusions: An identifiability issue in the linear ballistic accumulator model of decision-making. https://psyarxiv.com/2xu7f/

Evans, N. J. (2021). Think fast! The implications of emphasizing urgency in decision-making. *Cognition*, *214*, 104704. https://doi.org/10.1016/j.cognition.2021.104704

Evans, N. J., Bennett, A. J., & Brown, S. D. (2019). Optimal or not; depends on the task. *Psychonomic Bulletin & Review*, *26*(3), 1027–1034. https://doi.org/10.3758/s13423-018-1536-4

Evans, N. J., & Brown, S. D. (2018). Bayes factors for the linear ballistic accumulator model of decision-making. *Behavior Research Methods*, *50*(2), 589–603. https://doi.org/10.3758/s13428-017-0887-5

Evans, N. J., Brown, S. D., Mewhort, D. J., & Heathcote, A. (2018). Refining the law of practice. *Psychological Review*, *125*(4), 592–605. https://doi.org/10.1037/rev0000105

Evans, N. J., Dutilh, G., Wagenmakers, E.-J., & van der Maas, H. L. J. (2020). Double responding: A new constraint for models of speeded decision making. *Cognitive Psychology*, *121*, 101292. https://doi.org/10.1016/j.cogpsych.2020.101292

Evans, N. J., & Hawkins, G. E. (2019). When humans behave like monkeys: Feedback delays and extensive practice increase the efficiency of speeded decisions. *Cognition*, *184*, 11–18. https://doi.org/10.1016/j.cognition.2018.11.014

Evans, N. J., Hawkins, G. E., & Brown, S. D. (2020). The role of passing time in decision-making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 316–326. https://doi.org/10.1037/xlm0000725

Evans, N. J., Steyvers, M., & Brown, S. D. (2018). Modeling the covariance structure of complex datasets using cognitive models: An application to individual differences and the heritability of cognitive ability. *Cognitive Science*, *42*(6), 1925–1944. https://doi.org/10.1111/cogs.12627

Evans, N. J., & Wagenmakers, E.-J. (2019). Evidence accumulation models: Current limitations and future directions. https://osf.io/preprints/psyarxiv/74df9/

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292–303. https://doi.org/10.1038/nrn2258

Farrell, S., & Lewandowsky, S. (2018). *Computational Modeling of Cognition and Behavior*. Cambridge University Press.

Fengler, A., Bera, K., Pedersen, M. L., & Frank, M. J. (2022). Beyond drift diffusion models: Fitting a broad class of decision and reinforcement learning models with HDDM. *Journal of Cognitive Neuroscience*, *34*(10), 1780–1805. https://doi.org/10.1162/jocn_a_01902

Fiedler, S., & Glöckner, A. (2012). The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in Psychology*, *3*, 25643. https://doi.org/10.3389/fpsyg.2012.00335

Fific, M., Little, D. R., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review*, *117*(2), 309–348. https://doi.org/10.1037/a0018526

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*, *26*(4), 1099–1121. https://doi.org/10.3758/s13423-018-1554-2

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*(45), 17538–17542. https://doi.org/10.1073/pnas.0805903105

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*(1), 641–666. https://doi.org/10.1146/annurev-psych-122414-033645

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E.-J., Derrfuss, J., Imperati, D., & Brown, S. (2011). The speed-accuracy tradeoff in the elderly brain: A structural model-based approach. *Journal of Neuroscience*, *31*(47), 17242–17249. https://doi.org/10.1523/JNEUROSCI.0309-11.2011

Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: Opposites attract? *Trends in Cognitive Sciences*, *15*(6), 272–279. https://doi.org/10.1016/j.tics.2011.04.002

Frazier, P., & Yu, A. J. (2007). Sequential hypothesis testing under stochastic deadlines. *Advances in Neural Information Processing Systems*, *20*, 465–472.

https://proceedings.neurips.cc/paper_files/paper/2007/hash/9c82c7143c102b71c593d98d96093fde-Abstract.html

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429258411

Gilmore, R. O., Diaz, M. T., Wyble, B. A., & Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1396*(1), 5–18. https://doi.org/10.1111/nyas.13325

Glickman, M., & Usher, M. (2019). Integration to boundary in decisions between numerical sequences. *Cognition*, *193*, 104022. https://doi.org/10.1016/j.cognition.2019.104022

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Gomez, P., Ratcliff, R., & Childers, R. (2015). Pointing, looking at, and pressing keys: A diffusion model account of response modality. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1515–1523. https://doi.org/10.1037/a0039653

Gong, X., & Huskey, R. (2023). Moving behavioral experimentation online: A tutorial and some recommendations for drift diffusion modeling. *American Behavioral Scientist*, 00027642231207073. https://doi.org/10.1177/00027642231207073

Grasman, R. P., Wagenmakers, E.-J., & Van Der Maas, H. L. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology*, *53*(2), 55–68. https://doi.org/10.1016/j.jmp.2009.01.006

Gronau, Q. F., Heathcote, A., & Matzke, D. (2020). Computing Bayes factors for evidence-accumulation models using Warp-III bridge sampling. *Behavior Research Methods*, *52*(2), 918–937. https://doi.org/10.3758/s13428-019-01290-6

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian *t* -Tests. *The American Statistician*, *74*(2), 137–143. https://doi.org/10.1080/00031305.2018.1562983

Gunawan, D., Hawkins, G. E., Tran, M.-N., Kohn, R., & Brown, S. D. (2020). New estimation approaches for the hierarchical Linear Ballistic Accumulator model. *Journal of Mathematical Psychology*, *96*, 102368. https://doi.org/10.1016/j.jmp.2020.102368

Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2019). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science*, *42*, 13–31. https://doi.org/10.1007/s40614-018-00186-8

Harris, A., & Hutcherson, C. A. (2022). Temporal dynamics of decision making: A synthesis of computational and neurophysiological approaches. *WIREs Cognitive Science*, *13*(3), e1586. https://doi.org/10.1002/wcs.1586

Hawkins, G., Brown, S. D., Steyvers, M., & Wagenmakers, E.-J. (2012). Decision speed induces context effects in choice. *Experimental Psychology*, *59*(4), 206–215. https://doi.org/10.1027/1618-3169/a000145

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *Journal of Neuroscience*, *35*(6), 2476–2484. https://doi.org/10.1523/JNEUROSCI.2410-14.2015

Hawkins, G. E., & Heathcote, A. (2021). Racing against the clock: Evidence-based versus time-based decisions. *Psychological Review*, *128*(2), 222 263. https://doi.org/10.1037/rev0000259

Hawkins, G. E., Marley, A. A. J., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). The best of times and the worst of times are interchangeable. *Decision*, *1*(3), 192–214. https://psycnet.apa.org/doi/10.1037/dec0000012

Hawkins, G. E., Mittner, M., Boekel, W., Heathcote, A., & Forstmann, B. U. (2015). Toward a model-based cognitive neuroscience of mind wandering. *Neuroscience*, *310*, 290–305. https://doi.org/10.1016/j.neuroscience.2015.09.053

Hawkins, G. E., Mittner, M., Forstmann, B. U., & Heathcote, A. (2019). Modeling distracted performance. *Cognitive Psychology*, *112*, 48–80. https://doi.org/10.1016/j.cogpsych.2019.05.002

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An Introduction to Model-Based Cognitive Neuroscience* (pp. 25–48). Springer. https://doi.org/10.1007/978-1-4939-2236-9_2

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207. https://doi.org/10.3758/BF03212979

Heathcote, A., Lin, Y.-S., Reynolds, A., Strickland, L., Gretton, M., & Matzke, D. (2019). Dynamic models of choice. *Behavior Research Methods*, *51*(2), 961–985. https://doi.org/10.3758/s13428-018-1067-y

Heathcote, A., Loft, S., & Remington, R. W. (2015). Slow down and remember to remember! A delay theory of prospective memory costs. *Psychological Review*, *122*(2), 376–410. https://doi.org/10.1037/a0038952

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Psychology*, *3*, 24343. https://doi.org/10.3389/fpsyg.2012.00292

Heitz, R. P., & Schall, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron*, *76*(3), 616–628. https://doi.org/10.1016/j.neuron.2012.08.030

Ho, T. C., Brown, S., van Maanen, L., Forstmann, B. U., Wagenmakers, E.-J., & Serences, J. T. (2012). The optimality of sensory processing during the speed–accuracy tradeoff. *Journal of Neuroscience*, *32*(23), 7992–8003. https://doi.org/10.1523/JNEUROSCI.0340-12.2012

Ho, T. C., Brown, S., & Serences, J. T. (2009). Domain general mechanisms of perceptual decision making in human cortex. *Journal of Neuroscience*, *29*(27), 8675–8687. https://doi.org/10.1523/JNEUROSCI.5984-08.2009

Holmes, W. R., & Trueblood, J. S. (2018). Bayesian analysis of the piecewise diffusion decision model. *Behavior Research Methods*, *50*(2), 730–743. https://doi.org/10.3758/s13428-017-0901-y

Holmes, W. R., Trueblood, J. S., & Heathcote, A. (2016). A new framework for modeling decisions about changing information: The Piecewise Linear Ballistic Accumulator model. *Cognitive Psychology*, *85*, 1–29. https://doi.org/10.1016/j.cogpsych.2015.11.002

Howard, Z. L., Evans, N. J., Innes, R. J., Brown, S. D., & Eidels, A. (2020). How is multi-tasking different from increased difficulty? *Psychonomic Bulletin & Review*, *27*(5), 937–951. https://doi.org/10.3758/s13423-020-01741-8

Huang, Y.-T., Georgiev, D., Foltynie, T., Limousin, P., Speekenbrink, M., & Jahanshahi, M. (2015). Different effects of dopaminergic medication on perceptual decision-making in Parkinson's disease as a function of task difficulty and speed–accuracy instructions. *Neuropsychologia*, *75*, 577–587. https://doi.org/10.1016/j.neuropsychologia.2015.07.012

Huang-Pollock, C., Ratcliff, R., McKoon, G., Roule, A., Warner, T., Feldman, J., & Wise, S. (2020). A diffusion model analysis of sustained attention in children with attention deficit hyperactivity disorder. *Neuropsychology*, *34*(6), 641–653. https://psycnet.apa.org/doi/10.1037/neu0000636

Huang-Pollock, C., Ratcliff, R., McKoon, G., Shapiro, Z., Weigard, A., & Galloway-Long, H. (2017). Using the diffusion model to explain cognitive deficits in attention deficit hyperactivity disorder. *Journal of Abnormal Child Psychology*, *45*(1), 57–68. https://doi.org/10.1007/s10802-016-0151-y

Huseynov, S., & Palma, M. A. (2021). Food decision-making under time pressure. *Food Quality and Preference*, *88*, 104072. https://doi.org/10.1016/j.foodqual.2020.104072

Innes, R., Stevenson, N., Boag, R., & Heathcote, A. (2022). *Model-based sampling with EMC2: Extended models of choice*. https://bookdown.org/reilly_innes/EMC_bookdown/

Janczyk, M., & Lerche, V. (2019). A diffusion model analysis of the response-effect compatibility effect. *Journal of Experimental Psychology: General*, *148*(2), 237–251. https://doi.org/10.1037/xge0000430

Jeffreys, H. (1998). *The Theory of Probability*. Oxford University Press, Oxford. https://books.google.com/books?hl=en&lr=&id=vh9Act9rtzQC&oi=fnd&pg=PA1&ots=fgQxGSZ9jV&sig=fe_jVAcCQBAmXrzbF8pwVZwUpUs

Jepma, M., Wagenmakers, E.-J., & Nieuwenhuis, S. (2012). Temporal expectation and information processing: A model-based analysis. *Cognition*, *122*(3), 426–441. https://doi.org/10.1016/j.cognition.2011.11.014

Johnson, D. J., Cesario, J., & Pleskac, T. J. (2018). How prior information and police experience impact decisions to shoot. *Journal of Personality and Social Psychology*, *115*(4), 601–623. https://doi.org/10.1037/pspa0000130

Johnson, D. J., Stepan, M. E., Cesario, J., & Fenn, K. M. (2021). Sleep deprivation and racial bias in the decision to shoot: A diffusion model analysis. *Social Psychological and Personality Science*, *12*(5), 638–647. https://doi.org/10.1177/1948550620932723

Jones, M., Curran, T., Mozer, M. C., & Wilder, M. H. (2013). Sequential effects in response time reveal learning mechanisms and event representations. *Psychological Review*, *120*(3), 628–666. https://doi.org/10.1037/a0033180

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, *121*(1), 1–32. https://doi.org/10.1037/a0034190

Kadane, J., & Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society Series D: The Statistician*, *47*(1), 3–19. https://doi.org/10.1111/1467-9884.00113

Karayanidis, F., Mansfield, E. L., Galloway, K. L., Smith, J. L., Provost, A., & Heathcote, A. (2009). Anticipatory reconfiguration elicited by fully and partially informative cues that validly predict a switch in task. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(2), 202–215. https://doi.org/10.3758/CABN.9.2.202

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Katsimpokis, D., Hawkins, G. E., & Van Maanen, L. (2020). Not all speed-accuracy trade-off manipulations have the same psychological effect. *Computational Brain & Behavior*, *3*(3), 252–268. https://doi.org/10.1007/s42113-020-00074-y

Kelly, S. P., Corbett, E. A., & O'Connell, R. G. (2021). Neurocomputational mechanisms of prior-informed perceptual decision-making in humans. *Nature Human Behaviour*, *5*(4), 467–481. https://doi.org/10.1038/s41562-020-00967-9

Kirkpatrick, R. P., Turner, B. M., & Sederberg, P. B. (2021). Equal evidence perceptual tasks suggest a key role for interactive competition in decision-making. *Psychological Review*, *128*(6), 1051–1087. https://psycnet.apa.org/doi/10.1037/rev0000284

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology*, *93*(3), 353–368. https://doi.org/10.1037/0022-3514.93.3.353

Konovalov, A., & Krajbich, I. (2017). Revealed indifference: Using response times to infer preferences. *Judgment and Decision Making, 14*(4) 381-394. https://doi.org/10.2139/ssrn.3024233

Korteling, J. E., & Toet, A. (2022). Cognitive biases. *Encyclopedia of Behavioral Neuroscience,* 610–619. http://dx.doi.org/10.1016/B978-0-12-809324-5.24105-9

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. https://doi.org/10.1038/nn.2635

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., & Fehr, E. (2015). A common mechanism underlying food choice and social decisions. *PLoS Computational Biology*, *11*(10), e1004371. https://doi.org/10.1371/journal.pcbi.1004371

Krajbich, I., Lu, D., Camerer, C., & Rangel, A. (2012). The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in Psychology*, *3*, 23998. https://doi.org/10.3389/fpsyg.2012.00193

Krajbich, I., Oud, B., & Fehr, E. (2014). Benefits of neuroeconomic modeling: New policy interventions and predictors of preference. *American Economic Review*, *104*(5), 501–506. https://doi.org/10.1257/aer.104.5.501

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857. https://doi.org/10.1073/pnas.1101328108

Krimsky, M., Forster, D. E., Llabre, M. M., & Jha, A. P. (2017). The influence of time on task on mind wandering and visual working memory. *Cognition*, *169*, 84–90. https://doi.org/10.1016/j.cognition.2017.08.006

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Elsevier. https://books.google.com/books?hl=en&lr=&id=FzvLAwAAQBAJ&oi=fnd&pg=PP1&dq=kruschke+doing+bayesian+data+analysis&ots=ChqjV-vf1J&sig=-DqHDtn--RjUUc_a7lHWoDvzM6E

Kruschke, J. K. (2010). Bayesian data analysis. *WIREs Cognitive Science*, *1*(5), 658–676. https://doi.org/10.1002/wcs.72

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kvam, P. D. (2019a). A geometric framework for modeling dynamic decisions among arbitrarily many alternatives. *Journal of Mathematical Psychology*, *91*, 14–37. https://doi.org/10.1016/j.jmp.2019.03.001

Kvam, P. D. (2019b). Modeling accuracy, response time, and bias in continuous orientation judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(3), 301–318. https://doi.org/10.1037/xhp0000606

Kvam, P. D., Marley, A. A. J., & Heathcote, A. (2023). A unified theory of discrete and continuous responding. *Psychological Review*, *130*(2), 368–400. https://doi.org/10.1037/rev0000378

Kvam, P. D., & Turner, B. M. (2021). Reconciling similarity across models of continuous selections. *Psychological Review*, *128*(4), 766–786. https://psycnet.apa.org/doi/10.1037/rev0000296

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, *75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

Laming, D. R. J. (1968). *Information theory of choice-reaction times.* Academic Press. https://psycnet.apa.org/record/1968-35026-000

Larson, J. S., & Hawkins, G. E. (2023). Speed-accuracy tradeoffs in decision making: Perception shifts and goal activation bias decision thresholds. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(1), 1–32. https://doi.org/10.1037/xlm0000913

Lee, D. G., Daunizeau, J., & Pezzulo, G. (2023). Evidence or confidence: What is really monitored during a decision? *Psychonomic Bulletin & Review*, *30*(4), 1360–1379. https://doi.org/10.3758/s13423-023-02255-9

Lee, D. G., & Pezzulo, G. (2022). Choice-induced preference change under a sequential sampling model framework. *bioRxiv*, 2022–07. https://doi.org/10.1101/2022.07.15.500254

Lee, D. G., & Pezzulo, G. (2023). Changes in preferences reported after choices are informative, not merely statistical artifacts. *Decision*, *10*(2), 181–195. https://doi.org/10.1037/dec0000207

Lee, D. G., & Usher, M. (2023). Value certainty in drift-diffusion models of preferential choice. *Psychological Review*, *130*(3), 790–806. https://psycnet.apa.org/doi/10.1037/rev0000329

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7. https://doi.org/10.1016/j.jmp.2010.08.013

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., Matzke, D., Rouder, J. N., Trueblood, J. S., White, C. N., & Vandekerckhove, J. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*(3), 141–153. https://doi.org/10.1007/s42113-019-00029-y

Lee, M. D., & Dry, M. J. (2006). Decision making and confidence given uncertain advice. *Cognitive Science*, *30*(6), 1081–1095. https://doi.org/10.1207/s15516709cog0000_71

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*(1), 114–127. https://doi.org/10.3758/s13423-017-1238-3

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge university press. https://books.google.com/books?hl=en&lr=&id=Gq6kAgAAQBAJ&oi=fnd&pg=PR10&dq=Lee+%26+Wagenmakers,+2014&ots=tzDcGEFpqy&sig=ByOKcYNNtzvnXPpDyW_Ju0gEEtA

Lee, P.-S., & Sewell, D. K. (2024). A revised diffusion model for conflict tasks. *Psychonomic Bulletin & Review*, *31*(1), 1–31. https://doi.org/10.3758/s13423-023-02288-0

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making*, *6*(7), 651–687. https://doi.org/10.1017/S1930297500002680

Leontyev, A., & Yamauchi, T. (2021). Discerning mouse trajectory features with the drift diffusion model. *Cognitive Science*, *45*(10), e13046. https://doi.org/10.1111/cogs.13046

Lerche, V., & Voss, A. (2016). Model complexity in diffusion modeling: Benefits of making the model more parsimonious. *Frontiers in Psychology*, *7*, 186038. https://doi.org/10.3389/fpsyg.2016.01324

Lerche, V., & Voss, A. (2018). Speed–accuracy manipulations and diffusion modeling: Lack of discriminant validity of the manipulation or of the parameter estimates? *Behavior Research Methods*, *50*(6), 2568–2585. https://doi.org/10.3758/s13428-018-1034-7

Lerche, V., & Voss, A. (2019). Experimental validation of the diffusion model based on a slow response time paradigm. *Psychological Research*, *83*(6), 1194–1209. https://doi.org/10.1007/s00426-017-0945-8

Lerche, V., & Voss, A. (2020). When accuracy rates and mean response times lead to false conclusions: A simulation study based on the diffusion model. *The Quantitative Methods for Psychology*, *16*(2), 107–119. https://psycnet.apa.org/record/2020-28133-002

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods*, *49*(2), 513–537. https://doi.org/10.3758/s13428-016-0740-2

Lewis, J. R. (1989). Pairs of Latin squares to counterbalance sequential effects and pairing of conditions and stimuli. *Proceedings of the Human Factors Society Annual Meeting*, *33*(18), 1223–1227. https://doi.org/10.1177/154193128903301812

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*(1), 77–105. https://doi.org/10.1007/BF02291481

Little, D. R. (2012). Numerical predictions for serial, parallel, and coactive logical rule-based models of categorization response time. *Behavior Research Methods*, *44*(4), 1148–1156. https://doi.org/10.3758/s13428-012-0202-4

Little, D. R., Eidels, A., Fifić, M., & Wang, T. S. (2018). How do information processing systems deal with conflicting information? Differential predictions for serial, parallel, and coactive models. *Computational Brain & Behavior*, *1*, 1–21. https://doi.org/10.1007/s42113-018-0001-9

Liu, C. C., Wolfgang, B. J., & Smith, P. L. (2009). Attentional mechanisms in simple visual detection: A speed–accuracy trade-off analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(5), 1329–1345. https://doi.org/10.1037/a0014255

Logan, G. D., Lilburn, S. D., & Ulrich, J. E. (2023). The spotlight turned inward: The time-course of focusing attention on memory. *Psychonomic Bulletin & Review*, *30*(3), 1028–1040. https://doi.org/10.3758/s13423-022-02222-w

Loughnane, G. M., Brosnan, M. B., Barnes, J. J., Dean, A., Nandam, S. L., O'Connell, R. G., & Bellgrove, M. A. (2019). Catecholamine modulation of evidence accumulation during perceptual decision formation: A randomized trial. *Journal of Cognitive Neuroscience*, *31*(7), 1044–1053. https://doi.org/10.1162/jocn_a_01393

Luce, R. D. (1991). *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford University Press. https://books.google.com/books?hl=en&lr=&id=WSmpNN5WCw0C&oi=fnd&pg=PA1&dq=luce+1986&ots=XtFVO7a2gO&sig=2xyp_TUWSLYUqfz_hNiWaNJ6StI

Ludwig, C. J., Farrell, S., Ellis, L. A., & Gilchrist, I. D. (2009). The mechanism underlying inhibition of saccadic return. *Cognitive Psychology*, *59*(2), 180–202. https://doi.org/10.1016/j.cogpsych.2009.04.002

Lüken, M., Heathcote, A., Haaf, J. M., & Matzke, D. (2023). Parameter identifiability in evidence-accumulation models: The effect of error rates on the diffusion decision model and the linear ballistic accumulator. https://doi.org/10.31234/osf.io/wsgnt

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR Serious Games*, *4*(2), e5888. https://doi.org/10.2196/games.5888

Maier, S. U., Raja Beharelle, A., Polanía, R., Ruff, C. C., & Hare, T. A. (2020). Dissociable mechanisms govern when and how strongly reward attributes affect decisions. *Nature Human Behaviour*, *4*(9), 949–963. https://doi.org/10.1038/s41562-020-0893-y

Martone, M. E., Garcia-Castro, A., & VandenBos, G. R. (2018). Data sharing in psychology. *American Psychologist*, *73*(2), 111–125. https://doi.org/10.1037%2Famp0000242

Matzke, D., Hughes, M., Badcock, J. C., Michie, P., & Heathcote, A. (2017). Failures of cognitive control or attention? The case of stop-signal deficits in schizophrenia. *Attention, Perception, & Psychophysics*, *79*(4), 1078–1086. https://doi.org/10.3758/s13414-017-1287-8

Matzke, D., Logan, G. D., & Heathcote, A. (2020). A cautionary note on evidence-accumulation models of response inhibition in the stop-signal paradigm. *Computational Brain & Behavior*, *3*(3), 269–288. https://doi.org/10.1007/s42113-020-00075-x

Matzke, D., Love, J., & Heathcote, A. (2017). A Bayesian approach for estimating the probability of trigger failures in the stop-signal paradigm. *Behavior Research Methods*, *49*(1), 267–281. https://doi.org/10.3758/s13428-015-0695-8

McDougal, R. A., Bulanova, A. S., & Lytton, W. W. (2016). Reproducibility in computational neuroscience models and simulations. *IEEE Transactions on Biomedical Engineering*, *63*(10), 2021–2035. https://doi.org/10.1109/TBME.2016.2539602

McDougle, S. D., & Collins, A. G. E. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, *28*(1), 20–39. https://doi.org/10.3758/s13423-020-01774-z

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315372495

McVay, J. C., & Kane, M. J. (2012). Drifting from slow to "D'oh!": Working memory capacity and mind wandering predict extreme reaction times and executive-control errors. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(3), 525–549. https://doi.org/10.1037/a0025896

Mendonça, A. G., Drugowitsch, J., Vicente, M. I., DeWitt, E. E., Pouget, A., & Mainen, Z. F. (2020). The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs. *Nature Communications*, *11*(1), 2757. https://doi.org/10.1038/s41467-020-16196-7

Meng, X.-L. (1994). Posterior predictive p -values. *The Annals of Statistics*, *22*(3), 1142–1160. https://doi.org/10.1214/aos/1176325622

Michmizos, K. P., & Krebs, H. I. (2014). Reaction time in ankle movements: A diffusion model analysis. *Experimental Brain Research*, *232*(11), 3475–3488. https://doi.org/10.1007/s00221-014-4032-8

Miletić, S., Boag, R. J., & Forstmann, B. U. (2020). Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia*, *136*, 107261. https://doi.org/10.1016/j.neuropsychologia.2019.107261

Miletić, S., Boag, R. J., Trutti, A. C., Stevenson, N., Forstmann, B. U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks. *Elife*, *10*, e63055. https://doi.org/10.7554/eLife.63055

Miletić, S., Turner, B. M., Forstmann, B. U., & van Maanen, L. (2017). Parameter recovery for the leaky competing accumulator model. *Journal of Mathematical Psychology*, *76*, 25–50. https://doi.org/10.1016/j.jmp.2016.12.001

Miller, J. (2023). Outlier exclusion procedures for reaction time analysis: The cures are generally worse than the disease. *Journal of Experimental Psychology: General*. https://psycnet.apa.org/record/2023-93160-001

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, *5*(6), 437–449. https://doi.org/10.1017/S1930297500001285

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147. https://doi.org/10.1016/j.cogpsych.2015.01.002

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. https://psycnet.apa.org/doi/10.1037/a0024377

Mulder, M. J., Van Maanen, L., & Forstmann, B. U. (2014). Perceptual decision neurosciences–a model-based review. *Neuroscience*, *277*, 872–884. https://doi.org/10.1016/j.neuroscience.2014.07.031

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience*, *32*(7), 2335–2343. https://doi.org/10.1523/JNEUROSCI.4156-11.2012

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190–204. https://doi.org/10.1006/jmps.1999.1283

Myung, I. J., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*(3–4), 53–67. https://doi.org/10.1016/j.jmp.2013.05.005

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95. https://doi.org/10.3758/BF03210778

Myung, I. J., Pitt, M., Tang, Y., & Cavagnaro, D. R. (2009). Bayesian adaptive optimal design of psychology experiments. *Proceedings of the 2nd International Workshop in*

Sequential Methodologies (IWSM2009), 1–6. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=9df47ce222958be33a068d5af1f122b44dd46f77

Niwa, M., & Ditterich, J. (2008). Perceptual decisions between multiple directions of visual motion. *Journal of Neuroscience*, *28*(17), 4435–4445. https://doi.org/10.1523/JNEUROSCI.5564-07.2008

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S., Breckler, S., Buck, S., Chambers, C., Chin, G., & Christensen, G. (2016). Transparency and openness promotion (TOP) guidelines. https://doi.org/10.31219/osf.io/vj54c

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, *118*(2), 280–315. https://doi.org/10.1037/a0022494

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300. https://doi.org/10.1037/0033-295X.104.2.266

Nosofsky, R. M., & Palmeri, T. J. (2015). An exemplar-based random-walk model of categorization and recognition. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology* (pp. 142–164). Oxford University Press.

Nunez, M. D., Fernandez, K., Srinivasan, R., & Vandekerckhove, J. (2024). A tutorial on fitting joint models of M/EEG and behavior to understand cognition. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02331-x

Nunez, M. D., Schubert, A.-L., Frischkorn, G. T., & Oberauer, K. (2023). Cognitive models of decision-making with identifiable parameters: Diffusion Decision Models with within-trial noise. https://osf.io/preprints/psyarxiv/h4fde/

Oswal, A., Ogden, M., & Carpenter, R. H. S. (2007). The time course of stimulus expectation in a saccadic decision task. *Journal of Neurophysiology*, *97*(4), 2722–2730. https://doi.org/10.1152/jn.01238.2006

Palada, H., Neal, A., Strayer, D., Ballard, T., & Heathcote, A. (2019). Using response time modeling to understand the sources of dual-task interference in a dynamic environment. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(10), 1331–1345. https://doi.org/10.1037/xhp0000672

Palada, H., Neal, A., Tay, R., & Heathcote, A. (2018). Understanding the causes of adapting, and failing to adapt, to time pressure in a complex multistimulus environment. *Journal of Experimental Psychology: Applied*, *24*(3), 380–399. https://doi.org/10.1037/xap0000176

Palada, H., Searston, R. A., Persson, A., Ballard, T., & Thompson, M. B. (2020). An evidence accumulation model of perceptual discrimination with naturalistic stimuli. *Journal of Experimental Psychology: Applied*, *26*(4), 671–691. https://psycnet.apa.org/doi/10.1037/xap0000272

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 1. https://doi.org/10.1167/5.5.1

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Pashler, H. (1994). Overlapping mental operations in serial performance with preview. *The Quarterly Journal of Experimental Psychology: Section A*, *47*(1), 161–191. https://doi.org/10.1080/14640749408401148

Pedersen, M. L., & Frank, M. J. (2020). Simultaneous hierarchical Bayesian parameter estimation for reinforcement learning and drift diffusion models: A tutorial and links to neural data. *Computational Brain & Behavior*, *3*(4), 458–471. https://doi.org/10.1007/s42113-020-00084-w

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, *24*(4), 1234–1251. https://doi.org/10.3758/s13423-016-1199-y

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491. https://doi.org/10.1037/0033-295X.109.3.472

Plant, R. R., Hammond, N., & Whitehouse, T. (2002). Toward an experimental timing standards lab: Benchmarking precision in the real world. *Behavior Research Methods, Instruments, & Computers*, *34*(2), 218–226. https://doi.org/10.3758/BF03195446

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. https://doi.org/10.1037/a0019737

Pleskac, T. J., Cesario, J., & Johnson, D. J. (2018). How race affects evidence accumulation during the decision to shoot. *Psychonomic Bulletin & Review*, *25*(4), 1301–1330. https://doi.org/10.3758/s13423-017-1369-6

Popper, K. (2005). *The Logic of Scientific Discovery*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9780203994627/logic-scientific-discovery-karl-popper

Provost, A., & Heathcote, A. (2015). Titrating decision processes in the mental rotation task. *Psychological Review*, *122*(4), 735–754. https://doi.org/10.1037/a0039706

Qarehdaghi, H., & Amani Rad, J. (2022). An EZ-circular diffusion model of continuous decision processes. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/5z09c72m

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1226–1243. https://doi.org/10.1037/a0036801

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, *53*(3), 195–237. https://doi.org/10.1016/j.cogpsych.2005.10.002

Ratcliff, R. (2013). Parameter variability and distributional assumptions in the diffusion model. *Psychological Review*, *120*(1), 281–292. https://doi.org/10.1037/a0030775

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision*, *2*(4), 237–279. https://doi.org/10.1037/dec0000030

Ratcliff, R., & Hendrickson, A. T. (2021). Do data from mechanical Turk subjects replicate accuracy, response time, and diffusion modeling results? *Behavior Research Methods*, *53*(6), 2302–2325. https://doi.org/10.3758/s13428-021-01573-x

Ratcliff, R., & Kang, I. (2021). Qualitative speed-accuracy tradeoff effects can be explained by a diffusion/fast-guess mixture model. *Scientific Reports*, *11*(1), 15169. https://doi.org/10.1038/s41598-021-94451-7

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, *95*(3), 385–408. https://doi.org/10.1037/0033-295x.95.3.385

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 127–140. https://doi.org/10.1037//0096-1523.26.1.127

Ratcliff, R., Scharre, D. W., & McKoon, G. (2022). Discriminating memory disordered patients from controls using diffusion model parameters from recognition memory. *Journal of Experimental Psychology: General*, *151*(6), 1377–1393. https://doi.org/10.1037/xge0001133

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367. https://doi.org/10.1037/0033-295X.111.2.333

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59–83. https://doi.org/10.1037%2Fa0014086

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, *120*(3), 697–719. https://doi.org/10.1037/a0033152

Ratcliff, R., & Strayer, D. (2014). Modeling simple driving tasks with a one-boundary diffusion model. *Psychonomic Bulletin & Review*, *21*(3), 577–589. https://doi.org/10.3758/s13423-013-0541-x

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging*, *19*(2), 278–289. https://doi.org/10.1037%2F0882-7974.19.2.278

Ratcliff, R., Thapar, A., & Mckoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics*, *65*(4), 523–535. https://doi.org/10.3758/BF03194580

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50*(4), 408–424. https://doi.org/10.1016/j.jml.2003.11.002

Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin & Review*, *13*(4), 626–635. https://doi.org/10.3758/BF03193973

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481. https://doi.org/10.3758/BF03196302

Ratcliff, R., & Van Dongen, H. P. A. (2011). Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation. *Proceedings of the National Academy of Sciences*, *108*(27), 11285–11290. https://doi.org/10.1073/pnas.1100483108

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*(2), 261–300. https://doi.org/10.1037/0033-295X.106.2.261

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256. https://doi.org/10.1026/1618-3169.49.4.243

Rinkenauer, G., Osman, A., Ulrich, R., Müller-Gethmann, H., & Mattes, S. (2004). On the locus of speed-accuracy trade-off in reaction time: Inferences from the lateralized readiness potential. *Journal of Experimental Psychology: General*, *133*(2), 261–282. https://doi.org/10.1037/0096-3445.133.2.261

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367. https://doi.org/10.1037/0033-295x.107.2.358

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review*, *108*(2), 370–392. https://doi.org/10.1037/0033-295X.108.2.370

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19–26. https://doi.org/10.1177/2515245917745058

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Rouder, J. N., Kumar, A., & Haaf, J. M. (2023). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychonomic Bulletin & Review*, *30*(6), 2049–2066. https://doi.org/10.3758/s13423-023-02293-3

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*(2), 304–321. https://doi.org/10.1037/met0000057

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*(2), 491–513. https://doi.org/10.1007/s11336-013-9396-3

Sandry, J., & Ricker, T. J. (2022). Motor speed does not impact the drift rate: A computational HDDM approach to differentiate cognitive and motor speed. *Cognitive Research: Principles and Implications*, *7*(1), 66. https://doi.org/10.1186/s41235-022-00412-7

Schall, J. D. (2019). Accumulators, neurons, and response time. *Trends in Neurosciences*, *42*(12), 848–860. https://doi.org/10.1016/j.tins.2019.10.001

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to

working memory and intelligence. *Journal of Experimental Psychology: General*, *136*(3), 414–429. https://doi.org/10.1037/0096-3445.136.3.414

Schurr, R., Reznik, D., Hillman, H., Bhui, R., & Gershman, S. J. (2024). Dynamic computational phenotyping of human cognition. *Nature Human Behaviour*, 1–15. https://doi.org/10.1038/s41562-024-01814-x

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464. http://www.jstor.org/stable/2958889

Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., & De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *eLife*, *9*, e60705. https://doi.org/10.7554/eLife.60705

Servant, M., Logan, G. D., Gajdos, T., & Evans, N. J. (2021). An integrated theory of deciding and acting. *Journal of Experimental Psychology: General*, *150*(12), 2435–2454. https://psycnet.apa.org/doi/10.1037/xge0001063

Servant, M., van Wouwe, N., Wylie, S. A., & Logan, G. D. (2018). A model-based quantification of action control deficits in Parkinson's disease. *Neuropsychologia*, *111*, 26–35. https://doi.org/10.1016/j.neuropsychologia.2018.01.014

Servant, M., White, C., Montagnini, A., & Burle, B. (2016). Linking theoretical decision-making mechanisms in the Simon task with electrophysiological data: A model-based neuroscience study in humans. *Journal of Cognitive Neuroscience*, *28*(10), 1501–1521. https://doi.org/10.1162/jocn_a_00989

Sewell, D. K., Jach, H. K., Boag, R. J., & Van Heer, C. A. (2019). Combining error-driven models of associative learning with evidence accumulation models of decision-making. *Psychonomic Bulletin & Review*, *26*(3), 868–893. https://doi.org/10.3758/s13423-019-01570-4

Sewell, D. K., & Smith, P. L. (2012). Attentional control in visual signal detection: Effects of abrupt-onset and no-onset stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 1043–1068. https://doi.org/10.1037/a0026591

Sewell, D. K., & Stallman, A. (2020). Modeling the effect of speed emphasis in probabilistic category learning. *Computational Brain & Behavior*, *3*(2), 129–152. https://doi.org/10.1007/s42113-019-00067-6

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939. https://doi.org/10.1016/j.neuron.2016.04.036

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N., & Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS Computational Biology*, *15*(2), e1006803. https://doi.org/10.1371/journal.pcbi.1006803

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*(8), 1248–1284. https://doi.org/10.1080/03640210802414826

Singmann, H., Kellen, D., Mizrak, E., & Öztekin, I. (2018). Using ensembles of cognitive models to answer substantive questions. *The Society for the Improvement of Psychological Science*. https://doi.org/10.31234/osf.io/rg4sz

Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*(1), 167–183. https://doi.org/10.1016/j.jmp.2009.06.007

Smith, P. L. (2010). From Poisson shot noise to the integrated Ornstein–Uhlenbeck process: Neurally principled models of information accumulation in decision-making and

response time. *Journal of Mathematical Psychology*, *54*(2), 266–283. https://doi.org/10.1016/j.jmp.2009.12.002

Smith, P. L. (2016). Diffusion theory of decision making in continuous report. *Psychological Review*, *123*(4), 425–451. https://doi.org/10.1037/rev0000023

Smith, P. L. (2019). Linking the diffusion model and general recognition theory: Circular diffusion with bivariate-normally distributed drift rates. *Journal of Mathematical Psychology*, *91*, 145–158. https://doi.org/10.1016/j.jmp.2019.06.002

Smith, P. L. (2023). "Reliable organisms from unreliable components" revisited: The linear drift, linear infinitesimal variance model of decision making. *Psychonomic Bulletin & Review*, *30*(4), 1323–1359. https://doi.org/10.3758/s13423-022-02237-3

Smith, P. L., & Lilburn, S. D. (2020). Vision for the blind: Visual psychophysics and blinded inference for decision models. *Psychonomic Bulletin & Review*, *27*(5), 882–910. https://doi.org/10.3758/s13423-020-01742-7

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

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. https://doi.org/10.1016/j.tins.2004.01.006

Smith, P. L., & Ratcliff, R. (2022). Modeling evidence accumulation decision processes using integral equations: Urgency-gating and collapsing boundaries. *Psychological Review*, *129*(2), 235–267. https://doi.org/10.1037/rev0000301

Smith, P. L., & Ratcliff, R. (2024). An introduction to the diffusion model of decision-making. In B. U. Forstmann & B. M. Turner (Eds.), *An Introduction to Model-Based Cognitive*

Neuroscience (pp. 67–100). Springer International Publishing. https://doi.org/10.1007/978-3-031-45271-0_4

Smith, P. L., Saber, S., Corbett, E. A., & Lilburn, S. D. (2020). Modeling continuous outcome color decisions with the circular diffusion model: Metric and categorical properties. *Psychological Review*, *127*(4), 562–590. https://psycnet.apa.org/doi/10.1037/rev0000185

Smith, P. L., & Sewell, D. K. (2013). A competitive interaction theory of attentional selection and decision making in brief, multielement displays. *Psychological Review*, *120*(3), 589–627. https://doi.org/10.1037/a0033140

Smith, P. L., Sewell, D. K., & Lilburn, S. D. (2015). From shunting inhibition to dynamic normalization: Attentional selection and decision-making in brief visual displays. *Vision Research*, *116*, 219–240. https://doi.org/10.1016/j.visres.2014.11.001

Souza, A. S., & Frischkorn, G. T. (2023). A diffusion model analysis of age and individual differences in the retro-cue benefit. *Scientific Reports*, *13*(1), 17356. https://doi.org/10.1038/s41598-023-44080-z

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Starns, J. J. (2014). Using response time modeling to distinguish memory and decision processes in recognition and source tasks. *Memory & Cognition*, *42*(8), 1357–1372. https://doi.org/10.3758/s13421-014-0432-z

Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed–accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging*, *25*(2), 377–390. https://doi.org/10.1037/a0018022

Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language*, *70*, 36–52. https://doi.org/10.1016/j.jml.2013.09.005

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, *64*(1–2), 1–34. https://doi.org/10.1016/j.cogpsych.2011.10.002

Stefan, A. M., Evans, N. J., & Wagenmakers, E.-J. (2022). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*, *27*(2), 177–197. https://doi.org/10.1037/met0000354

Stevenson, N., Donzallaz, M., Innes, R. J., Forstmann, B., Matzke, D., & Heathcote, A. (2024). *EMC2: An R Package for Cognitive Models of Choice*. https://doi.org/10.31234/osf.io/2e4dq

Stevenson, N., Innes, R. J., Boag, R. J., Miletić, S., Isherwood, S. J. S., Trutti, A. C., Heathcote, A., & Forstmann, B. U. (2024). Joint modelling of latent cognitive mechanisms shared across decision-making domains. *Computational Brain & Behavior*, *7*(1), 1–22. https://doi.org/10.1007/s42113-023-00192-3

Steyvers, M., Hawkins, G. E., Karayanidis, F., & Brown, S. D. (2019). A large-scale analysis of task switching practice effects across the lifespan. *Proceedings of the National Academy of Sciences*, *116*(36), 17735–17740. https://doi.org/10.1073/pnas.1906788116

Stine, G. M., Zylberberg, A., Ditterich, J., & Shadlen, M. N. (2020). Differentiating between integration and non-integration strategies in perceptual decision making. *Elife*, *9*, e55365. https://doi.org/10.7554/eLife.55365

Stone, C., Mattingley, J. B., & Rangelov, D. (2022). On second thoughts: Changes of mind in decision-making. *Trends in Cognitive Sciences*, *26*(5), 419–431. https://doi.org/10.1016/j.tics.2022.02.004

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260. https://doi.org/10.1007/BF02289729

Strickland, L., Boag, R. J., Heathcote, A., Bowden, V., & Loft, S. (2023). Automated decision aids: When are they advisors and when do they take control of human decision making? *Journal of Experimental Psychology: Applied*, *29*(4), 849–868. https://doi.org/10.1037/xap0000463

Strickland, L., Loft, S., Remington, R. W., & Heathcote, A. (2018). Racing to remember: A theory of decision control in event-based prospective memory. *Psychological Review*, *125*(6), 851–887. https://doi.org/10.1037/rev0000113

Sullivan, N., Hutcherson, C., Harris, A., & Rangel, A. (2015). Dietary self-control is related to the speed with which attributes of healthfulness and tastiness are processed. *Psychological Science*, *26*(2), 122–134. https://doi.org/10.1177/0956797614559543

Taylor, G. J., Nguyen, A. T., & Evans, N. J. (2023). Does allowing for changes of mind influence initial responses? *Psychonomic Bulletin & Review*, 1–13. https://doi.org/10.3758/s13423-023-02371-6

Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological Review*, *120*(1), 1–38. https://doi.org/10.1037/a0030776

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging*, *18*(3), 415–429. https://doi.org/10.1037/0882-7974.18.3.415

Theisen, M., Lerche, V., Von Krause, M., & Voss, A. (2021). Age differences in diffusion model parameters: A meta-analysis. *Psychological Research*, *85*(5), 2012–2021. https://doi.org/10.1007/s00426-020-01371-8

Tillman, G., Strayer, D., Eidels, A., & Heathcote, A. (2017). Modeling cognitive load effects of conversation between a passenger and driver. *Attention, Perception, & Psychophysics*, *79*(6), 1795–1803. https://doi.org/10.3758/s13414-017-1337-2

Tillman, G., Van Zandt, T., & Logan, G. D. (2020). Sequential sampling models without random between-trial variability: The racing diffusion model of speeded decision making. *Psychonomic Bulletin & Review*, *27*(5), 911–936. https://doi.org/10.3758/s13423-020-01719-6

Todd, A. R., Johnson, D. J., Lassetter, B., Neel, R., Simpson, A. J., & Cesario, J. (2021). Category salience and racial bias in weapon identification: A diffusion modeling approach. *Journal of Personality and Social Psychology*, *120*(3), 672–693. https://doi.org/10.1037/pspi0000279

Tran, N.-H., van Maanen, L., Heathcote, A., & Matzke, D. (2021). Systematic parameter reviews in cognitive modeling: Towards a robust and cumulative characterization of psychological processes in the diffusion decision model. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.608287

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, *121*(2), 179–205. https://doi.org/10.1037/a0036137

Trueblood, J. S., Eichbaum, Q., Seegmiller, A. C., Stratton, C., O'Daniels, P., & Holmes, W. R. (2021). Disentangling prevalence induced biases in medical image decision-making. *Cognition*, *212*, 104713. https://doi.org/10.1016/j.cognition.2021.104713

Tsetsos, K., Usher, M., & Chater, N. (2010). Preference reversal in multiattribute choice. *Psychological Review*, *117*(4), 1275–1293. https://doi.org/10.1037/a0020580

Tsetsos, K., Usher, M., & McClelland, J. L. (2011). Testing multi-alternative decision models with non-stationary evidence. *Frontiers in Neuroscience*, *5*, 8457. https://doi.org/10.3389/fnins.2011.00063

Turner, B. M., Forstmann, B. U., & Steyvers, M. (2019). *Joint Models of Neural and Behavioral Data*. Springer International Publishing. https://doi.org/10.1007/978-3-030-03688-1

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206. https://doi.org/10.1016/j.neuroimage.2013.01.048

Turner, B. M., Palestro, J. J., Miletić, S., & Forstmann, B. U. (2019). Advances in techniques for imposing reciprocity in brain-behavior relations. *Neuroscience & Biobehavioral Reviews*, *102*, 327–336. https://doi.org/10.1016/j.neubiorev.2019.04.018

Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, *123*(1), 34–80. https://doi.org/10.1037/0096-3445.123.1.34

Ulrich, R., Schröter, H., Leuthold, H., & Birngruber, T. (2015). Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. *Cognitive Psychology*, *78*, 148–174. https://doi.org/10.1016/j.cogpsych.2015.02.005

Ulrich, R., & Stapf, K. H. (1984). A double-response paradigm to study stimulus intensity effects upon the motor system in simple reaction time experiments. *Perception & Psychophysics*, *36*(6), 545–558. https://doi.org/10.3758/BF03207515

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

Usher, M., Olami, Z., & McClelland, J. L. (2002). Hick's law in a stochastic race model with speed–accuracy tradeoff. *Journal of Mathematical Psychology*, *46*(6), 704–715. https://doi.org/10.1006/jmps.2002.1420

Van Maanen, L., Brown, S. D., Eichele, T., Wagenmakers, E.-J., Ho, T., Serences, J., & Forstmann, B. U. (2011). Neural correlates of trial-to-trial fluctuations in response caution. *Journal of Neuroscience*, *31*(48), 17488–17495. https://doi.org/10.1523/JNEUROSCI.2924-11.2011

Van Maanen, L., Forstmann, B. U., Keuken, M. C., Wagenmakers, E.-J., & Heathcote, A. (2016). The impact of MRI scanner environment on perceptual decision-making. *Behavior Research Methods*, *48*(1), 184–200. https://doi.org/10.3758/s13428-015-0563-6

van Ravenzwaaij, D., Brown, S. D., Marley, A. A. J., & Heathcote, A. (2020). Accumulating advantages: A new conceptualization of rapid multiple choice. *Psychological Review*, *127*(2), 186–215. https://doi.org/10.1037/rev0000166

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2017). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review*, *24*(2), 547–556. https://doi.org/10.3758/s13423-016-1081-y

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2012). A diffusion model decomposition of the effects of alcohol on perceptual decision making. *Psychopharmacology*, *219*(4), 1017–1025. https://doi.org/10.1007/s00213-011-2435-9

van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology*, *53*(6), 463–473. https://doi.org/10.1016/j.jmp.2009.09.004

Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1147–1166. https://doi.org/10.1037/0278-7393.30.6.1147

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–1026. https://doi.org/10.3758/BF03193087

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, *40*(1), 61–72. https://doi.org/10.3758/BRM.40.1.61

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. (2008). A Bayesian approach to diffusion process models of decision-making. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 1429–1434.

Vanunu, Y., & Ratcliff, R. (2023). The effect of speed-stress on driving behavior: A diffusion model analysis. *Psychonomic Bulletin & Review*, *30*(3), 1148–1157. https://doi.org/10.3758/s13423-022-02200-2

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Verbruggen, F., Aron, A. R., Band, G. P., Beste, C., Bissett, P. G., Brockett, A. T., Brown, J. W., Chamberlain, S. R., Chambers, C. D., & Colonius, H. (2019). A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *Elife*, *8*, e46323. https://doi.org/10.7554/eLife.46323

Verdonck, S., & Tuerlinckx, F. (2016). Factoring out nondecision time in choice reaction time data: Theory and implications. *Psychological Review*, *123*(2), 208–218. https://doi.org/10.1037/rev0000019

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, *13*(1), 37–58. https://doi.org/10.1080/00140137008931117

Vickers, D. (2014). *Decision Processes in Visual Perception*. Academic Press.

Visser, I., & Poessé, R. (2017). Parameter recovery, bias and standard errors in the linear ballistic accumulator model. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 280–296. https://doi.org/10.1111/bmsp.12100

Voskuilen, C., Ratcliff, R., & Smith, P. L. (2016). Comparing fixed and collapsing boundary versions of the diffusion model. *Journal of Mathematical Psychology*, *73*, 59–79. https://doi.org/10.1016/j.jmp.2016.04.008

Voss, A., Lerche, V., Mertens, U., & Voss, J. (2019). Sequential sampling models with variable boundaries and non-normal noise: A comparison of six models. *Psychonomic Bulletin & Review*, *26*(3), 813–832. https://doi.org/10.3758/s13423-018-1560-4

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*(7), 1206–1220. https://doi.org/10.3758/BF03196893

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*(4), 767–775. https://doi.org/10.3758/BF03192967

Voss, A., Voss, J., & Lerche, V. (2015). Assessing cognitive processes with diffusion model analyses: A tutorial based on fast-dm-30. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00336

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159. https://doi.org/10.1016/j.jml.2007.04.006

Wagenmakers, E.-J., van der Maas, H. L., Dolan, C. V., & Grasman, R. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review*, *15*, 1229–1235. https://doi.org/10.3758/PBR.15.6.1229

Wagenmakers, E.-J., Van Der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22. https://doi.org/10.3758/BF03194023

Wall, L., Gunawan, D., Brown, S. D., Tran, M.-N., Kohn, R., & Hawkins, G. E. (2021). Identifying relationships between cognitive processes across tasks, contexts, and time. *Behavior Research Methods*, *53*(1), 78–95. https://doi.org/10.3758/s13428-020-01405-4

Walsh, M. M., Gunzelmann, G., & Van Dongen, H. P. A. (2017). Computational cognitive modeling of the temporal dynamics of fatigue from sleep loss. *Psychonomic Bulletin & Review*, *24*(6), 1785–1807. https://doi.org/10.3758/s13423-017-1243-6

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(12). https://www.jmlr.org/papers/volume11/watanabe10a/watanabe10a.pdf

Weichart, E. R., Galdo, M., Sloutsky, V. M., & Turner, B. M. (2022). As within, so without, as above, so below: Common mechanisms can support between-and within-trial category learning dynamics. *Psychological Review*, *129*(5), 1104–1143. https://doi.org/10.1037/rev0000381

Weichart, E. R., Turner, B. M., & Sederberg, P. B. (2020). A model of dynamic, within-trial conflict resolution for decision making. *Psychological Review*, *127*(5), 749–777. https://doi.org/10.1037/rev0000191

Weigard, A., Huang-Pollock, C., Brown, S., & Heathcote, A. (2018). Testing formal predictions of neuroscientific theories of ADHD with a cognitive model–based approach. *Journal of Abnormal Psychology*, *127*(5), 529–539. https://doi.org/10.1037/abn0000357

Weindel, G., Anders, R., Alario, F., & Burle, B. (2021). Assessing model-based inferences in decision making with single-trial response time decomposition. *Journal of Experimental Psychology: General*, *150*(8), 1528–1555. https://doi.org/10.1037/xge0001010

Weindel, G., Gajdos, T., Burle, B., & Alario, F.-X. (2021). The decisive role of non-decision time for interpreting the parameters of decision making models. https://doi.org/10.31234/osf.io/gewb3

White, C. N., Ratcliff, R., & Starns, J. J. (2011). Diffusion models of the flanker task: Discrete versus gradual attentional selection. *Cognitive Psychology*, *63*(4), 210–238. https://doi.org/10.1016/j.cogpsych.2011.08.001

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology*, *54*(1), 39–52. https://doi.org/10.1016/j.jmp.2010.01.004

White, C. N., Servant, M., & Logan, G. D. (2018). Testing the validity of conflict drift-diffusion models for use in estimating cognitive processes: A parameter-recovery study. *Psychonomic Bulletin & Review*, *25*(1), 286–301. https://doi.org/10.3758/s13423-017-1271-2

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, *7*, 55610. https://doi.org/10.3389/fninf.2013.00014

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., & Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Wilson, M. K., Boag, R. J., & Strickland, L. (2019). All models are wrong, some are useful, but are they reproducible? Commentary on Lee et al. (2019). *Computational Brain & Behavior*, *2*, 239-241. https://doi.org/10.1007/s42113-019-00054-x

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Yang, J., Pitt, M. A., Ahn, W.-Y., & Myung, I. J. (2021). ADOpy: A python package for adaptive design optimization. *Behavior Research Methods*, *53*(2), 874–897. https://doi.org/10.3758/s13428-020-01386-4

Yang, X., & Krajbich, I. (2023). A dynamic computational model of gaze and choice in multi-attribute decisions. *Psychological Review*, *130*(1), 52–70. https://doi.org/10.1037/rev0000350

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53–79. https://doi.org/10.1037/a0024177

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zeelenberg, R., & Pecher, D. (2015). A method for simultaneously counterbalancing condition order and assignment of stimulus materials to conditions. *Behavior Research Methods*, *47*, 127–133. https://doi.org/10.3758/s13428-014-0476-9

Zhang, S., Lee, M. D., Vandekerckhove, J., Maris, G., & Wagenmakers, E.-J. (2014). Time-varying boundaries for diffusion models of decision making and response time. *Frontiers in Psychology*, *5*, 112331. https://doi.org/10.3389/fpsyg.2014.01364

Zhou, J., Osth, A. F., Lilburn, S. D., & Smith, P. L. (2021). A circular diffusion model of continuous-outcome source memory retrieval: Contrasting continuous and threshold accounts. *Psychonomic Bulletin & Review*, *28*(4), 1112–1130. https://doi.org/10.3758/s13423-020-01862-0