Int. J. Human–Computer Studies 185 (2024) 103220

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Towards the application of evidence accumulation models in the design of (semi-)autonomous driving systems – an attempt to overcome the sample size roadblock

Dominik Bachmann [a,b,*], Leendert van Maanen [b]

[a] Institute for Logic, Language and Computation, University of Amsterdam
[b] Department of Experimental Psychology, Utrecht University

ARTICLE INFO

Keywords:
Automated vehicles
Human-automated vehicle interaction
Evidence accumulation models
Decision Diffusion models
Bayesian hierarchical modeling
Sample size

ABSTRACT

For the foreseeable future, automated vehicles (AVs) will coexist on the roads with human drivers. To avoid accidents, AVs will require knowledge on how human drivers typically make high-stakes and time-sensitive decisions (e.g., whether or not to brake). Providing such insights could be statistical models designed to explain human information processing and decision making. This paper attempts to address a roadblock that prevents one class of such "cognitive models", evidence accumulation models (EAMs), from being widely applied in the design of AV systems: their high demands for data. Specifically, we investigate whether Bayesian hierarchical modeling can be used to determine a person's characteristics, if we only have limited data about their behavior but extensive data on other (comparable) people's behaviors. Leveraging a simulation study and a reanalysis of experimental data, we find that most parameters of Decision Diffusion Models (a class of EAMs) – representing information processing components – can be adequately estimated with as few as 20 observations, if prior information regarding the decision-making processes of the population is incorporated. Subsequently, we discuss the implications of our findings for the modeling of traffic situations.

1. Introduction

Some life-changing decisions have to be made in less than a second, and making a wrong choice is costly. Every day, many drivers face such decisions when they choose between hitting their brakes or waiting a tad longer to observe. Whichever choice they make, if it is the wrong one for the situation, it can lead to serious consequences: The driver who braked when there was no need (e.g., because they thought they perceived something) can cause a rear-end collision with their sudden and unpredictable behavior; so can the driver who does not immediately brake when they have to.

Human drivers differ in the skill with which they make these choices (i.e., their speed and accuracy in making such split-second decisions) as well as in their proclivities: Some drivers will err on the side of braking while others are hesitant to brake. If human and automated vehicles (AVs) are to coexist in traffic, (semi-)automated driving systems are confronted with such human decision making processes in at least two ways: Firstly, they need an accurate model of how human drivers make such choices – to anticipate and adjust to the behavior of human drivers that they encounter (e.g., drivers ahead of them that could brake, or drivers behind them that react to the AI system's braking). Secondly, systems that share driving duties with a human driver (the likely highest achievable level of automation for the foreseeable future; Janssen et al., 2022; Noy et al., 2018) need an accurate model of their (co-)driver's driving attributes and likely behaviors. For example, for levels of automation in which the AV is not constantly supervised by a human driver (e.g., levels 3-4 of SAE's taxonomy for driving automation; SAE International, 2014), there might be situations in which the AI system makes higher quality decisions (in speed and accuracy) than its driver (e.g., braking while the driver is tired), while the driver outperforms it in other situations (e.g., braking in snow) and the AI system would need to be able to accurately "perceive" these performance discrepancies.[1]

* Corresponding author.
E-mail addresses: d.bachmann@uva.nl (D. Bachmann), l.vanmaanen@uu.nl (L. van Maanen).

[1] While we will mostly stick with that example throughout our paper, braking is not the only part of driving for which a driving system would benefit from an understanding of how human drivers make time-sensitive high-stakes decisions on the road. For example, a driving system misjudging whether or when a human driver changes lanes (e.g., to overtake another car or to switch lanes on a multilane road) could similarly result in accidents.

https://doi.org/10.1016/j.ijhcs.2024.103220
Received 31 January 2023; Received in revised form 27 November 2023; Accepted 3 January 2024
Available online 11 January 2024

From cognitive psychological research, it is well-known that these split-second decisions share a common information processing mechanism. That is, in less than a second, the cognitive system accrues evidence for viable courses of actions and commits to the action which it perceives to have the highest chance of success (see e.g., Boag et al., 2022). These cognitive processes can be described by mathematical models. One class of such so-called *cognitive models* is especially well-suited for explaining and predicting human choices in time-sensitive situations (like split-second decisions about whether or not to brake). These so-called *Evidence Accumulation Models* (EAMs; Boag et al., 2022; Ratcliff et al., 2016) express individual differences in decision making aspects like speed, accuracy and decision tendencies (e.g., whether a particular driver is particularly "trigger-happy" with their brake). One way in which we could provide AI driving systems with an understanding of how human drivers make decisions could hence be to inform their assumptions about human drivers with insights stemming from evidence accumulation models.

1.1. The problem of sample sizes

While EAMs could provide a lot of insight about important facets of human (traffic) decision making, and have been employed amply to explain human traffic behaviors (Section 2.1), their application to the design of (semi-)autonomous driving systems has so far been limited. Integration (e.g., using insights from cognitive models to make automated driving systems more adaptable; Janssen et al., 2022) remains a challenge, and to further enable it, one roadblock will have to be addressed first: To fit EAMs, one requires a lot of data – routinely researchers collect 100 or more observations of a type of decision to fit these models on participants' choice behaviors. This severely limits the extent to which EAMs can be utilized to inform driving system about human decision making. Even if sophisticated driving systems are developed that can identify and analyze driving situations as they occur naturalistically during driving (e.g., overtaking, unexpected braking, dodging animals on the road), it would take months to years to gather sufficient data to adequately fit EAMs for some of these driving situations. High-stakes split-second decisions are critical to model but each individual driver rarely encounters them in daily driving.

The alternatives are also suboptimal: If data were collected upon the purchase of a car (e.g., to personalize human-system interactions before the buyer starts driving), drivers would need to spent hours performing (simulated) driving tasks, so that enough data can be collected to fit models for each relevant driving situation. Few drivers would be willing to endure such an ordeal, especially if calibration is only temporary and has to be repeated regularly (e.g., due to age-related changes in cognitive processing; Archambeau et al, 2020; Ratcliff & Vanunu, 2022). Another undesirable alternative is to neglect individual differences altogether when fitting EAMs (as do e.g., Pekkanen et al., 2022; Van Maanen et al., 2012; Zgonnikov et al., 2023). Then, AVs and driving system do not take into account different drivers' strengths and weaknesses but treat each driver as the amalgamated "average human driver". This both neglects one of EAMs' main strengths (i.e., accounting for individual differences) and can be dangerous in applied settings (e.g., if the driving system overestimates a human driver's perceptual abilities by assuming that they are average).

In this paper, we investigated a potential way in which individual differences could be modeled, without requiring so much data. Specifically, we perform a simulation study (Section 3) and reanalyze data from a perceptual study (Section 4) to assess the extent to which adequate model parameter values can be obtained with few trials, if we make use of Bayesian estimation techniques.

2. Background

Before describing our simulation study and reanalysis in more detail, this section provides necessary background information for understanding these. Specifically, we provide a short introduction into one class of Evidence Accumulation Models, Diffusion Decision Models (DDM; see Section 2.1) and discuss how it has been used to model driving behavior (Section 2.2). Subsequently, we describe Bayesian parameter estimation techniques (Section 2.3) which might decrease the amount of data that is required to fit Evidence Accumulation Models, before finally (Section 2.4) describing our contribution and setup.

As subsections 2.1 and 2.3 are merely introductory, seasoned cognitive and Bayesian modelers may elect to skip one or both of them.

2.1. Evidence Accumulation Models and the Diffusion Decision Model

Evidence Accumulation Models (EAMs) are cognitive models that describe human decision making. They have been used to model decision-making in a variety of contexts (see e.g., Ratcliff et al., 2016), including driving (see e.g., Ratcliff, 2015; Ratcliff & Vanunu, 2022). Conceptually, EAMs assume that decisions are made based on a gradual accumulation of evidence relevant to a particular choice (e.g., to brake or not to brake), until enough evidence has been accrued to commit to a decision, at which point the decided-on action (e.g., braking) is irrevocably initiated. EAMs assume that individuals differ in decision-making attributes like the efficiency of the evidence accumulation process (how quickly and precisely they can analyze relevant information), and the amount of evidence they need to commit to a choice. Individual differences on EAM's parameter values are assumed to reflect individual differences in cognitive processing and decision making attributes.

While much of our conceptual discussion here (e.g., of how EAM's insights are relevant to AV design) can be extended to other classes of EAMs, as EAMs model similar decision making processes (but e.g., focus on different types of decisions humans face), our discussion in this paper focuses on Diffusion Decision Models (DDMs; Ratcliff, 1978; Ratcliff & McKoon, 2008). We focus on this class of evidence accumulation models, because it is the seminal one, and because it has been used to model cognitive processes involved in several traffic situations that are relevant to AV systems design. DDMs are designed to explain situations in which humans make choices between two options (e.g., whether or not to brake). Like other EAMs, the model estimates unique values for each person on parameters that represent aspects of human decision making. In its simplest form (which we use in our reanalysis; Section 4), each person has unique values on four such parameters: The *drift rate* (v in Fig. 1) represents how efficiently people accumulate evidence (e.g., process information they perceive on the road) towards one of the two decision options (which are represented by the two dashed horizontal lines on the top and bottom of Fig. 1). Once the evidence that was accumulated meets one of the thresholds (i.e., touches either horizontal boundary line in Fig. 1), a decision has been made and the associated action (e.g., braking) is initiated.

The parameter *boundary separation* (a in Fig. 1) represents the distance between these two decision boundaries and hence expresses how much evidence people require to make decisions. Everything else being equal, people with smaller values for the boundary separation make rasher decisions, as they need to accumulate less evidence to decide on one of the two options. People with larger boundary separation values are less prone to errors, but also slower at making decisions.

People can have proclivities for one response option over the other, before even starting to accumulate evidence towards either option. These a priori preferences are modeled by a parameter representing the starting position of evidence accumulation (z in Fig. 1). If the starting position is closer to the upper boundary, the person has an a priori inclination towards the option associated with this upper boundary, as they need to accumulate less evidence to commit to it. This would, for example, be the case for drivers that, when faced with a split-second decision about whether or not to brake, err on the side of braking.

The *non-decision time* (t_0) parameter represents all time that is not part of the cognitive decision-making process which passes between the onset of a choice situation and the execution of a response. It includes

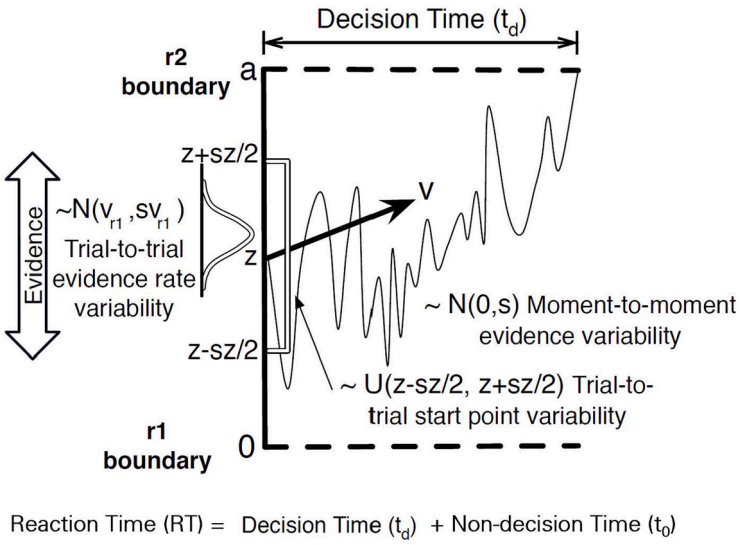

Fig. 1. Parameters of the Decision Diffusion Model (DDM)
Note. Figure adapted from Heathcote et al. (2019). Over time (from left to right) evidence is continuously accumulated towards one of two choice options (i.e., moving the line upwards or downwards). When sufficient evidence is accumulated (the line of evidence accumulation touches either the top or bottom horizontal line), a choice has been made. The nondecision time parameter (t_0) models time that passes before and after a decision has been made (e.g., time necessary for executing a response). The boundary separation (a) represents the length of the vertical line (i.e., how much evidence needs to be accumulated to make a choice). The start point (z) parameter represents where (on average) on the vertical line the evidence accumulation begins (i.e., whether closer to the top or the bottom decision).

processes such as identifying choice-relevant information (e.g., parsing a visual scene for information that is relevant towards making a choice between braking or not braking) and executing a behavioral response (like stepping on a brake) once it has been decided on.

With these four parameters, DDMs can jointly predict a distribution of response times as well as a distribution of choices: Evidence (x) starts each trial at z^*a and accumulates over time (t) as $dx = v^*dt + \varepsilon$, with ε representing standard normally distributed noise. The model emits a response as soon as $x \leq 0$or $x \geq a$ (i.e., the evidence exceeds either the lower or the upper boundary), at which point t_d represents the decision time (t_d in Fig. 1). The duration of a trial (or reaction time; RT) is the sum of its decision (t_d) and non-decision time (t_0). Often, the non-decision time is sampled from a uniform distribution whose shape depends on a mean (t_0) as well as a width parameter (s_{t0}). In this study, we simplified and just estimated a mean t_0 (i.e., a constant that gets added to each participants' reaction time), as it significantly sped up model fit.

While these four parameters are the main ones driving behavior, there are several ways in which DDMs can be extended. For example, DDMs can model the consistency with which individuals analyze information (i.e., the extent to which they show trial-by-trial drift rate variability; s_v), and how variable they are trial-by-trial in their starting point position (s_z; see also Fig. 1). Other examples of extensions of the DDM are models (discussed also in Section 2.2) which take into account how perceived stimuli change over time (e.g., how the speed and distance of a nearby car changes across multiple seconds; Zgonnikov et al., 2023) by allowing parameters (e.g., the drift rate) to change over time. Such DDMs are relevant especially to traffic decision-making which unfolds over multiple seconds (e.g., pedestrians choosing whether or not to cross a road; Pekkanen et al., 2022; or cars deciding whether or not to take a turn before another car; Zgonnikov et al., 2022) where road users continuously monitor (changes in) the relevant stimuli.

For all parameters (and their underlying cognitive processes), there is substantial inter- and intrapersonal variability that is relevant to predicting people's behavior on the road. As an example of systematic interpersonal differences, older adult drivers have been found to have lower drift rates (i.e., accumulate evidence less efficiently) than younger adult drivers, resulting in them making more mistakes on the road (Ratcliff & Vanunu, 2022). Additionally, a person's parameters differ based on their state (e.g., whether or not they are tipsy; Van Ravenzwaaij et al., 2012; or whether or not they are sleep deprived; Ratcliff & Van Dongen, 2009). Besides such systematic differences, there are also nonsystematic differences on parameters that are relevant to predicting how people behave on the road. For example, some people generally need more time to move their foot to the brake pedal (a tendency that would be modeled with a larger value for their average nondecision time t_0). In our view, one of the biggest conceptual strengths of Evidence Accumulation Models (like the DDM) is that they can account for such individual differences.

2.2. Using EAMs to model driving behavior

In preventing road accidents, both the speed with which humans make choices as well as their accuracy in making these choices are important. If road users make the wrong choice (e.g., steering to the side in an attempt to dodge an animal when such a choice is dangerous) it can endanger their lives and the lives of others. Conversely, the right choice is worth little if the road user took too long to arrive at it (e.g., drivers realizing they should brake while they are already colliding). As EAMs (such as DDMs) model both the speed and the accuracy with which people make choices, they have been widely applied to modeling behavior and decision making in traffic situations.

Research ranges from modeling general driving behaviors (e.g., braking and overtaking; Ratcliff, 2015; Ratcliff & Strayer, 2014) to modeling situations that might be particularly likely to result in accidents. For examples of the latter, several papers have used EAMs to investigate the effects of divided attention and cognitive load on driving (Castro et al., 2019; Engström et al., 2018; Tillman et al., 2017) or to investigate how drivers' behaviors change if they feel speed-pressure (Vanunu & Ratcliff, 2022). Additionally, there is variability in the types of analyzed data, ranging from data of simulated driving experiments (e.g., Ratcliff & Strayer, 2014; Xue et al., 2018) to data and recordings from actual (near-)crashes (e.g., Engström et al., 2022; Svärd et al., 2021). While the former allows for more stringent experimental control and for testing rare or hypothetical scenarios (e.g., how human drivers interact at intersections with fully automatic driver-less cars; Zgonnikov et al., 2023), it is unclear to what extent their results generalize to real-world driving where there is no unambiguous "trial onset", but drivers continuously observe the traffic until their

expectations (e.g., of other drivers' behaviors; Engström et al., 2022) are violated and they must react.

Recent work has focused on driving situations that unfold over multiple seconds. These include, for example, pedestrians choosing whether or when to cross the road (Giles et al., 2019; Markkula et al., 2018; Pekkanen et al., 2022) or drivers interacting with each other at intersections (Zgonnikov et al., 2022, 2023). As such traffic situations involve continuous (re-)assessments of stimuli over time (e.g., while pedestrians accumulate evidence about whether to cross a road before or after an incoming car, changes in the car's speed are relevant to that choice), these papers extend DDMs with kinematics-dependent components (e.g., have car acceleration-dependent drift rates; Zgonnikov et al., 2023). Recent work on braking also reflects this shift towards kinematics-dependent EAMs. Especially visual looming (i.e., the expansion over time of a stimulus on a person's retina; see e.g., Durrani & Lee, 2023; Xue et al., 2018) has received a lot of research attention, being part, for example, of models of drivers' braking when cruise control fails (Bianchi et al., 2020) and of models that account for evidence accumulation during off-road glances (Svärd et al., 2021).

As this study is meant as a proof of concept, we focus here on the conceptually simpler main parameterization of the DDM. While their kinematics-dependent extensions conceptually offer a lot of promise, we believe that the traditional (time-independent) DDMs still have a role to play when modelling traffic situations. The time-dependency of these kinematics-dependent models is particularly relevant for traffic situations that play out over multiple seconds (e.g., the nonverbal communication between a human driver that approaches an intersection and an AV decelerating to indicate that the driver's priority is being respected; Zgonnikov et al., 2023). For decision making on smaller timescales (e.g., unanticipated split-second decisions about whether or not to brake), we believe that the increases in model fit will not always be sufficient to justify introducing the additional complexity (over base DDMs). Then the choice between kinematics-dependent and traditional DDMs should be made based on rigorous model comparison.

The cost of additional model complexity is also evident from the fact that many studies utilizing such kinematics-dependent models (e.g., Pekkanen et al., 2022; Zgonnikov et al., 2023) neglect individual differences and pool data across participants (i.e., modeling the amalgamated "average human driver"; see Section 1.1). In situations where fitting kinematics-dependent models might not be practically viable (e.g., if human drivers are unwilling to spend hours on driving tasks for the adjustment of AI co-drivers), fitting simpler traditional DDMs could be a viable alternative. This would be the case, especially so, if we can decrease the sample size demands of traditional DDMs, using Bayesian estimation techniques.

Additionally, as we want to assess the effect of Bayesian estimation techniques (i.e., utilizing informative prior distributions; see next session), traditional DDMs were the obvious starting point. To the best of our knowledge, no Bayesian formulations of these novel kinematics-dependent models currently exist. We do believe that such Bayesian versions of the kinematics-dependent DDMs could be formulated; novel methods like amortized Bayesian inference (Radev et al., 2020) substantially simplified Bayesian parameter estimation of complex models. However, before creating such versions and investigating how well their parameters could be recovered with small sample sizes and Bayesian estimation, we first need an indication of whether using Bayesian estimation methods and informative prior distributions to decrease sample size demands is a promising avenue of research, to begin with. The aim of this study is to provide this initial indication.

2.3. Using Bayesian estimation techniques to reduce data demands

An important advancement in the field of parameter estimation that may decrease EAMs' data demands is the development of Bayesian parameter estimation techniques (see e.g., Lee & Wagenmakers, 2013). The general premise of Bayesian estimation is that the researcher first assumes a prior distribution (often abbreviated as "the prior") of probable and possible parameter values (e.g., how high or low a drift rate usually is for a driver in a particular situation and how high or low it could possibly be in this situation). From this prior distribution of parameter values, sets of values for the parameters are iteratively sampled. After each round of sampling, the likelihood of data that has been observed is determined under a particular set of parameter values (e.g., how likely we are to observe a specific set of data, if the DDM's parameters had the values that were sampled this round). Samples of parameter values that give high likelihoods to the data (i.e., that apparently explain the observed data well) are weighted more strongly in future rounds of sampling, increasing the probability that these samples will be drawn again. Repeated sampling in this way yields a posterior distribution for each parameter (e.g., one for the drift rate and one for boundary separation), which – in the ideal scenario – describes the optimal parameter values.

Researchers differ in which prior distributions they pick for their parameters. Many choose distributions that are maximally uninformative (i.e., priors that influence the sampling procedure as little as possible), reasoning that then any information in the resulting posterior distribution stems solely from the data. Recently, however, researchers have argued in favor of including prior information in the parameter estimation process (Lee & Vanpaemel, 2018; Tran et al., 2021; Vanpaemel, 2011), for example when data from relevant comparable situations (or experiments) is available. If we, for example, have no prior knowledge on how people make decisions while drunk driving, we could still make our sampling of parameter values more efficient (i.e., decreasing the proportion of iterations in which unreasonable parameter values are tested) by informing our priors with data from analogous situations (e.g., knowledge on the kinds of parameter values that are probable for drivers that are highly distracted). Even if no data from analogous situations exist, general knowledge about a particular model can be utilized: For example, relevant to our project, Tran et al. (2021) generated prior distributions for DDM parameters that summarize all parameter values that have been published in papers about DDMs.

Utilizing such *informative priors*, the sampling for reasonable DDM parameters can be made much more efficient, by guiding the algorithm that samples parameter estimates towards reasonable parameter values (e.g., emphasizing values that were commonly observed for drift rates and heavily deemphasizing values that have never been observed). For our driving application this means that we might need much less data on how drivers make choices than is usually gathered for DDMs, if we have a lot of data from comparable drivers (e.g., in age and experience) for the same type of driving situation.

2.4. This study

While it is known that informative priors can be used to reduce the amount of required data, it is not clear how small *sample sizes* (i.e., numbers of recorded choices per person we wish to model) can be, if we want our estimates for the DDM parameters to still be reasonably accurate. We aim to answer this question through a simulation study (Section 3) and a reanalysis of experimental data (Section 4). The simulation study assesses how well DDM parameters can be recovered under ideal circumstances. Herewith we aim to identify the lowest possible sample sizes that would still enable reasonable parameter recovery. With the reanalysis, we assess parameter recovery under more realistic circumstances, where the data is noisier and it is unclear how much the informative priors that we chose resemble people's true parameter distributions.

While several studies have assessed the effects of sample sizes on DDM parameter recovery (e.g., Lerche et al., 2017; Ratcliff & Childers, 2015; White et al., 2018; Wiecki et al., 2013), our study is, to the best of our knowledge, the first to assess such small sample sizes and to assess the impact of prior choice on parameter recovery with such small sample size DDMs. As experimental trials are much more time-consuming for

(simulated) driving tasks (Pekkanen et al., 2022) than for laboratory tasks that are traditionally modeled with DDMs (e.g., people will not feel like they are really driving, if the simulation makes them encounter too many critical driving situation in short succession), assessing sample sizes as small as we assess here is important for the driving domain. This is especially true if we wish to model individual differences in cognitive parameters rather than pool data across participants.

3. Simulation study to identify the lowest possible sample sizes

With this simulation study we tested how small sample sizes could be in the ideal situation in which the prior distribution near-perfectly describes the real distribution of parameters. The general setup of the simulation study was as follows: We simulated data for 100 individuals, assuming that the individuals' DDM parameters are distributed according to the distributions reported in Tran et al. (2021). These distributions are believed to reflect the true data-generating distributions across multiple tasks (including driving tasks), as they summarize values that were reported across the scientific DDM literature.

Specifically, we randomly drew values for the boundary separation (a), drift rate (v), start point (z), nondecision time (t_0), start point variability (s_z) and drift rate variability (s_v) parameters separately for each simulated individual (see Table 1 and Fig. 2 for the priors from Tran et al., 2021). Based on these parameters, we simulated 200 trials of choices and reaction times (RTs) for each participant. Next, we estimated each person's parameters to understand whether we could reliably recover these data-generating parameters. Imitating a realistic user modeling scenario, where data of individuals are sequentially acquired, the parameter estimation was also performed independently (i.e., not making use of the hierarchical structure in the data). As we did in the reanalysis of data (Section 4), we fit all DDMs with Dynamic Models of Choice (DMC; Heathcote et al., 2019), a package for Bayesian estimation of EAMs in R.[2]

We vary two aspects of the simulation: the number of observations per individual (5, 10, 20, 30, 50, 100, or 200) that is used to fit the DDMs and the type of priors that are used for the parameter estimations. The first aspect is the central question of the paper, namely what can be concluded when only a limited number of observations is acquired. For reference, we also include the total number of observations that we analyzed per person in our later reanalysis of experimental data (100) as well as a common number of trials in a psychological choice experiments that is often believed to suffice to reliably recover DDM parameters (200 trials; see e.g., Lerche et al., 2017).

The second aspect addresses the role of prior shape in estimating parameters for small samples. Accurate informative priors ("empirical priors") reflect knowledge of the population from which the individual stems that is being evaluated. Since knowledge about a person's population is relevant for that person, empirical priors may assist in estimating parameters for small samples. However, if the individual's true parameters are dissimilar to the ones favored by the prior, informative priors might bias parameter estimates. To assess the risk of choosing a wrong informative prior, we contrasted three different priors: Accurate empirical priors (i.e., the priors from Tran et al., 2021, which resemble the individuals' parameters, as those parameters were initially sampled from the same distribution), wide uniform (and hence uninformative) priors, and informative prior distributions that are misaligned with the actual distributions of parameters. Our misaligned "mismatch" priors were identical to the empirical ones, except that we added 20% to their central tendency parameters (compare the blue with the black distributions in Fig. 2). This way, these priors are shifted by 20% relative to the population distribution with which the data was generated (i.e., the correct distribution of parameters).

Table 1
Parameters of the empirical prior distributions, adapted from Tran et al. (2021)

Parameter	Distribution	Mean		Scale	df	Lower & Upper bounds
a	gamma		11.69	0.12	0.11	7.47
v	truncated normal	1.76	1.51		0.01	8.51
z	truncated t	0.5	0.05	1.85	0.04	0.96
s_v	truncated normal	1.36	0.69		0	3.45
s_z	truncated normal	0.33	0.22		0.01	0.85
t_0	truncated t	0.44	0.08	1.32	0	3.69

Note. The parameters of the "mismatch prior" are identical to the ones of the here-reported "empirical prior" distribution, except that z's mean value is 0.75 and that all other mean values are doubled (e.g., the mean value for the mismatch prior of t_0 is 0.88). All uninformative prior distributions were uniform distributions with the lower and upper bound values depicted here.

3.1. Parameter recovery in the simulation study

As an initial inspection of our data, we plotted posterior means of the individual DDMs' parameter estimates against participants' true parameter values. Unsurprisingly, sample size heavily influenced the success of the parameter recovery. For example, Fig. 3 depicts the posterior means of the boundary separation parameter, for the lowest (left) and highest (right) sample size in the simulation and for all three prior distributions. Data points that are close to the dotted diagonal line were recovered well (points on the line indicate perfect recovery), which indicates that the cognitive processes that generated the data of the individuals could be identified (e.g., Miletić et al., 2017; Van Maanen & Miletić, 2021). It is clear from the figure that the parameters can be recovered when a large number of datapoints (i.e., 200) is collected. Similarly evident is that an extremely low sample size yields too little information to reliably recover the boundary separation parameter (see the left panel of the figure). However, when the distribution of individuals is known (i.e., when an accurate empirical prior is used; black dots), the estimated parameter values are at least close to the diagonal line (i.e., close to good recovery).

In our view more meaningful than estimating the exact parameter values of an individual (e.g., proximity to the line in Fig. 3) is the ordering of individuals along the dimension of interest. For example, if the goal of our modeling is to identify the level of cautiousness of drivers, what matters most is whether a cautious driver is indeed estimated to be more cautious than a non-cautious one (for a similar line of argumentation, see Ratcliff & Childers, 2015). This is why we assessed the extent to which parameter recovery was successful by computing Spearman's rank order correlation coefficients between the medians of a parameter's posterior distributions (each stemming from an individual participant's DDM) and the individuals' true parameter values. In graphing these correlation coefficients, Fig. 4 generalizes our conclusion from Fig. 3 to all parameters and sample sizes: The correlation between ground truth (i.e., people's correct parameter values) and people's median parameter estimates increases with sample size (i.e., recovery improves). However, Fig. 4 also reveals that specifying an appropriate prior distribution helps in the recovery, especially for the boundary separation parameter – a parameter important for driving as it indicates how much information (and hence time) human drivers require to come to a decision.

For the between-trial variability parameters (s_v and s_z), recovery is poor (i.e., even for DDMs with 30 observations, the correlation coefficients were $< .4$), independently of the prior shape. This finding is

[2] We modified the DMC package to support truncated t-distributions, as some of Tran et al.'s (2021) priors have that shape. To ensure reproducibility, our modified version of the DMC package (alongside our analysis scripts) are available on https://osf.io/9y6ze .

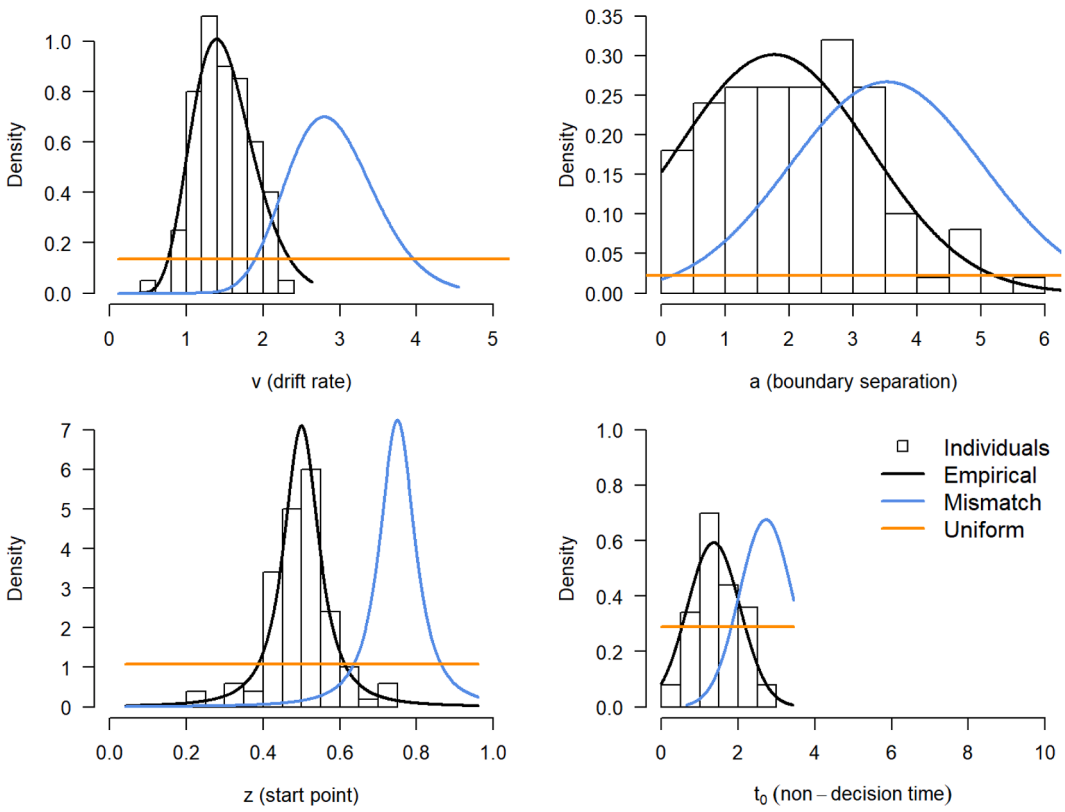

Fig. 2. Prior distributions for four of the parameters
Note. The colored lines show the different prior distributions that are used in our simulation study. The bars indicate the frequency with which the simulated individuals have parameters of that value. Note that the 100 simulated individuals were initially based on population distributions of parameters that are identical to the empirical prior distributions (i.e., while there is substantial individual variation, their overall parameter distributions closely resemble the empirical priors).

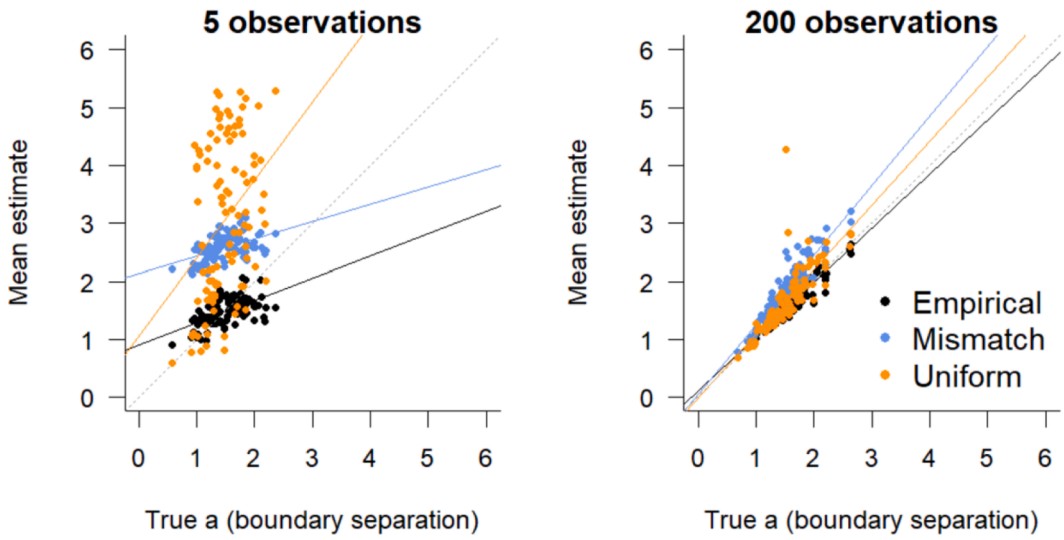

Fig. 3. Parameter recovery for the boundary separation parameter (a) with 5 and 200 observations
Note. Parameter recovery for the boundary separation parameter (a) was assessed using three different prior distributions (correct empirical prior, mismatched informative prior and uninformative uniform prior). Left shows recovery with small sample size of only 5 observations per participant; right shows a large sample size of 200 observations per participant.

unsurprising, given results from previous studies that found poor recovery of these parameters (e.g., Boehm et al., 2018), even for large sample sizes (e.g., correlations of $< .5$ for DDMs with 5000 trials; Lerche et al., 2017).

A sensible strategy in user modeling is to identify groups of individuals like "very cautious drivers" (e.g., so that one of several default options of a driver-assistance system can be chosen for a driver), rather than addressing all possible values on a continuum (e.g., of cautiousness). Therefore, we also analyzed the extent to which models based on few observations can faithfully recover which tertile of parameter values a simulated individual belongs to. To this end, we drew 50,000 samples from the posterior distributions of each parameter, and categorized the

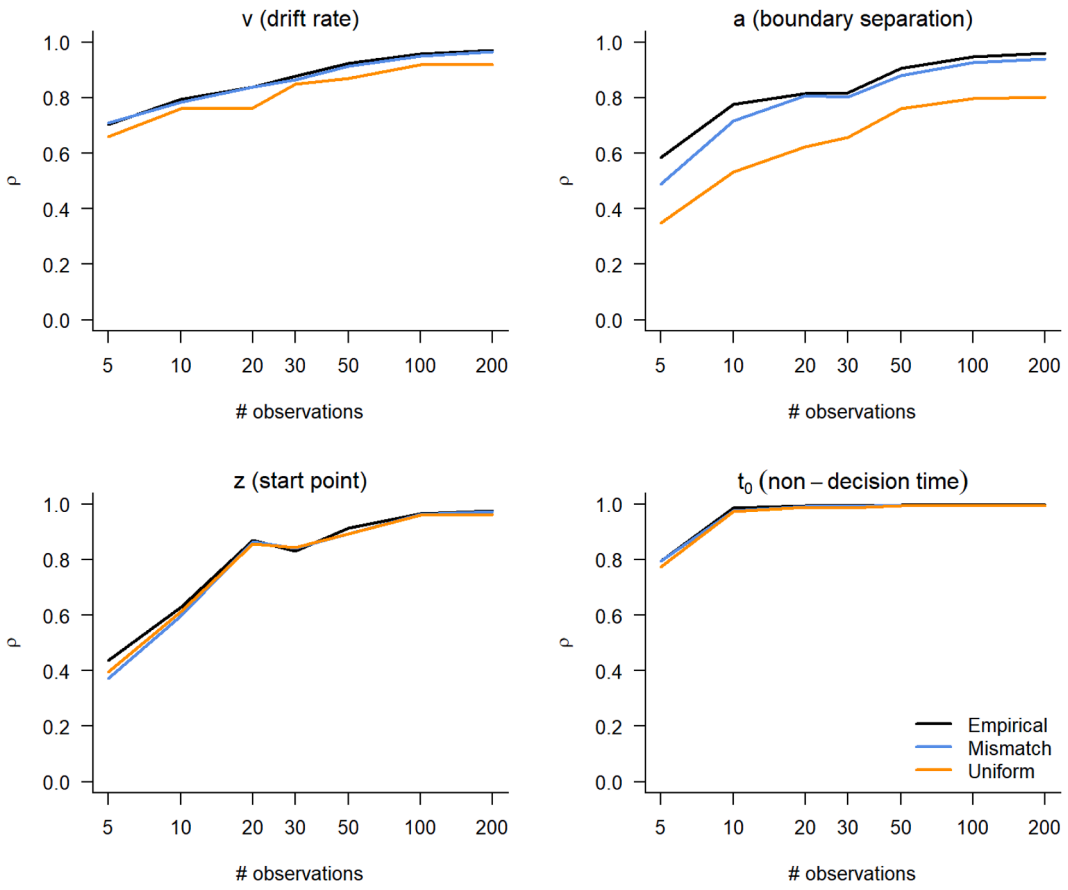

Fig. 4. Correlation coefficients between true and estimated DDM parameters
Note. Colors indicate different prior distributions (correct informative "empirical" prior, mismatched informative prior and uninformative uniform prior), y-axes depict the correlation coefficient and x-axes the number of observations (in log scale) used to fit the DDMs.

parameter values as either low, medium, or high. Next, we determined a confusion matrix from which we extracted the proportion of samples that the models (based on different sample sizes) classify correctly (compared to the ground truth distribution). Effectively, this analysis determines whether we can, with only limited observations, correctly categorize (simulated) individuals' values on a parameter as either low, medium, or high.

Our results revealed that specifying an appropriate prior distribution helps with the recovery of the drift rate and boundary separation parameters, in particular for models that are based on small sample sizes (Fig. 5). For example, the top left panel of Fig. 5 reveals that with as little as 5 observations, the correct classification of the drift rate (v) occurs on average in about .7 of cases; 95% confidence interval: [0.60, 0.79]. Similarly, the boundary separation parameter a (top middle panel) is correctly classified in .45 [CI: 0.35-0.55] of cases, and this accuracy increases to .60 [CI: 0.50-0.70] for models of 10 observations.

However, Fig. 5 also illustrates that it is of critical importance to specify an appropriate prior distribution: The Mismatch prior distribution of boundary separation, that is shifted relative to the population distribution from which individuals were sampled, yields a substantially lower classification accuracy, as compared to the other two prior conditions. This is mostly visible for a traditional sample size of 200 observations, where the 95% confidence intervals of the classifications do not even overlap across conditions. When the mismatch prior was used, the classification accuracy of the boundary separation (a) did not exceed chance performance for any of the small sample sizes that we considered. For the use case of categorizing participants' boundary separation parameters, it will hence be especially important to have an accurate prior distribution. While not to the same extent, the mismatch prior also underperforms the other priors in the classification of the drift rate.

To categorize other parameters, the nature of the prior distribution has little influence. For non-decision time (t_0) and start point (z) the recovery is reasonable for all priors. For the between-trial variability parameters (s_v and s_z), recovery was again poor and prior type did not impact recovery meaningfully.

3.2. Discussion of the simulation study

The results from the simulation study reveal that parameter recovery based on small sample sizes is, in principle, possible. Accurate informative priors improve the parameter recovery for all parameters and sample sizes and especially so for small numbers of observations. The influence of inaccurate informative priors compared to using uninformative priors was highly context-dependent: For the ordering of participants' parameter estimates (as indexed by rank-order correlations), the mismatch prior did not perform markedly worse than the uninformative prior (if anything, it outperformed the uninformative prior for some of the parameters, see e.g., drift rate and boundary separation in Fig. 4). For categorizing participants' parameters, however, the biasing influence of the mismatch prior could be observed clearly for the boundary separation and drift rate. There it seems to be particularly important for informative priors to be accurate. We note that these conclusions about incorrect informative priors might only hold for the type of mismatch we investigated here (i.e., a shift in the distributions' central tendency in the positive direction).

A severe limitation of our simulation study is that the conditions for the parameter recovery are unrealistically ideal. For example, in our simulation study each accurate informative prior near-perfectly

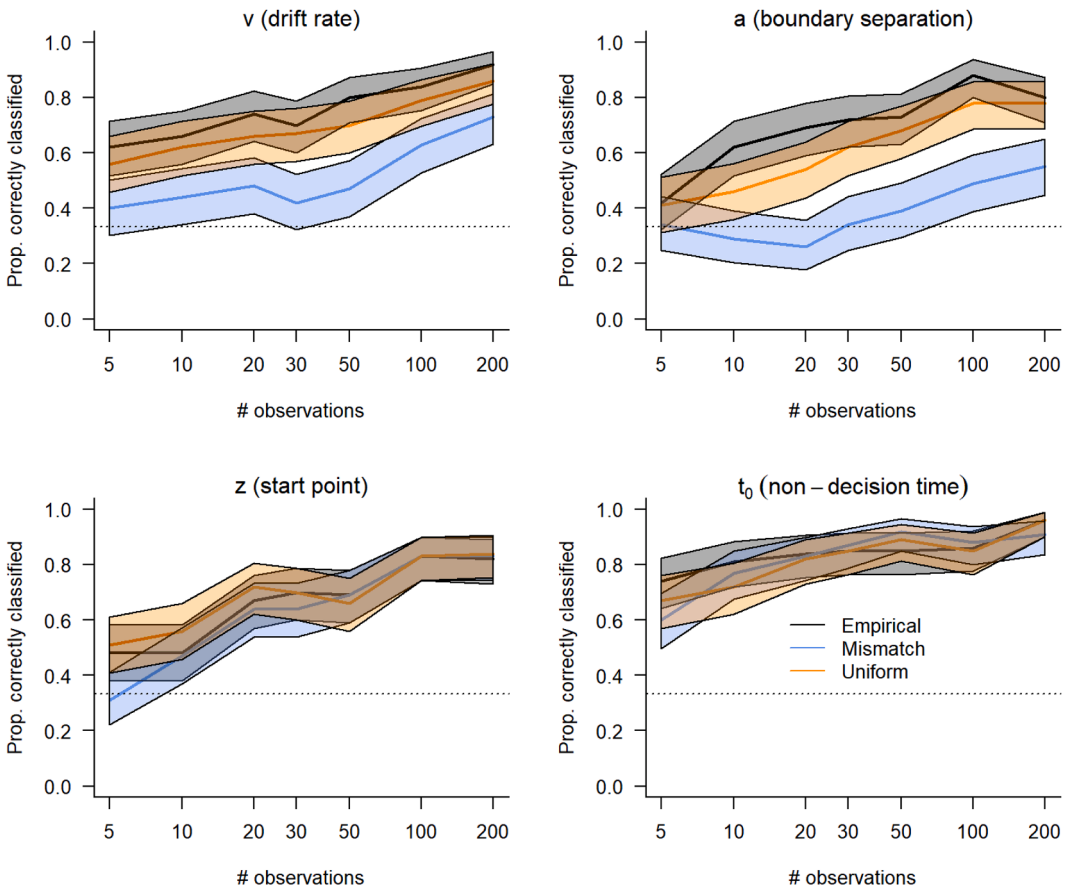

Fig. 5. Proportion of parameters that were correctly classified as either low, medium or high.
Note. Chance performance is indicated by the dotted line. 95%-confidence intervals are indicated by the shaded areas.

describes the distribution of participants' true parameter values (as their values were sampled from the same distribution) – a degree of precision we cannot expect in real-life applications. Additionally, variability in non-decision time (st_0) exists for real participants' data – the amount of nondecision time fluctuates trial-by-trial. This complicates the estimation of t_0 for real data, but not the simulation (where st_0 does not exist). Consequently, we reanalyzed experimental data to test whether the findings of our simulation study replicate under more realistic circumstances.

4. Data reanalysis to assess parameter recovery under realistic circumstances

The reanalysis of experimental data followed a similar structure to the simulation study with two notable differences: Firstly, we did not compare different prior distributions. As we cannot know to what extent our informative prior distributions are mismatched (i.e., differ from the true data-generating parameter distributions), a comparison between accurate informative, mismatched informative and uninformative distribution was not possible. Consequently, we merely assessed whether informative priors (which we hope to be accurate) led to good parameter recovery.

Secondly, unlike in the simulation study, we do not know participants' true parameter values. To have a benchmark by which we could evaluate the small sample size models' parameter recovery, we hence estimated parameter values based on all included trials with a Bayesian hierarchical model, a process we outline in Section 4.1.2.

4.1. Methods of the reanalysis

4.1.1. Dataset

The analyzed dataset stems from Ratcliff and Vanunu (2022) who made their data freely accessible on OSF. We chose this dataset, because its participants provided their responses in a simulated driving environment (although the task was perceptual) and because the dataset contained enough data to fit individual DDMs (i.e., at least 100 observations per participant).

In their experiment, Ratcliff and Vanunu (2022) tested whether and how adults of two different age groups (i.e., younger vs older adults) differed in their responses across three different tasks. The number of recorded observations differed per participant. We chose to model the data from Ratcliff and Vanunu's "two-choice clear task", as for that task participants had the highest minimum number of observations. In the "two-choice clear task" participants had to (via steering wheel, as well as gas and brake pedals) follow a simulated leading car on a computer screen until they were shown a color patch. Upon seeing the patch, participants had to overtake the leading car, either on the left or right, depending on whether the patch was blacker or whiter (e.g., consisted of 53% white or 53% black pixels).

Within the task, there were more or less ambiguous color patches (i.e., 53% or 57% of the pixels were of the dominant color) which influenced the difficulty of the trial. We chose to model the more ambiguous and hence harder trials (where the color patch was either 53% white and 47% black or the reverse), as we expected that participants would make more mistakes during this task. In addition, we limited our analysis to the group of younger adults. As the majority of participants in psychological experiments are young adults (e.g., psychology students), we assumed that the here-utilized Tran et al. (2021) priors are based on a

population that more closely resembles adults of Ratcliff and Vanunu's (2022) younger participant group (i.e., 19 to 29-year-olds) than their older group (i.e., 58 to 82-year-olds). As participants had inconsistent numbers of hard trials (ranging from 113 to 226 such trials), we randomly selected per participant a subset of 100 randomly-ordered hard trials. Initial inspection of the responses from the 30 young adult drivers revealed that two adults seemed to have acted randomly on the task or misunderstood it (i.e., on this task with two possible response options, they had accuracies of 27% and 46%, respectively). We excluded these participants from the analysis.

4.1.2. Establishing the ground truth

To obtain precise estimates for participants' individual parameters which can serve as performance benchmarks ("ground truths") for the small sample size DDMs, we fit a Bayesian hierarchical diffusion decision model (hDDM) on all included observations (i.e., 100 decisions per participant). Bayesian hierarchical models simultaneously estimate the parameters of all people whose data is included to fit the model. They do so by assuming that each individual level parameter (e.g., a person's v or t_0) stems from a distribution of parameters that is defined by a mean and a standard deviation. These mean and standard deviation *hyperparameters* are themselves modeled and estimated by the Bayesian hierarchical model with the help of prior distributions (i.e., so-called *hyperprior* distributions). We chose Tran et al.'s (2021) distributions as hyperprior distributions for the means and uninformative uniform priors as hyperpriors of the standard deviations. In estimating the parameters of a participant (e.g., the drift rate of the first participant), hierarchical models tend to be more accurate than nonhierarchical ones, because they do not only take into consideration data from the participant whose parameters they estimate, but additionally consider data from all other participants (Lee & Wagenmakers, 2013).

In fitting the "ground truth" model as in fitting any of the models in the following steps, we estimated four parameters: the drift rate (v), the response boundary separation (a), the start point (z) and the mean nondecision time (t_0). Initially, we also wanted to estimate the standard deviations of drift rate and start point (s_v and s_z) – as we had done in the simulation study. However, convergence of DDMs with these 6 parameters was too poor and too slow to be scalable (in light of the large number of models we wanted to fit in the following steps). With the Bayesian hierarchical DDM, we adjusted Tran et al.'s (2021) priors to arrive at 1000 converged chains of the posterior per parameter of the 28 participants. These chains serve as our approximations of the "true" posterior distribution of DDM parameters for each participant.

4.1.3. Fitting the models

To be able to fit individual DDMs (independently, as we had done in the simulation study), we first had to create appropriate informative priors; priors that contain information about a modeled participant's population without containing information about the participant themself. To that end, we fit a Bayesian hierarchical DDM (hDDM) for each participant in the same way as we had done to estimate the "ground truth" parameters (Section 4.1.2). Now we however fit these models without observations from the person whose driving attributes we wanted to subsequently model.

We obtained the means, standard deviations, minima and maxima across all samples of all hierarchical parameter estimates of the remaining participants to parameterize these priors; using them to define truncated normal (prior) distributions (after confirming with histograms and QQ-plots that the estimates from the samples were indeed approximately normally distributed). For example, to obtain a prior for the drift rate of the first participant, we fit a hDDM based on the experimental data of all participants but the first. We then obtained the mean, standard deviation, minimum and maximum of the 27×800 samples that the hDDM drew for the drift rates of these 27 participants. These four values defined the truncated normal distribution that we used as our informative prior for the drift rate of the first participant.

For the following reason, all our estimates about participants' parameters are based on the last 800 samples that our model drew: For each DDM that we fit (be they hierarchical or nonhierarchical), we sampled until the model converged[3] (i.e., until the convergence statistic Gelman-Rubin $\widehat{R}$ was below 1.05). As for some participants the model needed to sample additional times before converging, the number of drawn samples differed between participants. Only considering participants' last 800 samples to draw conclusions about parameters (e.g., to describe a prior distribution for drift rates) ensured that we considered the same number of samples per participant and that those samples stemmed from converged models.

After obtaining prior distributions for each participant's parameters, we used them to fit (nonhierarchical) DDMs for each of the 28 participants. We varied the number of observations from the modeled participant, fitting DDMs with 5, 10 or 20 of their observations. Whenever we fit a model, we randomly selected which of a participant's 100 observations to include. This random selection of trials and fitting of DDMs was repeated across 100 random seeds (see Section 4.2) to ensure that our conclusions are robust. We evaluated the individual models by the extent to which their estimated parameters were consistent with the ones that the ("ground truth") hDDM estimated.

4.2. Parameter recovery in the reanalysis

To evaluate the precision of a DDM's parameter estimate, we computed the Spearman correlation between the median of its posterior samples and the median of the ground truth samples for that participant's parameter. For example, to evaluate how well the boundary separation is estimated based on 5 observations, we obtained 28 medians – the medians of the last 800 samples for the boundary separation of the 28 participants' 5-observation DDMs. We correlated these medians with 28 medians from the ("ground truth") hDDM's last 800 samples for each of the participants' boundary separations (as the hDDM has samples for each parameter of each participant). A high positive correlation indicates good parameter recovery (i.e., high agreement between the DDMs' assessment of the participants' attributes and the "ground truth"). We tested whether these correlations are reliable by fitting small sample DDMs and correlating 100 times across different random seeds (thereby varying which observations are included, when fitting the models).

To test whether the few (5, 10 or 20) included observations meaningfully updated the priors towards the ground truth, we also checked what correlation coefficients are obtained when one randomly samples from the informative prior distributions. Specifically, we drew a parameter estimate per participant from their (truncated normal) prior distribution. We then correlated these 28 estimates with the ground truth estimates for that parameter. This process was repeated across 1000 random seeds. If the correlation coefficients from the small sample DDMs are not larger than the 95^{th} percentile of correlation coefficients from these random draws, the few observations from a driver did not meaningfully update the prior distribution towards the ground truth of that driver.

Results revealed that nondecision time (t_0) was not modeled well with any of the models that were based on small samples. Even for the 20-observation models did the mean correlation (.2993) only barely exceed the 95^{th} percentile of the control correlations based on the unchanged priors (.2737). Presumably, a larger number of observations is required to adequately estimate nondecision time.

The other three parameters could be estimated relatively well with

[3] *Convergence* means that the model consistently draws samples (of parameter values) from the same distribution, which likely is the (posterior) distribution of the correct parameter values. Interested readers can find more details on sampling for example in Lee and Wagenmakers's (2013) book on Bayesian estimation.

small sample size models. On average the 20 observation DDMs' estimates were moderately highly correlated with the estimates of the ground truth, $M_{r\_v} = .5536$, $M_{r\_z} = .6047$. The average recovery from 10 observation DDMs' estimates ($M_{r\_v} = .3605$, $M_{r\_z} = .4637$) was also better than the 95th percentile of random correlations ($r_{v\_95} = .2677$, $r_{z\_95} = .2835$), but given the high frequency of low correlations (see orange distributions in Fig. 6), we believe that 20 observations DDMs are preferable in practice.

With small sample sizes, the boundary separation could be estimated much better than the other parameters. Remarkably, estimates for that parameter already outperformed the control correlations for models that were based on only 5 observations, $r_{minimum\_a} = .3305$, $M_{r\_a\_5} = .6056$ compared to the random $r_{a\_95} = .2978$. For the 20 observations DDMs, this mean correlation rose to $M_{r\_a\_20} = .7300$.

Overall, with the exception of nondecision time, all parameters could be estimated well with 20 observations. That is markedly fewer than are usually obtained in psychological experiments and opens the door to applying EAMs for user modeling. For a more detailed breakdown of the distributions of observed correlations across models and random seeds, see Fig. 6. The figure demonstrates that parameters (with the exception of t_0) are estimated well with only 20 observations: The distributions of observed correlations based on 20 observation models (blue lines) have little overlap with the distributions of correlations that are observed randomly (i.e., the correlations between "ground truth" and random draws from the priors; black lines).

5. Discussion

For (semi-)autonomous driving systems to be able to coexist with human road users, they need a decent model for how humans make decisions. If they mis-anticipate a person's behavior, they could cause an accident or at least fail to prevent one. Similarly, a system that shares driving duties with a human driver needs an accurate model of how that person makes decisions on the road; what their strengths, weaknesses, and proclivities are (e.g., when in doubt, whether or not they err on the side of braking). Even in a future in which humans leave all driving to AI, driving systems will need to anticipate human traffic decision making (e.g., when an AV interacts with pedestrians at a crossing; Pekkanen et al., 2022).

Cognitive models seem like the perfect source for insights about such human decision-making. After all, these models were designed to explain how humans make decisions and how they perceive and analyze their surroundings. Additionally, cognitive models can predict how human drivers will behave under circumstances that are better simulated than tested. For example, evidence across several experimental paradigms indicates that drift rates decrease with less perceptual discriminability of task-relevant stimuli (see e.g., Donkin & Van Maanen, 2014; Palmer et al., 2005; Philiastides et al., 2006). Instead of gathering driving data from dangerous situations of low perceptual discriminability (e.g., when it is too foggy or snowy for human drivers to fully perceive all relevant information), driving systems could be informed with EAMs' predictions about what would change if the driver had a lower drift rate. Despite all this conceptual promise, several roadblocks present themselves to applying cognitive models to the driving domain, and cognitive models have not yet realized their potential for informing the design of AVs (Janssen et al., 2022).

One of the roadblocks of applying specifically evidence accumulation models in practice is that they require a lot of data. After all, we cannot expect even highly motivated drivers – which want their AI co-drivers to be well-informed about their driving tendencies – to sit through hours of driving simulations per type of high-stakes driving decision (e.g., braking, overtaking, etc.). This requirement becomes

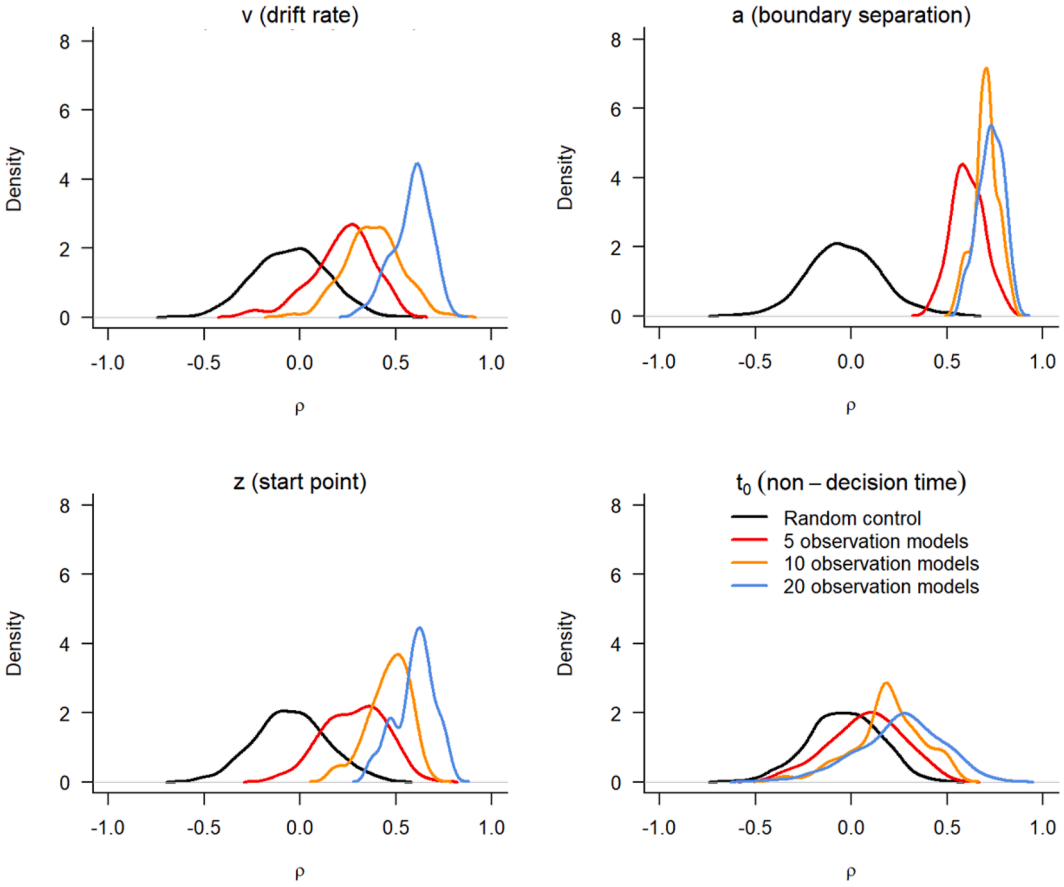

Fig. 6. Density distributions of correlation coefficients between ground truth and small sample size models' parameter estimates across random seeds

even more daunting, if we assume that assessments have to be repeated across conditions (e.g., to test drivers' behavior when stressed or sleepy) and time (e.g., due to age-related changes in the driver's cognitive processing; Ratcliff & Vanunu, 2022). Asking drivers to partake in such assessments of their driving attributes would be a lot more reasonable, if they took less time. In this paper, we consequently made a first attempt towards finding out how small sample sizes could be, if we want our model's parameter estimates to still be reasonably accurate.

5.1. Takeaways from our studies

The key takeaway of our simulation study is that little data is necessary to receive reasonable parameter estimates, if the prior distribution closely resembles the true distribution to which an individual belongs. In traffic situations in which this type of DDM is appropriate (e.g., split-second decisions, like whether or not to brake in response to suddenly perceiving a car one had previously overlooked), we can likely draw valid inferences about a driver's behavior with as few as 10 data points – if we have data on how other drivers that cognitively resemble our driver acted in such situations. More observations were required when an uninformative prior was used or when the informative prior distribution for a parameter was mismatched. This implies that obtaining accurate informative prior distributions will be very important for the applied setting, in which few observations can be obtained per driver.

Overall, the mismatched informative prior performed similarly to the uninformative one, in the recovery of parameters. At least for testing individual differences, informative priors (when mismatched as they were here) hence appear to be relatively save to choose over uninformative ones.[4] The situation is different for classifications: As an initial classification of the prior distributions determines which values are labeled as low, medium, or high, informative prior distributions that are as misaligned as ours lead to many misclassifications. For such applications of cognitive models, it will hence be very important to obtain accurate prior distributions, as mismatched priors can substantially bias our classifications.

Following our simulation study, we reanalyzed data from a perceptual task to test the effect of sample sizes on parameter recovery under more realistic situations. Similarly to our findings in the simulation study, we found that (with the help of informative prior distributions) most model parameters could be estimated well with few observations. We did however find notable differences between model parameters: While the boundary separation could be estimated with as few as 5 observations, drift rate and bias required 20, and nondecision time could not be estimated to a satisfactory level even with 20 observations. The latter finding stands in contrast to the results of our simulation study, where boundary separation was recovered well already with 10 observations (see Fig. 4). We suspect that the discrepancy between these results arises from the realistic (reanalyzed) data being noisier than the simulated data. Our DDM did not model variability in nondecision time (st_0). While no such variability exists in our simulated data, it does exist in real data and (as we do not model it) complicates recovery of the nondecision time constant (t_0). Nondecision time variability (st_0) might be especially high in the dataset that we reanalyzed, considering that participants had to multi-task (i.e., focus on their driving, while performing the perceptual task) and were not under immediate time pressure. As we only reanalyzed one dataset, it could also be the case that our prior distribution for the nondecision time was much more mismatched than the other priors (while the prior for t_0 was very accurate in the simulation study).

5.2. Future directions

There are several ways in which our work can be extended: First and foremost, future studies should investigate whether the observed differences in how well parameters recover (with small sample size DDMs) replicate or instead represent artifacts stemming from the data we analyzed. Such replications would ideally involve critical (simulated) driving situations (e.g., the sudden appearance of a car in one's lane or on the opposite side of the road) to ensure that the idiosyncrasies of driving situations (e.g., having to focus on keeping a lane and an appropriate speed, while observing the road) are taken into account, when estimating the minimum sample sizes. Simultaneously, such studies can also provide more on-task informative parameter distributions. In our study, the prior distributions were determined based on a wide array of decision-making paradigms (Tran et al., 2021), yielding relatively wide priors. More specific priors could improve precision of the estimates, decreasing the sample size requirements even further.

If differential trends for parameters prove reliable, the driving situation we want to model could inform the amount of data that is collected. For example, in some situations the choice of action is very obvious and drivers will immediately understand how they should act. Accurately estimating the nondecision time will be of crucial importance in such situations, as individual differences in motor response speed (i.e., how fast people can physically react) primarily determine the situations' outcomes. If t_0 continues to prove hard to estimate with small sample sizes, we would need to collect many observations for such driving situations. Conversely, if boundary separation's ease of recovery proves reliable, driving systems could be developed that actively monitor changes in a driver's cognitive states (e.g., to nudge the driver when they are prone to fast errors).

A second crucial extension of our work will be to assess the extent to which the parameters of kinematics-dependent DDMs can be recovered with small sample sizes. Many critical driving situations develop dynamically over multiple seconds (e.g., another car we were already aware of might suddenly force us to adjust our behavior by acting in ways we did not anticipate; Engström et al., 2022) – situations that the simpler DDM parameterizations we assessed here cannot account for. Kinematics-dependent models contain parameters that react to changes in observed stimuli (e.g., changes in the speed of another vehicle) and therefore hold substantial conceptual promise for modeling traffic situations that unfold over multiple seconds (see e.g., Giles et al., 2019; Zgonnikov et al., 2022; 2023). However, even for larger sample sizes (see e.g., Evans et al., 2020; White et al., 2018), successful parameter recovery – which is important to demonstrate that models are identifiable and specific (see e.g., Van der Velde et al., 2022; Van Maanen et al., 2021; Wilson & Collins, 2019) – can be tricky for EAMs with time-varying parameters (e.g., requiring additional model constraint; Evans et al., 2020). Consequently, the parameter recovery of more novel kinematics-dependent models should be established, before assessing the extent to which their parameters can be recovered with small sample sizes.

Additionally, Bayesian formulations of these new models will need to be developed, before we can hope to decrease their sample size demands with informative prior distributions. We believe that such Bayesian formulations hold substantial promise (and that they are achievable in light of successful Bayesian parameter recoveries for other time-varying DDMs, like those with collapsing boundaries; Evans et al., 2020). Pekkanen et al. (2022) point out that reaching a large number of experimental trials is hard for (simulated) driving studies, as trials often take 30 or more seconds. Additionally, we believe that having many trials occur in relatively short succession – compared to how much time passes between identical (critical) driving decisions in real life – is undesirable, as drivers will anticipate trials and train the (artificial) driving decision instead of (re-)acting naturally. Here, we believe, Bayesian parameter

[4] Our mismatched prior distributions were relatively starkly misaligned with the actual distribution of parameter values (Fig. 2); in most applied cases the misalignment will be smaller. That being said, we only assessed one type of misalignment and others might be more biasing compared to uninformative priors.

estimation techniques could play a crucial role, as they could decrease the sample size demands for modelling individual drivers' parameters. We note that Bayesian formulations of the kinematics-dependent models would suffice; Bayesian hierarchical models (while, in our view, likely leading to better priors) are not required to obtain informative prior distributions that improve parameter estimation.

Potentially interesting to assess could also be the effect of experimental conditions on parameter recovery with small sample sizes. For example, in the perceptual study from Ratcliff and Vanunu (2022) that we reanalyzed, participants had to react to more or less ambiguous stimuli with the type of stimuli likely influencing the drift rate. We chose against modeling these within-participant differences in stimulus ambiguity, as we believe that the influence of different decision conditions (if different decisions, in practice, can even be unambiguously assigned to one condition or another) on model parameters would be difficult to pinpoint, in practice. If the effect of different conditions could be defined clearly for a relevant traffic situation, assessing the impact of conditions on (small sample size) parameter estimation would be interesting, though: On the one hand, conditions could aid parameter recovery, as they introduce additional model constraint and make the model easier to identify. On the other hand (all else being equal), the lower sample size per condition (e.g., splitting 20 observations into 2 conditions of 10 observations, each) could hurt recovery, especially since our sample sizes are so small. Thus, more research is needed to address this potential trade-off.

5.3. Sensitivity of parameter estimates

Finally, it would be interesting to assess how well small sample size DDMs recover participants' parameter values in absolute terms (e.g., to assess the width of the credibility interval, or the amount of error that is introduced by having too few observations). In this paper, we instead focused on the successful recovery of individual differences relative to each other (in the simulation study and reanalysis) and the recovery of classifications of participants as low, medium or high on a parameter (in the simulation study). In our view, these relative differences are more important for the design of driving systems: Categories of different default options (e.g., an AV's standard reaction to a person with a relatively large boundary separation or a co-driver system's default ways of supporting drivers with such attributes) will be easier to design and quality control than AV behaviours that change continuously along multiple dimensions (e.g., along different EAM parameters). That being said, accurate parameter recovery will be important, if the exact prediction of an individual's reaction time and choice is most important.

Bayesian hierarchical modelling interacts with these evaluation metrics in different ways: Shrinkage (e.g., that each individual's drift rate estimate within a hierarchical model is adjusted towards the mean drift rate, with larger adjustments, the further the estimate is from the mean) affects absolute values more than relative ones. Consider, for example, a driver that made decisions so cautiously that they received the highest boundary separation estimate among the assessed drivers. Shrinkage will not impact their rank much, as it similarly affects other drivers with extreme boundary separation estimates. It will also not impact their classification, as they will be classified as "high on boundary separation", in either case. It will, however, heavily impact their exact parameter estimate for the boundary separation, as the estimate is substantially shrunk towards the group mean. In most cases, shrinkage will make the driver's parameter estimate more accurate (and is, in our view, desired) – extreme values, especially if they are based on few observations, are more likely to be outliers than fully representative of the driver. However, in cases where the driver is part of a different population than the other drivers (e.g., a sole beginner amongst a group of experienced drivers) adjustments to this driver's values can be highly biasing. If we have reasons to believe that a particular driver is highly dissimilar to other drivers, it might thus be wise to assess their parameter values with non-hierarchical models. Here, too, vague but informative priors (e.g., the ones from Tran et al., 2021) can be used to aid parameter estimation.

5.4. Concluding remarks

Conceptually, evidence accumulation models hold a lot of promise for modelling traffic situations. After all, they model choices and the time required to make them, express choice-relevant individual differences (e.g., in how much evidence people need to commit to a choice) and can be used to predict changes in decision making, based on temporary states (e.g., being sleep-deprived; Ratcliff & Van Dongen, 2009). Several other traffic-relevant applications (e.g., the live-monitoring of a driver's cognitive state, or easier assessments of individual differences with kinematics-dependent DDMs) would be possible, if EAMs could be fit with less data. Here, we present a step towards decreasing EAMs' sample size demands. We hope that future such works will open the doors to many exciting applications of EAMs to the domain of (semi-) autonomous driving.

CRediT authorship contribution statement

Dominik Bachmann: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Leendert van Maanen:** Conceptualization, Methodology, Software, Formal analysis, Writing – review & editing, Visualization, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We are reanalyzing data that other authors made accessible on OSF. Additionally, we simulate data (which we provide access to).

Acknowledgements

This publication is part of the project "The biased reality of online media - Using stereotypes to make media manipulation visible" (with project number 406.DI.19.059) of the research programme Open Competition Digitalisation-SSH, which is financed by the Dutch Research Council (NWO). Additionally, we want to thank our two anonymous peer reviewers for their thought- and insightful feedback.

References

Archambeau, K., Forstmann, B., Van Maanen, L., Gevers, W., 2020. Proactive interference in aging: A model-based study. Psychonomic Bulletin & Review 27, 130–138.

Bianchi Piccinini, G., Lehtonen, E., Forcolin, F., Engström, J., Albers, D., Markkula, G., Lodin, J., Sandin, J., 2020. How do drivers respond to silent automation failures? Driving simulator study and comparison of computational driver braking models. Human factors 62 (7), 1212–1229.

Boag, R.J., Strickland, L., Heathcote, A., Neal, A., Palada, H., Loft, S., 2022. Evidence accumulation modelling in the wild: understanding safety-critical decisions. Trends in Cognitive Sciences.

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., 2018. Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. Journal of Mathematical Psychology 87, 46–75.

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.

Donkin, C., Van Maanen, L., 2014. Piéron's Law is not just an artifact of the response mechanism. Journal of Mathematical Psychology 62, 22–32.

Durrani, U., Lee, C., 2023. Applying the Accumulator model to predict driver's reaction time based on looming in approaching and braking conditions. Journal of Safety Research 86, 298–310.
Engström, J., Liu, S. Y., Dinparastdjadid, A., & Simoiu, C. (2022). Modeling road user response timing in naturalistic settings: a surprise-based framework. arXiv preprint arXiv:2208.08651v2.
Engström, J., Markkula, G., Xue, Q., Merat, N., 2018. Simulating the effect of cognitive load on braking responses in lead vehicle braking scenarios. IET Intelligent Transport Systems 12 (6), 427–433.
Evans, N.J., Trueblood, J.S., Holmes, W.R., 2020. A parameter recovery assessment of time-variant models of decision-making. Behavior Research Methods 52, 193–206.
Giles, O., Markkula, G., Pekkanen, J., Yokota, N., Matsunaga, N., Merat, N., Daimon, T., 2019. At the zebra crossing: Modelling complex decision processes with variable-drift diffusion models. In: Proceedings of the 41st annual meeting of the cognitive science society, pp. 366–372. Cognitive Science Society.
Heathcote, A., Lin, Y.S., Reynolds, A., Strickland, L., Gretton, M., Matzke, D., 2019. Dynamic models of choice. Behavior Research Methods 51, 961–985.
Janssen, C.P., Baumann, M., Oulasvirta, A., Iqbal, S.T., Heinrich, L., 2022. Computational Models of Human-Automated Vehicle Interaction (Dagstuhl Seminar 22102). Dagstuhl Reports 12 (3). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
Lee, M.D., Vanpaemel, W., 2018. Determining informative priors for cognitive models. Psychonomic Bulletin & Review 25, 114–127.
Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian modeling for cognitive science: A practical course*. Cambridge UP.
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, 513–537.
Markkula, G., Romano, R., Madigan, R., Fox, C.W., Giles, O.T., Merat, N., 2018. Models of human decision-making as tools for estimating and optimizing impacts of vehicle automation. Transportation Research Record 2672 (37), 153–163.
Miletić, S., Turner, B.M., Forstmann, B.U., Van Maanen, L., 2017. Parameter recovery for the Leaky Competitive Accumulator model. Journal of Mathematical Psychology 76, 25–50.
Noy, I.Y., Shinar, D., Horrey, W.J., 2018. Automated driving: Safety blind spots. Safety Science 102, 68–78.
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, 376–404.
Pekkanen, J., Giles, O.T., Lee, Y.M., Madigan, R., Daimon, T., Merat, N., Markkula, G., 2022. Variable-drift diffusion models of pedestrian road-crossing decisions. Computational Brain & Behavior 5, 60–80.
Philiastides, M.G., Ratcliff, R., Sajda, P., 2006. Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. Journal of Neuroscience 26 (35), 8965–8975.
Radev, S.T., Mertens, U.K., Voss, A., Ardizzone, L., Köthe, U., 2020. BayesFlow: Learning complex stochastic models with invertible neural networks. IEEE Transactions on Neural Networks and Learning Systems 33 (4), 1452–1466.
Ratcliff, R., 1978. A Theory of Memory Retrieval. Psychological Review 85, 59–108.
Ratcliff, R., 2015. Modeling one-choice and two-choice driving tasks. Attention, Perception, & Psychophysics 77 (6), 2134–2144.
Ratcliff, R., Childers, R., 2015. Individual differences and fitting methods for the two-choice diffusion model of decision making. Decision 2 (4), 237–279.
Ratcliff, R., McKoon, G., 2008. The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. Neural Computation 20 (4), 873–922.
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.
Ratcliff, R., Strayer, D., 2014. Modeling simple driving tasks with a one-boundary diffusion model. Psychonomic Bulletin & Review 21, 577–589.
Ratcliff, R., Van Dongen, H.P.A., 2009. Sleep deprivation affects multiple distinct cognitive processes. Psychonomic Bulletin & Review 16 (4), 742–751.
Ratcliff, R., Vanunu, Y., 2022. The effect of aging on decision-making while driving: A diffusion model analysis. Psychology and Aging 37 (4), 441.
SAE International, 2014. J3016: Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. SAE International, Warrendale, PA, USA.
Svärd, M., Markkula, G., Bärgman, J., Victor, T., 2021. Computational modeling of driver pre-crash brake response, with and without off-road glances: Parameterization using real-world crashes and near-crashes. Accident Analysis & Prevention 163, 106433.
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.
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, 608287.
Van der Velde, M., Sense, F., Borst, J.P., van Maanen, L., Van Rijn, H., 2022. Capturing Dynamic Performance in a Cognitive Model: Estimating ACT-R Memory Parameters With the Linear Ballistic Accumulator. Topics in Cognitive Science 14 (4), 889–903.
Van Maanen, L., Grasman, R.P., Forstmann, B.U., Keuken, M.C., Brown, S.D., Wagenmakers, E.J., 2012. Similarity and number of alternatives in the random-dot motion paradigm. Attention, Perception, & Psychophysics 74, 739–753.
Van Maanen, L., Heiden, R.V.D., Bootsma, S., Janssen, C.P., 2021. Identifiability and Specificity of the Two-Point Visual Control Model of Steering. In: Proceedings of the Annual Meeting of the Cognitive Science Society, 43.
Van Maanen, L., Miletić, S., 2021. The interpretation of behavior-model correlations in unidentified cognitive models. Psychonomic Bulletin & Review 28, 374–383.
Van Ravenzwaaij, D., Dutilh, G., Wagenmakers, E.J., 2012. A diffusion model decomposition of the effects of alcohol on perceptual decision making. Psychopharmacology 219, 1017–1025.
Vanpaemel, W., 2011. Constructing informative model priors using hierarchical methods. Journal of Mathematical Psychology 55 (1), 106–117.
Vanunu, Y., Ratcliff, R., 2022. The effect of speed-stress on driving behavior: A diffusion model analysis. Psychonomic Bulletin & Review 1–10.
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, 286–301.
Wiecki, T.V., Sofer, I., Frank, M.J., 2013. HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. Frontiers in Neuroinformatics 7, 14.
Wilson, R.C., Collins, A.G., 2019. Ten simple rules for the computational modeling of behavioral data. Elife 8, e49547.
Xue, Q., Markkula, G., Yan, X., Merat, N., 2018. Using perceptual cues for brake response to a lead vehicle: Comparing threshold and accumulator models of visual looming. Accident Analysis & Prevention 118, 114–124.
Zgonnikov, A., Abbink, D., Markkula, G., 2022. Should I stay or should I go? Cognitive modeling of left-turn gap acceptance decisions in human drivers. Human Factors, 00187208221144561.
Zgonnikov, A., Beckers, N., George, A., Abbink, D., & Jonker, C. (2023). Subtle motion cues by automated vehicles can nudge human drivers' decisions: Empirical evidence and computational cognitive model. [Preprint] Retrieved from https://osf.io/3cu8b/.

Dominik Bachmann is a PhD student at the Institute for Logic, Language and Computation (ILLC) at the University of Amsterdam. He is also guest researcher at the Department of Experimental Psychology at Utrecht University

Leendert van Maanen is Associate Professor Cognitive Aspects of Artificial Intelligence at the Department of Experimental Psychology at Utrecht University. He completed his PhD in 2009 at the University of Groningen.