# Jumping to Conclusion?
# A Lévy Flight Model of Decision Making

Eva Marie Wieschen [a] ✉ ⓘ , Andreas Voss [a] ⓘ & Stefan Radev [a] ⓘ

[a]Heidelberg University

**Abstract** ■ The diffusion model is one of the most prominent response time models in cognitive psychology. The model describes evidence accumulation as a stochastic process that runs between two boundaries until a threshold is hit, and a decision is made. The model assumes that information accumulation follows a Wiener diffusion process with normally distributed noise. However, the model's assumption of Gaussian noise might not be the optimal description of decision making. We argue that Lévy flights, incorporating more heavy-tailed, non-Gaussian noise, might provide a more accurate description of actual decision processes. In contrast to diffusion processes, Lévy flights are characterized by larger jumps in the decision process. To further examine this proposal, we compare the fit of the basic diffusion model and the full diffusion model (including inter-trial variability of starting-point, drift rate and non-decisional processes) to the fit of a simple and a complex version of a Lévy flight model. In the latter model, the heavy-tailedness of noise distributions was estimated by an additional free stability parameter alpha. Participants completed 500 trials of a color discrimination task and 400 trials of a lexical decision task. Results indicate that a complex Lévy flight model – including inter-trial variability parameters and alpha – shows the best fit in both tasks. Importantly, alpha-values correlated across tasks, indicating a trait-like nature of this parameter.

**Keywords** ■ Lévy flights, diffusion model, decision making.

## Introduction

Speed response time tasks are one of the most frequently used tasks in cognitive psychology. The diffusion model provides a valuable method to analyze response time data and its advantages have become increasingly clear during the last decades (Wagenmakers, 2009). The model describes evidence accumulation as a Wiener diffusion process that runs in a corridor between two thresholds, representing two response alternatives. In addition to a systematic drift in the decision process, normally distributed noise is assumed (Ratcliff, 1978). Although the diffusion model shows a reasonably good fit to data from a wide variety of cognitive tasks and its validity has been demonstrated across different paradigms (Ratcliff, 2002; Ratcliff & Rouder, 1998; Ratcliff, Thapar, & McKoon, 2001, 2003), recent studies suggest that models with more heavy-tailed noise distributions might be superior to classical diffusion models (Voss, Lerche, Mertens, & Voss, 2019). Stochastic processes assuming heavy-tailed noise are called Lévy flights. In these Lévy flights, the probability of extreme events is strongly increased, which results in jumps in evidence accumulation. Lévy flight models have been applied to a variety of contexts in different fields of science, including animal foraging (e.g., Reynolds, 2012), economic processes (e.g., Mantegna, 1991), as well as human perception and cognition (Liberati et al., 2017; Montez, Thompson, & Kello, 2015; Rhodes & Turvey, 2007).

The present study aims at providing further evidence for the applicability of Lévy flights to human decision making. Similar to the approach of Voss et al. (2019), we compare different evidence accumulation models with different noise distributions for the decision process. However, in contrast to the study of Voss et al. (2019), the present study is based on data with notably larger trial numbers per participant, and, importantly, we now employ exper-

imental paradigms that have been often used in the field of diffusion modeling. Furthermore, we now apply a fully Bayesian parameter estimation method. The Bayesian information criterion (BIC) that penalizes model complexity is computed to allow for a rigorous model comparison. Additionally, we address the question of whether there is a psychologically meaningful inter-individual variance in the stability parameter alpha, which maps the heavy-tailedness of the noise distribution.

### The Diffusion Model

Speeded response time tasks are a common type of paradigm in cognitive psychology. Typically, in two-alternative forced choice (2AFC) tasks, either mean response time or accuracy of responses is used as a measure of performance. However, such separate analyses entail the problem that a common metric for performance is missing, which is especially problematic when it comes to speed-accuracy trade-offs between experimental conditions. In addition, these traditional analyses utilize only a small amount of the available information, since information from many trials of an experimental condition is summarized by one single number, such as the mean response time.

The diffusion model (Laming, 1968; Ratcliff, 1978) addresses these problems. Due to its advantages over traditional analytic strategies for RT data, the model has remarkably grown in popularity over the last decades (Voss, Nagler, & Lerche, 2013). The model provides a theoretical account for the composition of response time distributions in binary decision making, considering location and shape of response time distributions for correct responses and errors and accuracy of responses. According to the model, during a binary decision, information is accumulated continuously, and this evidence accumulation is described by a Wiener diffusion process. This process comprises a systematic component, the so-called drift rate, and Gaussian noise. Whereas the drift rate determines average speed and direction of information accumulation, the random noise is responsible for variance of response times and response outcomes when the same stimulus is processed repeatedly by the same person. As soon as the diffusion process hits an upper or lower threshold, a decision for one or the other response is made.

### Parameters of the Diffusion Model

In a diffusion model analysis, several parameters are estimated from the empirical response time distributions, and these parameters are associated with different cognitive processes. Distinct psychological interpretations have been assigned to all model parameters. The basic diffusion model, as described by Ratcliff (1978), comprises four

parameters: drift rate $v$, threshold separation $a$, starting point $z$ and duration of non-decisional processes $t_0$.

The drift rate $v$ reflects the average slope of the diffusion process. It depends both on the speed of a participant´s information processing and the task difficulty. Accordingly, drift rates closer to zero indicate a slower processing of information or more difficult tasks (Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007; Voss, Rothermund, & Voss, 2004). Threshold separation $a$ describes the amount of information that is needed to draw a conclusion. High values of this parameter indicate a conservative decisional style with slow responses and high accuracy, whereas low values represent a liberal style with fast responses and higher error rates (Ratcliff, Thapar, Gomez, & McKoon, 2004). The starting point $z$ of the information accumulation is located between the two thresholds. In case of an a priori bias, it can be located closer to the threshold corresponding to the preferred response (Voss, Rothermund, & Brandtstadter, 2008; Voss et al., 2004). Finally, the duration of extra-decisional processes $t_0$, such as encoding and motoric response processes, is added to the decision times determined by the diffusion process (Ratcliff, Spieler, & McKoon, 2000). Note that some researchers question the validity of the psychological interpretation of parameters in evidence accumulation modeling (Jones & Dzhafarov, 2014), while others argue in favor of it (e.g., Heathcote, Wagenmakers, & Brown, 2014).

In the full version of the diffusion model (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002) that is often used in psychological applications, additional inter-trial variabilities are estimated for drift, starting point, and non-decision time. Note, however, that parameter estimation is often more accurate when more parsimonious models are used, suggesting that these inter-trial variability parameters might lead to overfitting (Lerche & Voss, 2016).

### The Lévy Flight Model

Recent studies suggest that the noise in evidence accumulation might be better described by heavy-tailed distributions (Voss et al., 2019). Heavy-tailed distributions like the Cauchy distribution or the Lévy distribution are characterized by an high likelihood for extreme events, compared to a normal distribution (Voss et al., 2019). Lévy flights have been applied to a variety of contexts. For example, they have proven useful to model animal foraging behavior. The Lévy flight foraging hypothesis states that in certain environments (truncated) Lévy flights optimize random searches. According to this hypothesis, biological organisms have evolved to exploit Lévy flights for their wandering movements during foraging (Viswanathan, Raposo, & da Luz, 2008). For example, Reynolds (2012) reports evidence for Lévy flights in the fishy-scented olfactory-

**Figure 1 ■** Basic version of the Lévy flight model. An information accumulation process starts at a starting point $z$ and runs over time with the mean slope $v$ until it hits an upper ($a$) or lower (0) threshold. Because of random noise, the process durations and outcomes vary from trial to trial. Outside the threshold decision-time distributions are shown. Due to a heavy-tailed noise distribution, sudden large jumps can be observed in the information accumulation process.
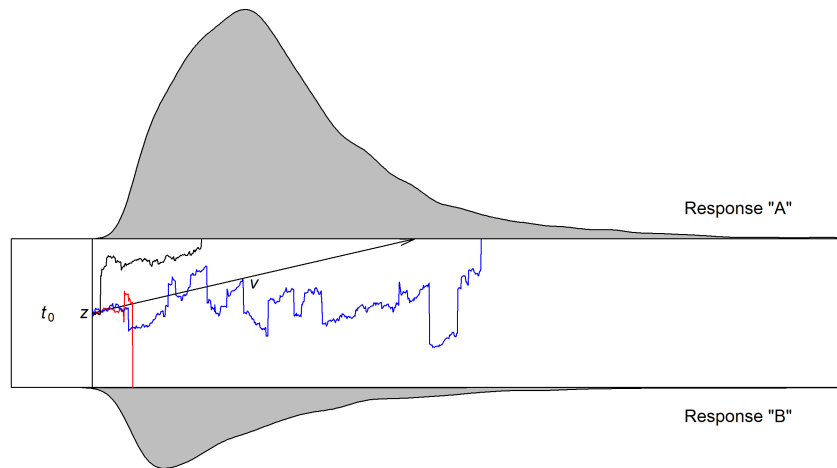


mediated prey detection of wandering albatross regarding the length of flights between landings. The author concludes that natural selection lead to Lévy flight searching as this is the optimal pattern when prey is sparsely and randomly distributed, whereas Brownian motion (i.e., a process with Gaussian noise) is only effective when prey is abundant. Similar search patterns have been observed in other kinds of species, including several open-ocean predatory fish (Humphries et al., 2010).[1] Lévy flights are also used in other scientific domains, as for example, to describe economic processes. Here, it has been shown that price indices in a stock exchange have statistical properties that are compatible with a Lévy random walk (Mantegna, 1991).

In the field of cognitive research, there are only few studies that provide evidence of Lévy flight processes in human perception and cognition. Montez et al. (2015) applied Lévy flights to searching and clustering in semantic memory. In a first experiment, Rhodes and Turvey (2007) had participants recall as many animal names as possible within 20 minutes. Inter-response intervals were recorded. In a second experiment by Montez et al. (2015), other participants had to arrange magnets with animal names, taken from the previous experiment, on a whiteboard with spatial distances representing the similarity of the species. Inter-response intervals from the first experiment correlated with spatial distances from the second experiment and distributions of both variables approxi-

mately followed predications from Lévy flights (Montez et al., 2015). Lévy flights have also been observed in the field of human perception. Liberati et al. (2017) conducted an eye-tracking study in which they showed typically developed children and children with autism spectrum disorder images of an adult gazing toward one of two objects. Scan-paths of gaze position of both groups were characterized by a probability distribution geometrically equivalent to Lévy flights.

Voss et al. (2019) applied a Lévy flight information accumulation model to analyze data from decision making based on a number-letter classification task. In a simple single-stimulus task, participants had to classify presented stimuli as numbers vs. letters, whereas in a more difficult multiple-stimulus task, participants estimated whether there were more letters or more numbers in a set of simultaneously presented stimuli. Both tasks were administered under speed and accuracy instructions. A model with an additional parameter that indicated the heaviness in the tails of the noise distribution (corresponding to Model 2 in the present study), fit the data better than a model with Gaussian noise. Moreover, larger jumps in the decision process were observed in the single-stimulus condition compared to the more complex multi-stimulus condition and under speed instructions compared to accuracy instructions.

Thus, Voss et al. (2019) provided first evidence that a Lévy flight model might be applicable to human decision

---

[1]Note, however, that the Lévy flight foraging hypothesis has been put into question (Palyulin, Chechkin, & Metzler, 2014).

making. In contrast to the diffusion model, the Lévy flight model allows for jumps in the decision process (see Figure 1). More heavy-tailed random noise in an evidence accumulation model might mirror cognitive processes in binary decision tasks even better than normally distributed noise.

Distributions such as the normal distribution, the Cauchy distribution or the Lévy distribution are special cases of a class of distributions called Lévy alpha-stable distributions. The heaviness in the tails of the distribution is described by the stability parameter $\alpha \in [0, 2]$. The normal distribution is indicated by a value of $\alpha = 2$ and the more heavy-tailed Cauchy distribution by $\alpha = 1$. In the present paper, we compare four different models:[2] Model 1 (standard diffusion model; $\alpha = 2$) assumes the noise to be normally distributed, whereas in Model 2 alpha is not fixed but estimated as an additional free parameter. Besides the fixed or free stability parameter $\alpha$, the two models comprise the free parameters $v_0$ and $v_1$ (drift rates for two stimulus types), $a$ (threshold separation), $z$ (starting point) and $t_0$ (non-decision time), as described above. Inter-trial variabilities were fixed to zero. Models 3 and 4 are full versions of the described models that additionally comprise $s_v$ (inter-trial variability of drift), $s_z$ (inter-trial variability of starting point) and $s_t$ (inter-trial variability of non-decision time).

### Research Questions and Hypotheses

Following the results of Voss et al. (2019), we expect a model with stability parameter alpha as an additional free parameter to fit the data better than a model with Gaussian noise, both for the color discrimination task and for the lexical decision task. Thus, we assume alpha to take average values between 1.0 (Cauchy noise) and 2.0 (Gaussian noise). Additionally, we expect a positive correlation between individual alpha-values across the two tasks, indicating meaningful inter-individual variance in this parameter.

### Method

#### *Participants*

The sample consisted of 40 participants (33 female, 6 male and 1 non-binary; mean age=21; range: 18-38) who were recruited with flyers at the Institute of Psychology at Heidelberg University. Consequently, the majority of participants were undergraduate students majoring in Psychology. All participants were German native speakers and none of them had impaired color vision. They gave an informed consent prior to the experiment and were granted partial course credit or 6 Euros as compensation.

#### *Design*

The design comprised the within-participant factors "task" (color discrimination vs. lexical decision) and "stimulus type". For both tasks, there were two different stimulus types (orange vs. blue or word vs. non-word for color discrimination and lexical decision, respectively). The order of tasks was counterbalanced across participants.

#### *Apparatus and Stimuli*

Stimuli were presented on a 17-inch computer screen. For color discrimination (Voss et al., 2004), we used colored squares (approximately 40 × 50 mm) consisting of 150 × 200 pixels. Each pixel was either orange or light blue. The proportion of colors within each stimulus was 47:53. Pixels were randomly intermixed. For half of the stimuli, orange was the dominant color; for the other half, blue was dominant.

For the lexical decision task, we re-used stimuli from Lerche and Voss (2017). Two-hundred German nouns with one or two syllables and four to six letters served as word stimuli. All words had low frequency in German language. For each word stimulus, a non-word was created by random replacement of all vowels. Stimuli were presented centered on the screen in Arial 20 pt font.

After the experimental tasks, a German adaptation of the UPPS Impulsive Behavior Scale was used to assess participants' impulsivity (Schmidt, Gay, d'Acremont, & Van der Linden, 2008).[3]

#### *Procedure*

The experiment was conducted in five computer-based group sessions with 6 to 10 participants. Participants were instructed to respond as quickly as possible, even if this would lead to some mistakes. Thereby, we intended to generate higher error rates that were sufficiently large for a robust parameter estimation. After every 100 trials, the opportunity for a short break was given.

For each of the two tasks, participants started with a training block of 12 trials. Only in this training block, accuracy feedback was given after each response. Thereafter, participants completed five experimental blocks of the color discrimination task and four blocks of the lexical decision task (order of tasks was counterbalanced) with 100 trials per block. Each trial had the following sequence: First, a white screen was presented for 500 ms. Afterwards, a fixation cross appeared for 250 ms at the center of the

---

[2] In addition to the described models, we analyzed the fit of a simple and a complex version of a model with Cauchy-distributed noise ($\alpha = 1.0$). As this model yielded an inferior fit in comparison to the other models, we did not describe the model in detail here to improve readability.

[3] By assessing participants' self-rated impulsivity, we intended to exploratively analyze the correlations between parameter alpha and UPPS scores. As the analysis did not yield any statistically significant correlations, results are not reported in detail here.

screen before it was replaced by a colored square in the color discrimination task or by a letter sequence in the lexical decision task. The stimulus remained on the screen until a decision was made. Responses were given by pressing the A-key for "blue" or "non-word" and the L-key for "orange" or "word" on a standard keyboard. Labels for the assignment of keys were presented at the bottom-corners of the screen throughout the experiment. The total duration of the experiment was approximately 45 minutes.

### Fitting the Models to the Data

The models' parameters were estimated for each task and each participant separately. This requires a multidimensional search for a set of parameter estimates that provide an optimal fit between predicted and empirical response time distributions. Whereas for the diffusion model, probability density functions (PDF) for response times are known (e.g., Navarro & Fuss, 2009; Voss & Voss, 2008), this is not the case for the model with alpha as an additional free parameter. Therefore, a direct calculation of the likelihood is not possible. Voss et al. (2019) applied a somewhat cumbersome simulation based approach to approximate the likelihood. Recent work from our lab showed that a deep learning approach for likelihood-free parameter estimation is much more efficient (Radev, Mertens, Voss, Ardizzone, & Köthe, 2019). In the following section, we describe the rationale of this new approach briefly.

Our estimation method draws on recent advances in deep probabilistic modeling (Ardizzone, Lüth, Kruse, Rother, & Köthe, 2019; Grover, Dhar, & Ermon, 2017; Kingma & Dhariwal, 2018; Radev et al., 2019). It involves two neural networks (a summary network that learns to extract the most informative summary statistics from raw data and an invertible network that learns the relation of these summary statistics to the true parameter values) which jointly learn a probabilistic mapping from data to parameters without assumptions on the parametric form of the posterior distributions of all parameters. The networks are trained from simulated data by implicitly minimizing the Kullback-Leibler (KL) divergence between the approximate posterior deduced by the networks and the true posterior of the model parameters. Moreover, the method recovers the true posterior exactly under optimal performance of the networks, as proved by Radev et al. (2019). Once trained with a sufficient amount of simulated data, the converged networks can be used to perform rapid likelihood-free Bayesian inference, thus essentially amortizing costs of training. Specifically, the method involves the following steps:

1. A broad enough prior on the models parameters is specified, such that the prior captures a realistic range of all parameter values.

2. Data are simulated on the fly by repeatedly drawing from the prior and generating artificial response time data.
3. The simulated data are fed into the networks which iteratively minimize the KL divergence between the approximate posterior (deduced by the networks) and the true posterior over parameters.
4. The trained networks are applied to the observed data in order to approximate parameter posteriors.

By splitting the process of parameter estimation into a training and an inference phase, the computational load is "outsourced" into the training phase. Subsequent inference involves only passing observed data sets through the trained networks, which is computationally cheap and very efficient. Moreover, the trained networks can be stored and re-used for estimating the parameters of the model they have been trained on.

The method described above provides posterior distributions for all parameters. To allow for model comparisons, we then employed the simulation approach by Voss et al. (2019) to approximate the likelihood at the mean parameter values for each model and each person. This allowed us to calculate the BIC. This information criterion is defined as follows with smaller values indicating a better fit (Voss, Voss, & Lerche, 2015):

$$BIC = -2LL - P \cdot \ln(M),$$

where $LL$ is the log-likelihood, $P$ is the number of free parameters and $M$ is the number of observations (i.e., trials).

### Results

### Data Pre-Treatment

Data from trials with logarithmized response times that fell more than 1.5 interquartile ranges below a participant's first quartile or more than 1.5 interquartile ranges above a participant's third quartile were removed prior to all analyses. This criterion led to an exclusion of 3.59% of trials in the color discrimination task and of 2.64% of trials in the lexical decision task.

### Response Times and Accuracy

Mean correct and error response times (RT) and accuracy-values for the different types of stimuli in the two tasks are presented in Table 1. Additionally, response times and accuracy-values are visualized in Figures 2 and 3, respectively.

As can be seen, for the color discrimination task, response times for errors were slightly slower that for correct responses. In the lexical decision task, we observed fast errors in response to non-word stimuli, but slower errors in response to word stimuli. Accuracy-values suggest

**Figure 2** ∎ Reaction times in milliseconds for correct responses and error responses to non-word and word stimuli in the lexical decision task and to predominantly blue and predominantly orange stimuli in the color discrimination task for the 40 participants.
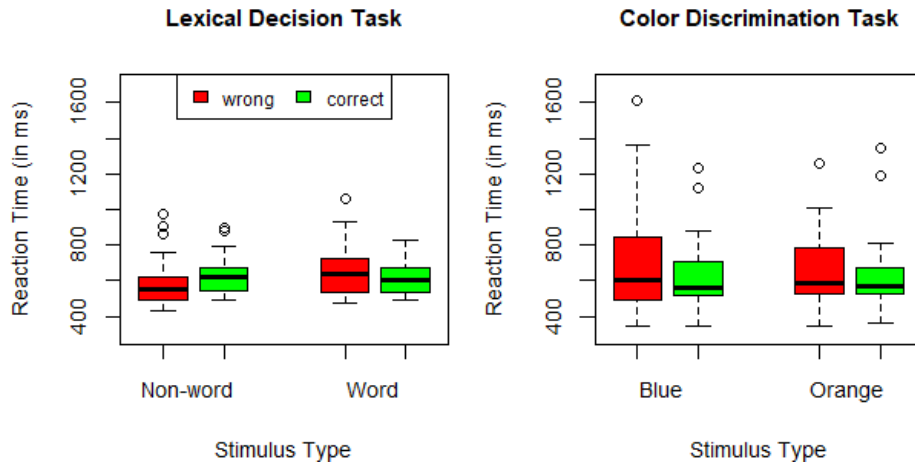


**Table 1** ∎ Mean response times (RT in ms) and accuracy in percent (standard deviations in parentheses) as a function of stimulus type, response accuracy, and task.

| Task | Color Discrimination | | Lexical Decision | |
|---|---|---|---|---|
| Stimulus type | Blue | Orange | Word | Non-Word |
| RT (correct) | 621 (179) | 621 (179) | 617 (88) | 628 (98) |
| RT (error) | 702 (283) | 675 (275) | 650 (128) | 582 (123) |
| Accuracy (%) | 85.2 (12.3) | 86.7 (9.9) | 87.2 (7.8) | 95.3 (3.7) |

*Note.* $N = 40$, except for non-word stimuli, where one participant had to be excluded as she did not make any mistakes.

that the lexical decision task was overall somewhat easier than the color discrimination task. This difference is largely based on high accuracy rates for non-words.

While response times do not differ largely between the different conditions in the two tasks, the variance in performance between participants regarding response time and accuracy is larger in the color discrimination task in comparison to the lexical decision task. Lastly, a larger variance in response times can also be observed for error responses compared to correct responses.

### Model Fit

To assess model fit, BIC was calculated separately for all participants for the two models and the two tasks. We decided for the report of BIC values instead of AIC values (Akaike Information Criterion) as BIC is consistent, whereas AIC might tend to select complex models that overfit the data (Vandekerckhove, Matzke, & Wagenmak-

ers, 2015). Summarized BIC values are presented in Table 2 with smaller values indicating a better fit. Additionally, the number of participants for whom the respective simple and full models had the best fit are shown.[4]

These analyses reveal that overall Model 4 (i.e., complex Lévy flight with freely estimated stability of noise and inter-trial variability) shows the best fit. When comparing simple models (without inter-trial-variability parameters) to each other, Model 2 (i.e., a model with alpha as an additional free parameter) performs better than the first model for both tasks.

When comparing the complex models, which allow for inter-trial variability in starting point, drift rate and non-decisional processes, Model 4 with alpha as an additional free parameter has a better fit than the full diffusion model (Model 3) across tasks. Note that the differences in BIC are huge, which is indicated by a performed transformation of BIC values to Schwarz weights (Vandekerckhove et

---

[4]We also calculated AIC values for model comparison. The analysis yielded identical results as the analysis performed with BIC values.

**Figure 3** ■ Accuracy values for word stimuli and non-word stimuli in the lexical decision task and for predominantly blue and predominantly orange stimuli in the color discrimination task for the 40 participants.
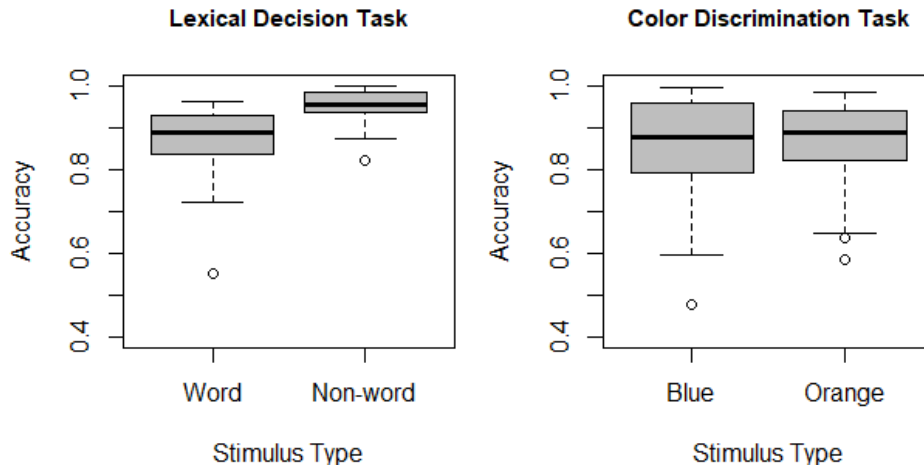


**Table 2** ■ Model fit assessed with BIC (number of participants for whom each model had the best fit in parentheses) for all tasks and models.

| Model | Color Discrimination | Lexical Decision |
|---|---|---|
| 1. Standard Diffusion | 152,880 (8) | 110,364 (1) |
| 2. Simple Lévy Flight | **151,177 (32)** | **108,376 (39)** |
| 3. Complex Diffusion | 145,791 **(22)** | 107,022 **(28)** |
| 4. Complex Lévy Flight | **145,618** (18) | **106,961** (12) |

*Note.* $N = 40$. BIC: Bayesian information criterion. Smaller values indicate a better fit. The best fit in each column –separately for simple and complex models– is printed in bold font.

al., 2015). As a result of this analysis, weights of nearly 1 for the superior model were observed. Results also suggest notable inter-individual differences in the type of information sampling, as different models are superior for different participants.

As Model 1 (the standard diffusion model with a fixed parameter $\alpha = 2$) is nested in Model 2 (the Lévy model with alpha as an additional free parameter), we additionally performed a likelihood ratio test to compare the fit of these models. Highly significant advantages were observed for the Lévy flight model for data from the lexical decision task, $\chi^2(40) = 4,058$, $p < .001$, and from the color discrimination task, $\chi^2(40) = 1,950$, $p < .001$.

### Parameter Estimates

Averages from the mean posterior distributions of all parameters for the four tested models are presented in Table 3. Additionally, distributions of mean alpha-values are presented in Figure 4 for the simple Lévy flight model. For

the 40 participants, alpha is approximately uniformly distributed in a range from $\alpha = 1.0$ (i.e., a model with Cauchy-distributed noise) to $\alpha = 2.0$ (i.e., the Ratcliff diffusion model with normally distributed noise).

### Correlations of Alpha-Values with Accuracy and Response Times

To improve the understanding, which characteristics of behavioral data are indicative of alpha, mean posterior values for alpha from Model 2 were correlated with response times for correct responses as well as accuracy rates. To increase normality, response time values were log transformed and accuracy-values were arcsin transformed prior to the analysis. Correlations are presented in Table 4 and visualized in Figure 5.

Moderate to high positive correlations were observed between response times and alpha-values across all tasks and stimuli, indicating that larger jumps in the decision process are associated with faster responses. Alpha-values

**Table 3** ■ Mean parameter values (standard deviations in parentheses) for simple and complex versions of the diffusion model and the model with stable noise in the color discrimination and the lexical decision task.

| Task | $a$ | $z_r$ | $v_0$ | $v_1$ | $t_0$ | $\alpha$ | $s_z$ | $s_v$ | $s_t$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | Model 1: Standard Diffusion | | | | | | |
| Color Discrimination | 1.24 | 0.49 | -1.92 | 2.00 | 0.34 | 2 (0) | 0 (0) | 0 (0) | 0 (0) |
| | (0.31) | (0.06) | (1.09) | (0.77) | (0.05) | | | | |
| Lexical Decision | 1.19 | 0.52 | -2.89 | 2.29 | 0.41 | 2 (0) | 0 (0) | 0 (0) | 0 (0) |
| | (0.28) | (0.05) | (0.64) | (0.58) | (0.03) | | | | |
| | | | Model 2: Simple Lévy Flight | | | | | | |
| Color Discrimination | 1.06 | 0.49 | -1.68 | 1.77 | 0.35 | 1.60 | 0 (0) | 0 (0) | 0 (0) |
| | (0.42) | (0.08) | (0.98) | (0.77) | (0.05) | (0.28) | | | |
| Lexical Decision | 1.05 | 0.52 | -2.51 | 2.10 | 0.41 | 1.61 | 0 (0) | 0 (0) | 0 (0) |
| | (0.36) | (0.06) | (0.64) | (0.59) | (0.03) | (0.21) | | | |
| | | | Model 3: Complex Diffusion | | | | | | |
| Color Discrimination | 1.17 | 0.49 | -2.46 | 2.61 | 0.40 | 2 (0) | 0.30 | 0.76 | 0.19 |
| | (0.46) | (0.08) | (1.22) | (0.99) | (0.06) | | (0.15) | (0.50) | (0.06) |
| Lexical Decision | 1.00 | 0.52 | -3.46 | 2.66 | 0.46 | 2 (0) | 0.33 | 0.37 | 0.16 |
| | (0.26) | (0.07) | (0.80) | (0.68) | (0.03) | | (0.12) | (0.27) | (0.04) |
| | | | Model 4: Complex Lévy Flight | | | | | | |
| Color Discrimination | 1.12 | 0.49 | -2.27 | 2.38 | 0.39 | 1.80 | 0.26 | 0.69 | 0.17 |
| | (0.50) | (0.08) | (1.18) | (0.96) | (0.06) | (0.20) | (0.10) | (0.50) | (0.06) |
| Lexical Decision | 0.99 | 0.52 | -3.29 | 2.51 | 0.46 | 1.87 | 0.35 | 0.35 | 0.15 |
| | (0.31) | (0.07) | (0.77) | (0.64) | (0.04) | (0.13) | (0.09) | (0.23) | (0.04) |

*Note.* $N = 40$. $a$ = threshold separation; $z_r$ = relative starting point; $v_0$ = drift rate for blue stimuli or non-word stimuli; $v_1$ = drift rate for orange stimuli or word stimuli; $t_0$ = non-decisional time; $\alpha$ = stability parameter of noise distribution; $s_z$ = across-trial variability in starting point; $s_v$ = across-trial variability in drift rate; $s_t$ = across-trial variability in non-decisional processes.

were also positively related to accuracy, showing significant correlations between alpha-values and accuracy in color discrimination, $r(38) = .34$, 95% CI [.03; .59], as well as between accuracy in lexical decision and alpha-values in color discrimination, $r(38) = .35$, 95% CI[.04; .59].[5]

Importantly, also a significant correlation was observed between alpha-values from color discrimination and lexical decision, $r(38) = .41$, 95% CI [.12; .64]. This inter-task correlation of alpha suggests a stable, trait-like component of the quality of information accumulation. Correlations of parameters in the simple Lévy flight model (Model 2) across both tasks are shown in Table 5.

### Discussion

**Summary and Interpretation of Results.** In the present study, we compared the fit of four evidence accumulation models applied to a color discrimination and a lexical decision task: In the color discrimination task, participants had to indicate whether there were more orange or more blue pixels in the presented stimuli. In the lexical deci-

sion task, participants had to decide whether a presented letter-string was an existing German word or a meaningless letter-string. The examined models included (1) a parsimonious version of the diffusion model with Gaussian noise (Ratcliff, 1978), and (2) a model with a stable distribution for the noise of the information accumulation, where the heavy-tailedness was modelled with an additional free parameter. Besides the distribution of noise in evidence accumulation (sometimes also denoted as intra-trial variability of drift), the models were identical. Additionally, (3) a full version of the diffusion model with inter-trial variability for drift, starting point and non-decision time (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002), and (4) a model with alpha as a free parameter and inter-trial variabilities were considered. Evidence accumulation in the models with non-normal noise follows a so-called Lévy flight, that is characterized by larger jumps in the stochastic process (Voss et al., 2019). The Gaussian distribution is a special case of the family of Lévy alpha-stable distribution with a stability value of $\alpha = 2$. Lower values of stabil-

---

[5]Please note that the correlations of alpha with reaction times and accuracy rates are of exploratory nature and have to be interpreted in a cautious manner. In this exploratory phase, we did not correct for multiple comparisons as we did not want to discard upcoming evidence to early and easily.

**Figure 4 ■** Distribution of estimated alpha values (means of posterior distributions) for the 40 participants in the lexical decision task and the color discrimination task. The values are nearly uniformly distributed between $\alpha = 1.0$ (Cauchy-distributed noise) and $\alpha = 2.0$ (normally distributed noise).
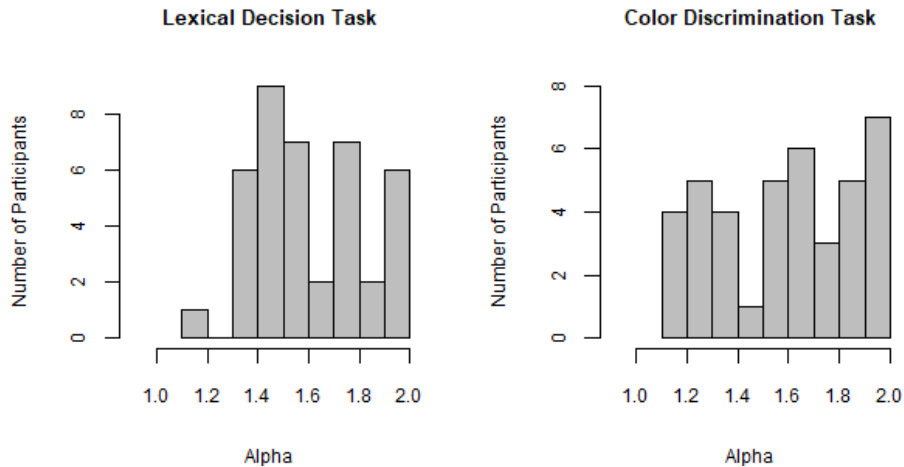


ity parameter alpha indicate a higher probability of jumps and also larger jumps in the decision process.

Accuracy rates were slightly higher in the lexical decision task than in the color discrimination task, suggesting that the latter was more difficult. Responses to non-word stimuli exhibited larger accuracy rates than the responses to word stimuli in the lexical decision task. This suggests that participants might not have been familiar with some of the used words due to their low frequency in German language.

**Comparison of Models.** Model 1 is a standard diffusion model. Five parameters ($a$, $z_r$, $v_0$, $v_1$ and $t_0$) were estimated for each participant and task. The stability parameter was fixed to a value of $\alpha = 2$, that is, Model 1 includes Gaussian noise in evidence accumulation. In Model 2 a more general stable distribution (with free parameter $\alpha$) is assumed for the random noise. Due to their heavy tails, the distribution allows the occurrence of extreme events, that is, jumps in the evidence accumulation process. Models 3 to 4 are equivalent to Models 1 and 2, respectively, except for the additional inclusion of inter-trial variabilities in drift, starting point and non-decision times.

Model parameters were fitted by using a novel Bayesian approach based on machine learning with neural networks (BayesFlow; Radev et al., 2019). This method provides accurate posterior distributions for parameters of mathematical models without requiring an explicit calculation of the likelihood. Rather, the neural networks learn the relation of data to plausible parameters from simulated data.

To compare fit of the four models, the BIC information criterion (Kuha, 2004) was computed. In accordance with previous results (Voss et al., 2019), the model with stable distributed noise (Model 2) had the best fit within the simple models (without inter-trial-variability) for both tasks. Within the complex models, that allow for inter-trial variability of starting point, drift rate and non-decision time, the complex Lévy model performed better than the complex diffusion model.

Unlike results from Voss et al. (2019) suggesting a superior fit of the simple Lévy flight model compared to the full diffusion model, in the present study the complex model had a superior fit compared to the simple Lévy flight model. All full models had a better fit than the simple models. We assume that these differences between the findings from Voss et al. (2019) and the present study might be based on the larger trial number: Possibly, the longer experimental duration causes more fluctuations in performance which in turn makes it necessary to include inter-trial variability parameters in the cognitive models.

**Inter-individual differences in alpha.** Results indicate that average values for alpha in the simple Lévy model are around 1.6 for both tasks, thus falling between a standard diffusion model and a model with Cauchy noise. This finding suggests that models with stable noise are applicable across different paradigms that have been applied in diffusion modeling.

In addition to the question, what model fit the data best, we were also interested to test whether alpha measures stable inter-individual differences in decision making. Substantial correlations of alpha-values were observed across the two completely different tasks. This find-

**Table 4** ∎ Correlations of alpha, accuracy and mean response time ($BF_{10}$ - i.e., Bayes factors in favor of an existing correlation - in parentheses and 95% confidence intervals in square brackets).

|  | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1. $\alpha$ in Color Discrimination | – |  |  |  |  |
| 2. $\alpha$ in Lexical Decision | .41 (5.83) [.12; .64] | – |  |  |  |
| 3. Accuracy in Color Discrimination | .34 (1.63) [.03; .59] | .24 (0.58) [-.08; .51] | – |  |  |
| 4. Accuracy in Lexical Decision | .35 (1.64) [.04; .59] | .32 (1.32) [.00; .57] | -.04 (0.20) [-.35; .27] | – |  |
| 5. RT in Color Discrimination | .33 (1.94) [.02; .58] | .48 (20.39) [.19; .69] | -.03 (0.20) [-.33; .29] | .46 (14.27) [.17; .67] | – |
| 6. RT in Lexical Decision | .33 (1.69) [.02; .58] | .74 (357,830.) [.56; .86] | .13 (0.27) [-.19; .42] | .23 (0.52) [-.09; .50] | .49 (28.63) [.21; .70] |

*Note.* $N = 40$.

**Table 5** ∎ Correlations with 95% confidence intervals between parameters in the color discrimination and the lexical decision task for the simple Lévy flight model.

| Parameter | $\alpha$ | $a$ | $z_r$ | $v_0$ | $v_1$ | $t_0$ |
|---|---|---|---|---|---|---|
| Correlation Coefficient | .41 | .67 | .13 | -.07 | .01 | .15 |
| 95% Confidence Interval | [.12; .64] | [.45; .81] | [-.19; .43] | [-.37; .25] | [-.30; .32] | [-.17; .44] |

*Note.* $N = 40$. $\alpha$ = stability parameter of noise distribution; $a$ = threshold separation; $z_r$ = relative starting point; $v_0$ = drift rate for blue stimuli or non-word stimuli; $v_1$ = drift rate for orange stimuli or word stimuli; $t_0$ = non-decisional time.

ing provides first evidence for a trait-like quality of the stability parameter. However, it remains unclear what the differences in alpha indicate in psychological terms. On a theoretic account, two opposing hypotheses are possible: On the one hand, increased jumps in evidence accumulation could reflect an irrational "jumping to conclusion" (McKay, Langdon, & Coltheart, 2006), that is, a premature and inefficient decision strategy (inefficient jumping hypothesis). On the other hand, it is also possible that a Lévy flight style of decision making reflects an efficient way to process information: Especially for easy, perceptual decisions that are conducted under time pressure, the processing of single pieces of information should change the subjective beliefs about the stimulus immediately. Following this thought, a gradual, diffusion-like, decision making style might reflect an inefficient use of information, while efficient decision making might be characterized by more jumps in the evidence accumulation process (efficient jumping hypothesis).

Following the inefficient jumping hypothesis, larger jumps – indicated by lower alpha-values – would be expected in populations with dysfunctional decision making or high values of impulsivity, such as people suffering from schizophrenia, borderline personality disorder or

substance abuse (Oshri et al., 2018; Richard-Lepouriel et al., 2019; Sterzer, Voss, Schlagenhauf, & Heinz, 2019). According to the efficient jumping hypothesis, larger jumps would be expected in individuals with efficient decision making and presumably high intelligence. The present data do not allow us to differentiate between both hypotheses. Therefore, future Lévy flight studies should include other measures of cognitive performance like, intelligence, working memory functioning, or cognitive flexibility.

### Evaluation of the Present Study and Implications for Further Research

The present study is one of the first attempts to apply Lévy flights in the context of decision making. For this reason, its results shed first light on a nearly unstudied field. In comparison to a previous study by Voss et al. (2019), we used notably larger trial numbers here. Thus, the present study allows for a more reliable assessment and comparison of the model fit. Correlations of alpha-values across tasks provide evidence for a meaningful inter-individual variance in this parameter. The inclusion of the parameter alpha in decision modeling provides further insight into human binary decision making as it helps to identify the exact processes that underlie fast reaction times and in-

**Figure 5 ■** Correlations of alpha, arcsin transformed accuracy and log transformed mean response time (RT) for the 40 participants in the lexical decision task (Lexical) and the color discrimination task (Color).
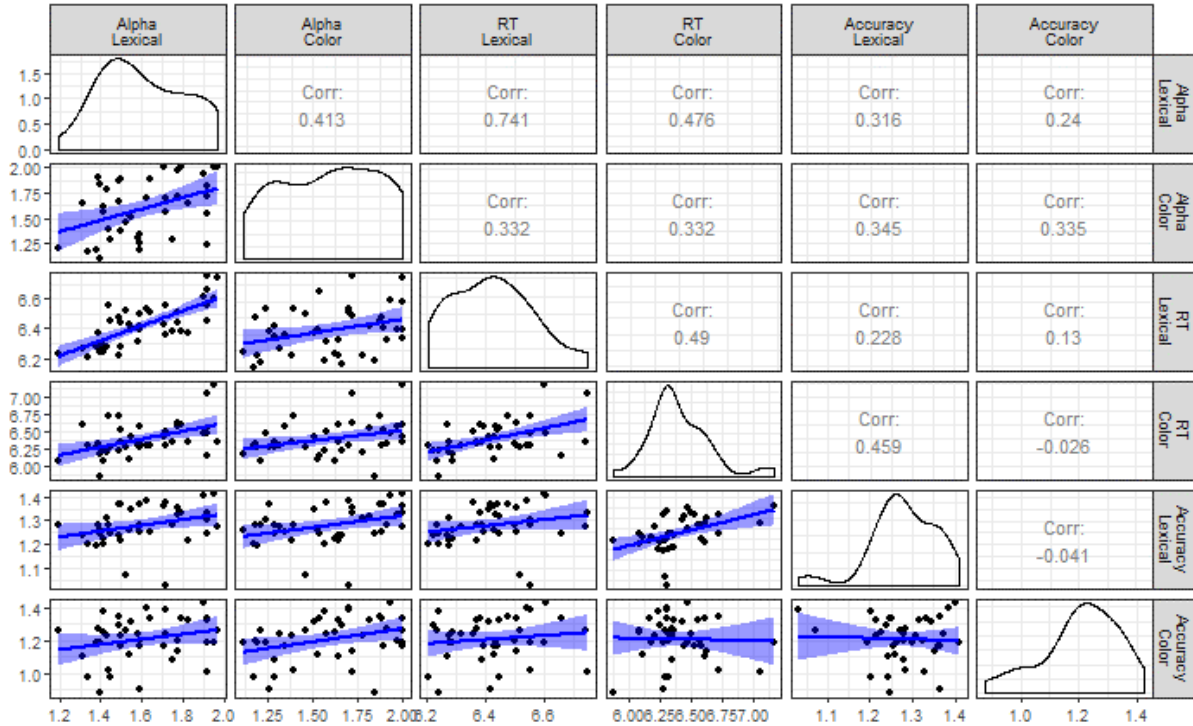


creased rates of fast errors. Consequently, the Lévy flight model allows the differentiation between different explanations for fast responses, which could alternatively occur due to stronger drift, lower decision thresholds, faster non-decision processes or a combination of these parameter values.

Some limitations of the presented study need to be addressed. Firstly, the present sample is limited in size and consists mainly of female psychology students in their early twenties. The study of this highly selective group could lead to a restricted variance of one or several of the assessed variables and therefore to an underestimation of correlations (e.g., between alpha and impulsivity).

Secondly, alternative criteria to BIC should be considered for model comparison. An adequate punishment of complexity and potential problems of overfitting of the complex models have to be analyzed more carefully.

Thirdly, models with alternative combinations of free parameters might show a superior model fit and should be considered in future research. For example, a model with alpha as a free parameter and inter-trial variability for non-decisional processes only might be a more parsimonious alternative to the complex Lévy model with inter-trial variability for starting point, drift rate and non-decisional processes. At the same time, such a model could possibly provide a better way to accommodate and explain fast errors compared to the full diffusion model.

Fourthly, in the present research we assume that jumps toward the correct and the error response boundary have the same probability. However, there are theoretical reasons to expect that jumps toward the correct boundary are more probable than jumps in the opposite direction. For example, one could assume Lévy flights to represent a kind of sudden insight into the solution of a problem and would therefore expect a higher probability of jumps toward the correct response boundary. Future studies should test models that allow for asymmetric proportions of jumps toward the correct and the error boundary in the noise distribution.

Lastly, the psychological meaning of stability parameter alpha requires further examination. As self-reported measures of impulsivity did not show any significant correlations with parameter alpha, experimental measures of impulsivity should be considered. Careful theories are to be developed to address the question of cognition or personality related correlates of alpha. Subsequently, the re-

sulting hypotheses have to be backed up by further empirical work.

## References

Ardizzone, L., Lüth, C., Kruse, J., Rother, C., & Köthe, U. (2019). Guided image generation with conditional invertible neural networks. *arXiv preprint*, 1907.02392.

Grover, A., Dhar, M., & Ermon, S. (2017). Flow-gan: Combining maximum likelihood and adversarial learning in generative models. Retrieved, from https://ui.adsabs.harvard.edu/abs/2017arXiv170508868G

Heathcote, A., Wagenmakers, E. J., & Brown, S. D. (2014). The falsifiability of actual decision-making models. *Psychological Review*, *121*(4), 676–678. doi:10.1037/a0037771

Humphries, N. E., Queiroz, N., Dyer, J. R., Pade, N. G., Musyl, M. K., Schaefer, K. M., . . . Sims, D. W. (2010). Environmental context explains levy and brownian movement patterns of marine predators. *Nature*, *465*(7301), 1066–1069. doi:10.1038/nature09116

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

Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. *arXiv e-prints*, *1*, 1–99. Retrieved from https://ui.adsabs.harvard.edu/abs/2018arXiv180703039K

Kuha, J. (2004). Aic and bic: Comparisons of assumptions and performance. *Sociological methods & research*, *33*(2), 188–229. doi:10.1177/0049124103262065

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. Oxford: Academic Press.

Lerche, V., & Voss, A. (2016). Model complexity in diffusion modeling: Benefits of making the model more parsimonious. *Frontiers in psychology*, *7*, 1324–1326. doi:10.3389/fpsyg.2016.01324

Lerche, V., & Voss, A. (2017). Retest reliability of the parameters of the ratcliff diffusion model. *Psychological Research*, *81*(3), 629–652. doi:10.1007/s00426-016-0770-5

Liberati, A., Fadda, R., Doneddu, G., Congiu, S., Javarone, M. A., Striano, T., & Chessa, A. (2017). A statistical physics perspective to understand social visual attention in autism spectrum disorder. *Perception*, *46*(8), 889–913. doi:10.1177/0301006616685976

Mantegna, R. N. (1991). Lévy walks and enhanced diffusion in milan stock exchange. *Physica A: Statistical Mechanics and its Applications*, *179*(2), 232–242. doi:10.1016/0378-4371(91)90061-G

McKay, R., Langdon, R., & Coltheart, M. (2006). Need for closure, jumping to conclusions, and decisiveness in delusion-prone individuals. *The Journal of nervous and mental disease*, *194*(6), 422–426. doi:10.1097/01.nmd.0000221353.44132.25

Montez, P., Thompson, G., & Kello, C. T. (2015). The role of semantic clustering in optimal memory foraging. *Cognitive science*, *39*(8), 1925–1939. doi:10.1111/cogs.12249

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in wiener diffusion models. *Journal of Mathematical Psychology*, *53*(4), 222–230. doi:10.1016/j.jmp.2009.02.003

Oshri, A., Kogan, S. M., Kwon, J. A., Wickrama, K. A. S., Vanderbroek, L., Palmer, A. A., & MacKillop, J. (2018). Impulsivity as a mechanism linking child abuse and neglect with substance use in adolescence and adulthood. *Development and Psychopathology*, *30*(2), 417–435. doi:10.1017/S0954579417000943

Palyulin, V. V., Chechkin, A. V., & Metzler, R. (2014). Lévy flights do not always optimize random blind search for sparse targets. *Proceedings of the National Academy of Sciences*, *111*(8), 2931–2936. doi:10.1073/pnas.1320424111

Radev, S., Mertens, U. K., Voss, A., Ardizzone, L., & Köthe, U. (2019). *Bayesflow: Learning complex stochastic models with invertible neural networks*. Manuscript submitted for publication.

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

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*(2), 278–291. doi:10.3758/BF03196283

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

Ratcliff, R., Spieler, D., & McKoon, G. (2000). Explicitly modeling the effects of aging on response time. *Psychonomic Bulletin & Review*, *7*(1), 1–25. doi:10.3758/BF03210723

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. doi:10.1037/0882-7974.19.2.278

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging*, *16*(2), 323–341. doi:10.1037/0882-7974.16.2.323

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. doi:10.3758/BF03194580

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychon Bull Rev*, *9*(3), 438–481. doi:10.3758/bf03196302

Reynolds, A. M. (2012). Olfactory search behaviour in the wandering albatross is predicted to give rise to levy flight movement patterns. *Animal Behaviour*, *83*(5), 1225–1229. doi:10.1016/j.anbehav.2012.02.014

Rhodes, T., & Turvey, M. T. (2007). Human memory retrieval as levy foraging. *Physica a-Statistical Mechanics and Its Applications*, *385*(1), 255–260. doi:10.1016/j.physa.2007.07.001

Richard-Lepouriel, H., Kung, A.-L., Hasler, R., Bellivier, F., Prada, P., Gard, S., . . . Etain, B. (2019). Impulsivity and its association with childhood trauma experiences across bipolar disorder, attention deficit hyperactivity disorder and borderline personality disorder. *Journal of Affective Disorders*, *244*, 33–41. doi:10.1016/j.jad.2018.07.060

Schmidt, R. E., Gay, P., d'Acremont, M., & Van der Linden, M. (2008). A German adaptation of the upps impulsive behavior scale: Psychometric properties and factor structure. *Swiss Journal of Psychology / Schweizerische Zeitschrift für Psychologie / Revue Suisse de Psychologie*, *67*(2), 107–112. doi:10.1024/1421-0185.67.2.107

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. doi:10.1037/0096-3445.136.3.414

Sterzer, P., Voss, M., Schlagenhauf, F., & Heinz, A. (2019). Decision-making in schizophrenia: A predictive-coding perspective. *Neuroimage*, *190*, 133–143. doi:10.1016/j.neuroimage.2018.05.074

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer & Z. Wang (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–319). doi:10.1093/oxfordhb/9780199957996.001.0001

Viswanathan, G. M., Raposo, E. P., & da Luz, M. G. E. (2008). Levy flights and superdiffusion in the context of biological encounters and random searches. *Physics of Life Reviews*, *5*(3), 133–150. doi:10.1016/j.plrev.2008.03.002

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*, 1–20. doi:10.3758/s13423-018-1560-4

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology*, *60*(6), 385–402. doi:10.1027/1618-3169/a000218

Voss, A., Rothermund, K., & Brandtstadter, J. (2008). Interpreting ambiguous stimuli: Separating perceptual and judgmental biases. *Journal of Experimental Social Psychology*, *44*(4), 1048–1056. doi:10.1016/j.jesp.2007.10.009

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Mem Cognit*, *32*(7), 1206–1220. doi:10.3758/bf03196893

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology*, *52*(1), 1–9. doi:10.1016/j.jmp.2007.09.005

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*, 336–339. doi:10.3389/fpsyg.2015.00336

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*(5), 641–671. doi:10.1080/09541440802205067

**Citation**