Response-time data provide critical constraints on dynamic models of multi-alternative, multi-attribute choice

Nathan J. Evans^{1,2} · William R. Holmes³ · Jennifer S. Trueblood¹

Published online: 8 February 2019
© The Psychonomic Society, Inc. 2019

Abstract

Understanding the cognitive processes involved in multi-alternative, multi-attribute choice is of interest to a wide range of fields including psychology, neuroscience, and economics. Prior investigations in this domain have relied primarily on choice data to compare different theories. Despite numerous such studies, results have largely been inconclusive. Our study uses state-of-the-art response-time modeling and data from 12 different experiments appearing in six different published studies to compare four previously proposed theories/models of these effects: multi-alternative decision field theory (MDFT), the leaky-competing accumulator (LCA), the multi-attribute linear ballistic accumulator (MLBA), and the associative accumulation model (AAM). All four models are, by design, dynamic process models and thus a comprehensive evaluation of their theoretical properties requires quantitative evaluation with both choice and response-time data. Our results show that response-time data is critical at distinguishing among these models and that using choice data alone can lead to inconclusive results for some datasets. In conclusion, we encourage future research to include response-time data in the evaluation of these models.

Keywords Decision-making · Multi-attribute choice · Context effects · Bayesian methods

Introduction

Multi-alternative, multi-attribute choice forms a fundamental part of everyday human life, ranging from simple decisions such as choosing today's lunch, to complex decisions such as selecting a retirement portfolio. In particular, decades of research have been devoted to understanding the cognitive processes that underlie these types of decisions using computational modeling. Over the years, many different

models of multi-alternative, multi-attribute choice have been developed, but these models are typically assessed in isolation. Only recently have attempts been made to compare and contrast their ability to account for empirical data (e.g., Trueblood, Brown, & Heathcote, 2014; Turner, Schley, Muller, & Tsetsos, 2018). However, conclusions from these model comparisons have been limited. The results of these comparisons often arrive at different conclusions (i.e., one model is preferred in one comparison and another in a different comparison). We believe these limitations are the result of using insufficient data for constraining the models. In the present paper, we illustrate that the predictions of different models can be clearly distinguished when constrained by the entire response-time distribution.

The study of multi-alternative, multi-attribute choice is a cornerstone of psychology, neuroscience, and economics research. Initially, formalized mathematical theories explained the process through utility models, where people decide in favor of the alternative with the highest subjective value. Importantly, most of these models obey a property called simple scalability where the probability of choosing an alternative is an increasing function of the difference between the utility of that alternative and the utility of the other options. An important consequence of simple scalability is independence among alternatives. For example, when

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13423-018-1557-z>) contains supplementary material, which is available to authorized users.

✉ William R. Holmes
william.holmes@vanderbilt.edu

✉ Jennifer S. Trueblood
jennifer.s.trueblood@vanderbilt.edu

¹ Department of Psychology, Vanderbilt University, Nashville, TN 37235, USA

² Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

³ Department of Physics and Astronomy, Department of Mathematics, Vanderbilt University, Nashville, TN 37235, USA

choosing between two laptops (A and B), independence dictates that the relative preference between A and B is unaffected by the introduction new alternatives.

However, violations of independence have been commonly observed within empirical data, with three key effects being the most often investigated: the attraction (Huber, Payne, & Puto, 1982), similarity (Tversky, 1972), and compromise effects (Simonson, 1989). Our study aims to provide a comprehensive assessment of four prominent models of multi-attribute choice using empirical data probing these three “context effects”: multi-alternative decision field theory (MDFT; Roe, Busemeyer, & Townsend, 2001; Hotaling, Busemeyer, & Li, 2010; Berkowitsch, Scheibehenne, & Rieskamp, 2014), the leaky competing accumulator (LCA; Usher & McClelland, 2004; Usher, Elhalal, & McClelland, 2008; Tsetsos, Usher, & Chater, 2010), the multi-attribute linear ballistic accumulator (MLBA; Trueblood et al., 2014), and the associative accumulation model (AAM; Bhatia, 2013). Going beyond previous quantitative evaluations of these models, which have focused on their ability to account for the response proportions of empirical data or qualitative benchmarks (Trueblood et al., 2014; Turner et al., 2018), we compare them on their ability to simultaneously account for response choices as well the full distribution of response times. Importantly, the results of our study show that the focus on response proportions has been a key shortcoming of the previous literature, and the additional constraint of the response-time distributions allow for a clear distinction between the models in their ability to account for the context effects. To ensure robustness of our conclusions, we test these theories/models using a broad range of empirical data, covering 12 total experiments from six previously published studies.¹

The three context effects describe how preferences between two alternatives can change with the introduction of a third new alternative. They differ in the location of the third alternative within attribute space and how it influences preferences between the original two options. The attraction effect occurs when the third alternative is similar to but objectively inferior to an existing alternative, resulting in an increased preference for the now dominating alternative over the other existing alternative (see Fig. 1, alternatives D_A and D_B). Consider the laptop example again. Suppose two explicit attributes of each alternative are the battery life and the processing speed, where laptop A has a long battery life, but a slow processor, while laptop B has a faster processor, but a shorter battery life. The attraction effect would involve the introduction of a new laptop (e.g., D_A in the figure), which has an equal battery life to laptop A, but has a slower processing speed. The introduction of such a laptop would result in an increased preference

Fig. 1 A pictorial depiction of the three key context effects—the attraction, similarity, and compromise effects—for the laptops example discussed in the main text. The *x*-axis provides the attribute of processing speed, with greater distances from the origin indicating greater processing speeds. The *y*-axis provides the same for the attribute of battery life. Within the plot, the black letters, A and B, provide the two initial alternatives, where A has a longer battery life than B, and B has a faster processing speed than A. The *broken grey line* provides the “line of indifference”, where all alternatives placed along this line have an equal overall objective value within the attribute space. The *red letters*, D_A and D_B , provide examples of the attraction effect. The *blue letters*, C_A and C_B , provide examples of the compromise effect. The *green letters* S_A and S_B , provide examples of the similarity effect

for laptop A over laptop B. The similarity effect occurs when the third alternative is similar to one of the existing alternatives, while being objectively equal to it, resulting in an increased preference for the dissimilar alternative over the similar alternative (see Fig. 1, alternatives S_A and S_B). The compromise effect occurs when the third alternative is objectively equal to the other alternatives, but is sufficiently extreme in its attribute values to turn one of the existing alternatives into a compromise between the other two, resulting in an increased preference for the compromise alternative (see Fig. 1, alternatives C_A and C_B).

These three context effects (along with other effects, such as the phantom decoy, see Trueblood & Pettibone, 2017) violate rational decision models (although see Howes, Warren, Farmer, El-Deredy, & Lewis, 2016 for a possible rational account) and have lead to the development of more complex models of multi-attribute decision-making such as MDFT, LCA, MLBA, and AAM (Usher & McClelland, 2004; Roe et al., 2001; Trueblood et al., 2014; Bhatia, 2013). Although these models add complexity beyond that of the simple utility framework, they remain quite functionally constrained by the data, as the preference for each alternative is informed by stimuli values (i.e., the attribute values of the options). Theoretically, this means that the models should be able to account

¹ It should also be noted that we attempted to obtain the data from one further study, noted in our “Experiments” section, but these data were never provided by the original author.

for multiple context effects with only a single set of values for their free parameters (Tsetsos, Chater, & Usher, 2015, though the simultaneous occurrence of all three effects in humans is questioned, see Trueblood et al., 2015; Liew, Howe, & Little, 2016). Although these models each share some underlying components, each encodes different hypothesized mechanisms to explain the three context effects (Tsetsos et al., 2010, 2015; Hotaling et al., 2010; see the “Models” section for details).

Thus far, very little research has contrasted the ability of these models to account for empirical data. For the most part, previous research has assessed each of the models in isolation, and has focused on exploring the range of parameter values that can qualitatively produce the three context effects (Tsetsos et al., 2010; Hotaling et al., 2010). While recent efforts have been made to carry out such comparison (Berkowitsch et al., 2014; Trueblood et al., 2014; Cohen, Kang, & Leise, 2017; Turner et al., 2018), these assessments have several limitations. Firstly, many assessments focus purely on whether the models can qualitatively account for the three effects (Usher & McClelland, 2004; Usher et al., 2008). However, the ability of the models to predict the existence of the effects (i.e., the correct response ordering between the alternatives) does not necessarily indicate their ability to successfully account for the precise response proportions seen within empirical data, or the relationship between the effects. For example, Trueblood et al. (2015) suggested that the standard model benchmark of producing all three effects with a single set of parameters is misleading since very few participants (approximately 23%) produced all three effects within a single experiment.

Secondly, every assessment to date has focused exclusively on response proportions, despite the fact that all four models were explicitly developed as dynamic models that predict both choice and response-time distributions (though see Cohen et al., 2017 for a comparison of MLBA to expected utility models in the prediction of response times for non-context effects data). Importantly, it is well known that relationships exist between choices and their associated response times. In particular, previous research has shown that the magnitude of three context effects increases with deliberation time (Pettibone, 2012). In addition, these assessments have exclusively used an “external stopping rule” for MDFT, the LCA, and the AAM, where the time at which the decision is made remains constant and is controlled by the experimenter, rather than an “internal stopping rule”, where the participant is free to decide at any time (Trueblood et al., 2014; Cohen et al., 2017; Turner et al., 2018). However, the use of an external stopping rule within the models has been inconsistent with the experimental designs used, where the decision time has been controlled by participants, rather than the experimenter. Although the

use of an external stopping rule can serve as a useful computational simplification for these models, allowing for a tractable likelihood function when fitting response proportions (e.g., see Roe et al., 2001), the inconsistencies between the applications of these models and the experimental designs may lead to misleading results regarding how well the models can account for empirical data.

Our study aims to provide a comprehensive comparison between four prominent models of multi-attribute choice—MDFT, LCA, MLBA, AAM—on empirical choice response-time distributions of the three context effects that they were designed to account for, across 12 total experiments from six previously published studies. These studies cover a broad range of data, including both within and between subjects manipulations of the three effects across domains ranging from perceptual to risky decision-making in both humans and non-human primates.

Our study also improves upon some of the key limitations present within previous assessments of the models. Firstly, and most importantly, our study provides a much greater constraint on the predictions of the models than previous research by fitting them to the choice response-time distributions. That is, we are fitting them to the type of data they were designed to predict. As our results later show, this provides a key source of distinction between the models, especially with regard to the compromise and similarity effects. As no analytic solution for the likelihood function exists for the choice response-time distributions of MDFT, LCA, or AAM, we fit these models using state-of-the-art probability density approximation techniques (PDA), which allow a synthetic likelihood to be generated for a model using simulated data (see Holmes, 2015; Turner & Sederberg, 2014 for the theoretical details and analysis of this method and Miletić, Turner, Forstmann, & van Maanen, 2017; Holmes & Trueblood, 2018; Holmes et al., 2016; Trueblood et al., 2018 for example applications). Secondly, we assess the models based upon the data of individual participants, avoiding the potential averaging issues associated with only assessing group-averaged data (Estes, 1956; Heathcote, Brown, & Mewhort, 2000; Evans, Brown, Mewhort, & Heathcote, 2018). In order to do this, we fit the models using Bayesian hierarchical modeling, which fits the model to participants’ individual data while constraining the parameters to follow some group-level distribution. This combination of diverse, independent sources of data coupled with state-of-the-art response-time modeling methods makes this study the most systematic and wide-ranging comparison of these theories carried out to date.

Models and methods

In this section, we provide a general conceptual and mathematical description of all models, with further specifics

in the appendices. All of these models belong to the broader framework known as “evidence accumulation models” (see Ratcliff, Smith, Brown, & McKoon, 2016 for a review, and Ho, Yang, Wu, Cassey, Brown, Hoang, & Yang, 2014; Evans & Brown, 2017; van Ravenzwaaij, Dutilh, & Wagenmakers, 2012; Dutilh, Annis, Brown, Cassey, Evans, Grasman, & Donkin, 2018; Evans, Hawkins, Boehm, Wagenmakers, & Brown, 2017a, Evans, Rae, Bushmakin, Rubin, & Brown, 2017c, 2018 for applications). These models propose that decisions are made through a process where evidence accumulates over time at some rate (known as the “drift rate”) until a decision criterion is met. Given the type of experimental data available, we will focus on variants of these models utilizing an internal stopping rule criterion, where evidence accumulates until the amount of evidence for one of the alternatives crosses some threshold level of evidence (known as the “decision threshold”), triggering a decision.

Our study makes some important alterations to the functional forms of MDFT, LCA, and AAM. Previous studies applied these three theories as “random walk” models for computational simplicity (Roe et al., 2001; Tsetsos et al., 2010; Usher et al., 2008; Usher & McClelland, 2004, though see Busemeyer & Townsend, 1992 and Busemeyer & Townsend, 1993 for descriptions of DFT as an Ornstein–Uhlenbeck process, and Busemeyer & Diederich 2002, and Huang et al., 2012 for descriptions of MDFT as a multi-dimensional diffusion model, though all without application), and studied choice proportions only (Tsetsos et al., 2010, 2015; Trueblood et al., 2014; Berkowitsch et al., 2014; Bhatia, 2013; Turner et al., 2018). Here, we will be fitting these models to choice and response-time (RT) values. We thus convert them to a “stochastic differential equation” (SDE) formalism, an analogue of a random walk that explicitly takes into account the timescale of accumulation dynamics (see Appendix A for details of the conversion). Although random walks can be made to approximate stochastic differential equations using certain methods (e.g., treating steps as exponentially distributed random variables, see Nosofsky & Palmeri, 2015), we felt the use of SDEs was the more natural choice given their common usage within the RT modeling literature (e.g., the diffusion model, Ratcliff, 1978). In the following sub-sections, we describe the mathematical details of each model and augmentations that we made to each model to facilitate RT modeling.

Generally, the preference for the vector of alternatives P (e.g., for the choice alternatives X, Y, and Z, $P = [P_X, P_Y, P_Z]'$) obeys a (potentially nonlinear) differential equation with the form:

$$d\vec{P} = \vec{\mu} dT + \sigma d\vec{W}, \quad (1)$$

where the first term represents deterministic aspects of accumulation based on evidence with rate $\vec{\mu}$, and the second term is a standard Brownian noise/diffusion term accounting

within trial variability. While this framework more naturally describes MDFT, LCA and AAM (which have within trial noise), MLBA can be described using the same framework, but with no within trial variability (i.e., $\sigma = 0$). In the following sections, we will outline how the components of each model are incorporated into this framework, primarily by specifying the dependencies of the drift rate vector ($\vec{\mu}$) for each model.

Additionally, we made a minor change to all models (MLBA included) to make them dimensionally well posed. Specifically, we added a scaling parameter (γ) to the incoming evidence for each alternative on each time step. While this adjustment is not necessary for fitting choice data, the models are mathematically mis-specified without it and (based on previous fitting), we have found it necessary to account for response-time distributions. Critically, we made the same adjustment to all four models so as not to bias the comparison process.

Before assessing these models in their ability to account for empirical data, we firstly performed a recovery assessment to determine the reliability of inferences from these models and methods. This is a necessary step to have confidence in any inferences (parametric or otherwise) from the results of model fits. Two important reliability benchmarks are assessed to establish whether a model is able to (1) recover its own data (i.e., data recovery) and (2) whether it can recover its own parameters (i.e., parameter recovery). In the former, data are generated out of model $M(P)$ (MDFT for example) with parameters P and the model is fit to the resulting data to determine if it can converge to a set of parameters or region of parameter space that matches the data. In the second, those fits are used to determine if the fitting procedure converges to the specific parameters P that generated the synthetic data. All models performed very well at data recovery, thus giving us confidence to proceed with quantitative fitting. However, the models had difficulty recovering specific parameter values. This is not surprising given previous parameter recovery assessments of LCA (Miletić et al., 2017) that showed strong trade-offs between several of its parameters. In general, complex models in a number of scientific domains are well known to have issues with parameter identifiability (termed “sloppy models” or “weakly identifiable models”, Holmes, 2015; Gutenkunst, Waterfall, Casey, Brown, Myers, & Sethna, 2007).

The consequences of this are twofold. First, the models can be used to determine whether their underlying theories account for experimental data, which will be our focus here. Given two weakly identifiable/sloppy models, one of which fits data well and one of which does not, the well-fitting model provides the more plausible theory (based on that data alone of course). These models cannot however be used for parameter inference, as it is not possible to reliably

interpret the specific values of any of their parameters. Thus, unfortunately, it is not possible at this stage to use specific parameter values to investigate underlying cognitive processes. However, this does not mean that the models cannot provide important insights about specific cognitive mechanisms. In order to probe specific processes, we use a knock-out/add-in modeling approach. The basic idea is to remove or add specific model components one at a time in order to understand their impact on the model's descriptive adequacy. If knocking-out a specific mechanism leads to severely degraded performance, then we can conclude that the mechanism is a critical part of the model. In our results section below, we use this type of approach to gain more detailed insight into the models' performance without resorting to parameter-based inferences.

MDFT

MDFT is comprised of the following critical elements: (1) leaky evidence accumulation, (2) mutual inhibition between alternatives, with this inhibition being some function of the distance between the alternatives on a unit plane of the attribute values ("closer" alternatives inhibiting each other more), and (3) a drift rate for each alternative that is based on the values of a single attribute, with attention switching between attributes over time. The latter two of these are crucial elements of the model that are necessary to explain the context effects.

The first of these can be described by a simple Ornstein–Uhlenbeck decay term in the drift rate. To describe the second, we use the same distance function reported in Hotelling et al. (2010) (which is mathematically identical to that of Berkowitsch et al., 2015 in the case of two attributes). A third of these requires a new assumption to be included in the model. In its original random walk formulation, MDFT assumed that the attribute on which evidence accumulation was based could switch at every step of the random walk. Using this previous mechanism in a stochastic differential equation creates two issues. Firstly, it would be akin to assuming that attention switches very rapidly (every few milliseconds). Secondly, it would assume that the duration of attention on an attribute is time step size-dependent. This could lead to a mathematical problem where simulation results are time step size-dependent (e.g., 1-ms and 5-ms time steps could lead to different results). Instead of assuming rapid attention switching at every time step, we assume the duration of time spent attending to an attribute is exponentially distributed. Thus, the probability of switching on each step is equal to $1 - \exp(-k \times \Delta t)$, where k is a free rate parameter of the model, and Δt is the time step size. We also made one final alteration, reducing the overly complex noise term present within previously versions of MDFT (which depended on irrelevant attributes

or other factors) with the simpler Wiener process. While these represent significant mathematical changes to the original model formulation, this model faithfully includes the core assumptions (1–3) of MDFT listed above (personal communication with Jerome Busemeyer).

Formally, these assumptions are encoded in the following MDFT drift rate function

$$\mu = [S \times P(t) + V(t)] \quad (2)$$

where S is an $n \times n$ matrix (where n is the number of alternatives) whose diagonal elements encode the rate of evidence leakage and off diagonal elements describe the distance-dependent lateral inhibition between alternatives. $P(t)$ is the preference for the vector of alternatives at the time t , and $V(t)$ is the vector of valences for each alternative at time t based upon the attribute being attended to.

Overall, our definition of MDFT gives the model eight free parameters: a (the decision threshold), k_A (the attention duration parameter for attribute 1), k_B (the attention duration parameter for attribute 2), ϕ_1 (a parameter of the Gaussian distance function, which controls the overall strength of the lateral inhibition), ϕ_2 (a parameter of the Gaussian distance function, which controls the amount of evidence leakage), β (the multiplier applied to the dominating dimension of the unit plane), σ (the standard deviation of the Wiener process), and γ (a multiplier that scales the valance vector). The full details of the model can be found in Appendix B.

Two key aspects of this model are required to explain context effects: the "distance-dependent" inhibition between alternatives, and attention-based switching process (Roe et al., 2001; Hotelling et al., 2010; Tsetsos et al., 2010). Specifically, the attraction effect is explained by the decoy providing "negative inhibition" to the dominating alternative, boosting it above the dissimilar alternative. The similarity effect is explained by the "correlation" in evidence accumulation between the similar alternatives created by attention switching across attributes. This results in the dissimilar alternative being chosen more often, as the similar alternatives share the choices where their strongest attribute gains the most attention. The compromise effect is explained by the "correlation" in evidence accumulation between the extreme alternatives created by each having an "inhibitory link" with the compromise alternative, but not each other. This results in the compromise alternative being chosen more often, as the extreme alternatives share the choices where they successfully inhibit the compromise alternative.

LCA

LCA is comprised of the following critical elements: (1) leaky evidence accumulation, (2) mutual inhibition between alternatives (though unlike in MDFT, this inhibition is

unrelated to the distance between the alternatives), (3) a drift rate for each alternative that is based on the values of a single attribute, where the attribute being attended to changes over time, and (4) a process of ‘loss aversion’ incorporated into the drift rates, where differences between attribute values that are negative (i.e., where an alternative is worse on the attribute) are weighted more heavily than differences that are positive. Attention switching and loss aversion are particularly critical to this model’s ability to account for context effects.

Here we describe an augmented form of the LCA model posed as a SDE. We note that this is not a direct translation of the original random walk version of the LCA (Usher & McClelland, 2004). In converting the original version to an SDE, we found a parameter degeneracy issue associated with the manner in which leakage was incorporated. We thus make a minor adjustment to the way in which leakage is described to produce a model suitable for RT/choice investigation (e.g., the single attribute LCA of Usher & McClelland, 2001). Loss aversion and inhibition are accounted for in the same way as the original random walk version (Usher & McClelland, 2004; Tsetsos et al., 2010); only the treatment of leakage is adjusted. For further details, see Appendix C.

As with MDFT, leakage can be readily accounted for with a linear decay term in the drift rate function. Similarly, we treat attention switching in LCA just as was done in MDFT; the time spent attending to an attribute is exponentially distributed. Formally, the drift rate for the LCA is defined as:

$$\mu = [S \times A(t) + I(t)] \quad (3)$$

where S is an $n \times n$ matrix (where n is the number of alternatives) that contains both the amount of lateral inhibition between the alternatives and the rate of evidence leakage for each alternative. $A(t)$ is the activation for the vector of alternatives at the time t , $I(t)$ is the vector of inputs for each alternative at time t based upon the attribute being attended to, and σ is the standard deviation of the noise process, which is a free parameter of the model.

Overall, our definition of the LCA gives the model eight free parameters: a (the decision threshold), k_F (the attention duration parameter for attribute 1), k_G (the attention duration parameter for attribute 2), I_0 (the baseline level of activation), λ (the amount of evidence leakage), β (the amount of global inhibition), σ (the standard deviation of the Wiener process), and γ (a multiplier that scales the activation vector). The full details of the model can be found in Appendix C.

The LCA also relies on two key parts of its functional form to explain the context effects: the attention switching process and loss aversion (Usher & McClelland, 2004; Usher et al., 2008; Tsetsos et al., 2010). Like MDFT, the

LCA explains the similarity effect through the ‘‘correlation’’ in evidence accumulation between the similar alternatives, which split their winnings. In contrast to MDFT, the LCA explains both the attraction and compromise effects through its built-in process of loss aversion, where people aim to avoid selecting alternatives that have large losses attached to them. Specifically, the attraction effect is the result of the similar alternatives (i.e., the target and decoy) providing the dissimilar alternative (i.e., the competitor) with a large loss, and one of the similar alternatives (i.e., the target) being objectively dominant. The compromise effect is the result of each extreme alternative providing the other with a large loss, and the compromise alternative only having smaller losses.

MLBA

The MLBA models the preference of different alternatives of multi-attribute choice through a process of independent, leakless, noiseless evidence accumulation. Unlike MDFT and LCA, MLBA contains no leakage, no within-trial noise ($\sigma = 0$), and no attention switching. Drift rates for all alternatives depend jointly on all attributes and are fixed within a trial, though they are assumed to vary across trials (as usual with LBA models, Brown & Heathcote, 2008). The key to this model is in how those drift rates depend on the characteristics of the alternatives. The key elements of this dependence are: (1) the drift rate for each alternative is a weighted sum of pairwise comparisons among alternatives, (2) objective attribute values are transformed by a non-linear function into subjective values, (3) each pairwise comparison is weighted by the similarity of the options, modeled as an exponential decay of the distance between attribute values (i.e., similar options receive more attention), and (4) different decay functions exist for positive and negative differences, allowing either negative differences to be weighted more heavily (reflecting loss aversion) or positive differences to be weighted more heavily (possibly reflecting a confirmation bias). The similarity-based attention weights and the asymmetry in these weights (i.e., positive and negative differences are treated differently) are particularly critical in accounting for context effects.

Formally, the drift rate (μ) for the MLBA is defined as:

$$\mu \sim TN(d, s, 0, Inf) \quad (4)$$

$$d_i = I_0 + \sum_{j \neq i}^n V_{i,j} \quad (5)$$

where μ is defined by a distribution due to the between-trial variability in drift rate, TN is the normal distribution truncated to positive numbers (0 to positive infinity), d is the mean drift rate that is determined by a transformation of

the attribute values, d_i is the drift rate for each alternative determined by a baseline drift rate (I_0) and the sum of the pairwise comparisons between that alternative and each of the other alternatives ($V_{i,j}$), and s is the standard deviation of the drift rate distribution. As mentioned earlier, the model has no within-trial noise, meaning that it is the equivalent of the general stochastic differential equation form with σ fixed to 0.

Overall, our definition of the MLBA gives the model nine free parameters. There are three standard LBA parameters: b (the decision threshold), A (the top of the uniform distribution of random starting points of evidence), and t_0 (the time attributed to non-decision related processes). For the front-end process that translates attribute information into drift rates, there are six: I_0 (the baseline level of mean drift rate), λ_1 (the decay parameter for positive distances), λ_2 (the decay parameter for negative distances), β (attribute importance weight), m (controls the transformation of the objective values to subjective values), and γ (a multiplier that scales the mean drift rates). The full details of the model can be found in Appendix D.

The MLBA relies on three parts of its functional form to explain the context effects: similarity-based attention, asymmetric attention for positive and negative comparisons, and subjective attribute values (Trueblood et al., 2014; Tsetsos et al., 2015). The attraction effect is explained by similarity-based attention, as more similar alternatives (i.e., target and decoy) are given more attention, and one of the similar alternatives (i.e., target) is objectively superior. The similarity effect is explained by asymmetry in the attention weights where there is more weight on positive comparisons as compared to negative comparisons. Like the attraction effect, the compromise effect is explained by similarity-based attention, as the compromise option is more similar to the two extremes than the two extremes are to each other. The compromise effect can also be enhanced by the subjective attribute function, which can weight central values more than extremes (i.e., extremeness aversion).

AAM

AAM is comprised of the following critical elements: (1) leaky evidence accumulation, (2) mutual inhibition between alternatives (though unlike in MDFT, this inhibition is unrelated to the distance between the alternatives), (3) a drift rate for each alternative that is based on the values of a single attribute, where the attribute being attended to changes over time based on the attributes' relative levels of "activation", (4) subjective attribute values that are a non-linear transformation of the objective values. Attention switching and the level of activation are particularly critical to this model's ability to account for context effects.

For AAM, we treat leakage and attention switching in the same manner as MDFT and the LCA, though the specific parameters governing the attention switch time were slightly different to allow the integration of the attribute activation levels (see Appendix E for more details). Formally, the drift rate for the AAM is defined as:

$$\mu = [S \times A(t) + I(t)] \quad (6)$$

where S is an $n \times n$ matrix (where n is the number of alternatives) that contains both the amount of lateral inhibition between the alternatives and the rate of evidence leakage for each alternative. $A(t)$ is the activation for the vector of alternatives at the time t , $I(t)$ is the vector of inputs for each alternative at time t based upon the attribute being attended to, and σ is the standard deviation of the noise process, which is a free parameter of the model.

Overall, our definition of the AAM gives the model nine free parameters: a (the decision threshold), k_F (the attention duration parameter for attribute 1), k_G (the attention duration parameter for attribute 2), k_{scale} (the scaling parameter for the switching durations to allow for the attribute activations used within AAM), α (the parameter of the subjective attribute function), λ (the amount of evidence leakage), β (the amount of global inhibition), σ (the standard deviation of the Wiener process), and γ (a multiplier that scales the activation vector). The full details of the model can be found in Appendix E.

The AAM also relies on two key parts of its functional form to explain the context effects: the attention switching process, and the relative activation of the different attributes (Bhatia, 2013). Like MDFT and the LCA, the AAM explains the similarity effect through the "correlation" in evidence accumulation between the similar alternatives, which split their winnings. In contrast, AAM explains the attraction and compromise effects through the level of activation of each attribute, which determines how much attention is paid to the attribute. For both of these effects, the addition of the new alternative increases the activation for the attribute that the target is strongest on, increasing its response proportion relative to the competitor.

Bayesian hierarchical modeling

In this paper, we fit these models to the data from six separate studies. For five of those, we use hierarchical Bayesian methods. Standard methodologies fit the free parameters of models to each individual separately, meaning that the parameters of one individual are completely independent of another. The hierarchical structure places extra constraints on the models, assuming that each person's parameters follow some group-level distribution, creating some

dependence between the parameter values by assuming that individuals are members of a population with some structure. We estimate both the group-level and individual-level parameters for hierarchical extensions of each of these models. The benefit of this approach is that it provides information about the uncertainty in the parameter values by providing a distribution (known as the “posterior”) of values for the parameter, rather than just the single, point-estimate value provided by other methods (e.g., maximum likelihood).

To estimate the posterior distributions for these experiments, we used differential evolution Markov chain Monte Carlo (DE-MCMC; Ter Braak, 2006; Turner et al., 2013), with $3k$ parallel chains, where k is the number of free parameters per participant, 2000 iterations of burn-in followed by 1500 samples from the posterior per chain. As no analytic solution currently exists for the likelihood function for MDFT, LCA, and AAM under an internal stopping rule, we used the recently developed probability density approximation (PDA) method to create approximations to the likelihood function that can be used in the MCMC framework (Holmes, 2015). This involves: (1) generating a large number (10,000) of synthetic trials from the model for the parameter values currently being evaluated, (2) using those samples to generate an approximate density function, and (3) using that approximate density function to calculate a log likelihood. The hierarchical structure for each model can be found in Appendix F.

For one of the studies Farmer et al. (2016), we used maximum likelihood estimation in place of the Bayesian hierarchical modeling, due to this study containing a large number of experimental conditions. The computational burden of any fitting method grows with the complexity of the data set, which is particularly true of simulation-based Bayesian methods, since Bayesian methods require sampling large numbers of parameters to generate distributions and PDA requires an independent set of simulations for every experimental condition. Given our finite computational resources, hierarchical Bayesian estimation was not tractable for this data set. Instead, we used maximum likelihood estimation to find the best parameter values, with a differential evolution optimizer that used $5k$ parallel particles, where k is the number of free parameters, and 500 iterations. Starting points were randomly generated from a uniform distribution spanning the parameter boundaries, which can be found in Appendix G.

Experiments

We examine 12 experiments across six studies, spanning context effects in human participants for perceptual decision-making, criminal inference, and risky decision-making, as well as perceptual decision-making for non-human primate participants. In this section, we will briefly

describe each of these six studies, and the experiments that they consist of, and refer the interested reader to the appropriate articles for further details. It should also be noted that we attempted to include data from two other studies, Liew et al. (2016) and Berkowitsch et al. (2014); however, we never received the data from the original authors of Liew et al. (2016). Berkowitsch et al. (2014) provided their choice data, but they did not record response times and thus it could not be included.

Trueblood et al. 2013 (E1)

We assessed three experiments from Trueblood et al. (2013): an attraction effect experiment (E1a; 53 participants, six relevant conditions, 90 trials per condition), a compromise effect experiment (E1b; 63 participants, two relevant conditions, 180 trials per condition), and a similarity effect experiment (E1c; 62 participants, four relevant conditions, 135 trials per condition). These experiments all used human participants with perceptual stimuli, with participants having to judge which rectangle was the largest from three alternatives that differed on height and width. The key finding of this study was that the three context effects generalize from high-level decision-making to perceptual stimuli.

Trueblood et al. 2015 (E2)

We assessed one experiment from Trueblood et al. (2015), which manipulated all three effects as different within-subject conditions (75 participants, six relevant conditions, 80 trials per condition). This experiment used human participants with perceptual stimuli similar to Trueblood et al. (2013). The key finding of this study was that the three context effects can be quite fragile, with only 23.6% of participants displaying all three context effects.

Trueblood et al. 2012 (E3)

Our study assessed three experiments from Trueblood (2012): an attraction effect experiment (E3a; 47 participants, six relevant conditions, 20 trials per condition), a compromise effect experiment (E3b; 52 participants, three relevant conditions, 40 trials per condition), and a similarity effect experiment (E3c; 51 participants, four relevant conditions, 30 trials per condition). These experiments all used human participants with a “criminal inference” paradigm, where participants were presented with three criminal suspects and had to choose which one was guilty of having committed the crime, based upon two differing eye-witnesses ratings. This study was the first to establish that all three context effects can occur within a non-value-based experimental paradigm (as opposed to the standard consumer choice paradigm).

Trueblood et al. 2014 (E4)

Our study assessed one experiment from Trueblood et al. (2014), which manipulated all three effects as different within-subjects conditions (68 participants, eight relevant conditions, 10–20 trials per condition). This experiment used human participants with the same criminal inference paradigm as Trueblood (2012). This experiment was used within the original study as a comparison data set between MLBA and MDFT on their ability to predict group-level summary statistics, finding the MLBA to be superior.

Farmer et al. 2016 (E5)

Our study assessed three experiments (Experiments 2a, 2b, and 2c) from Farmer et al. (2016). All were attraction effect experiments with human participants, though the paradigm differed between experiments: risky decision-making through perceptual stimuli (E5a; 50 participants), risky decision-making through standard text presentation (E5b; 52 participants), and perceptual decision-making similar to Trueblood et al. (2013) (E5c; 41 participants). Each experiment contained 32 relevant conditions, with eight trials per condition. The key finding of this study was that the attraction effect occurs in situations where there is an objectively correct answer (e.g., one option has a higher expected value in the risky choice tasks).

Parrish et al. 2015 (E6)

Our study assessed one experiment from Parrish et al. (2015), which manipulated the attraction effect for non-human primate participants with perceptual stimuli similar to Trueblood et al. (2013) (seven participants, 32 relevant conditions, 18–30 trials per condition). The key finding of this study was that the attraction effect also occurs within monkeys.

Results

Firstly, we note that although some participant exclusions were applied to some of the original analysis of these experiments, we chose not to remove any participants from our analysis. Generally, the results appeared to show two very clear trends, which we will briefly summarize below. We then provide a more detailed description of the results from the experiments of each study. Lastly, we performed a variety of additional model fits using the knock-out/add-in approach, as well as an assessment of the models in their ability to explain only response proportions (i.e., ignoring the response-time data). These results indicate that the ability of the models to explain certain effects can be

linked to specific parts of their function, and that the use of the entire choice response-time distributions is a key factor in distinguishing the models and their predictions. Additionally, details on the fits as well as the additional variants mentioned in the text that were fit can all be found in the [Supplementary Materials](#).

When assessing the attraction effect, all four models appear to do a good, and essentially indistinguishable, job of explaining the empirical data. An example of this can be seen in Figs. 2 and 3. Figure 2 displays the choice proportions in the empirical data (*x*-axis of each panel) versus the model predictions (*y*-axis), with a single data point indicating a single condition for a single participant. Figure 3 displays the quantiles of the response-time distribution in the empirical data (*x*-axis of each panel) versus the model predictions (*y*-axis). In each figure, each row of plots indicates a different experiment, and each column indicates the fit for a different model. These results demonstrate that all three models provide an equally good accounting of that attraction effect.

However, the ability of the models to predict the empirical data appears to be clearly differentiated when assessing the compromise and similarity effects (Figs. 2 and 3). When assessing the response-time quantiles, each model provides a similarly accurate match to the empirical data. However, they differ significantly in their ability to account for choice proportions. MLBA fits the choice proportion data well while MDFT, the LCA, and the AAM do not. In particular, these three models predict the majority of participants to have identical response proportions across conditions: a trend that is not supported by the data. Overall, it appears that the MLBA is able to provide a good account of all three effects observed within empirical data, whereas MDFT, the LCA, and AAM only provide a good account of the attraction effect. This is also seen in the DIC (Spiegelhalter et al., 2002) results,² which are included in the respective figures for the fits to each experiment. MLBA has substantially lower DIC values than MDFT and LCA for seven of the nine experiments where Bayesian fits were performed. LCA has lower values for the remaining two. Critically, those two experiments are attraction-affect-only experiments where all models perform well, and the DIC differences are much smaller in these two cases. This reinforces the observation from figures that all models perform similarly on attraction data (with LCA performing a little better according to DIC), but that similarity and compromise data distinguish MLBA from the others.

²It should be noted that DIC is not a well-behaved test statistic in cases where models are weakly identifiable/sloppy. We have, however, included this information by request.

Fig. 2 Probability–probability plots for the empirical response proportions (x-axis of each panel) and the model predicted response proportions (y-axis of each panel) of experiments E1 and E2. The rows of panels give the different experiments, being attraction only (ATT; E1a), compromise only (COM; E1b), similarity only (SIM; E1c), and

all effects together (ALL; E2). The columns of panels give the different models, being the MLBA, MDFT, the LCA, and the AAM. Different symbols represent different experimental conditions. The colors of the symbols denote different types of alternatives: green represents the target, blue represents the competitor, and red represents the decoy

Trueblood et al. 2013 (E1)

All of the models use stimuli values as inputs. In the original experiment, different trials in the same condition had slightly different attribute values. We used the average attribute values for each condition as the inputs to the

models. As all of the models assume an additive relationship between attribute values, and the rectangle area is a multiplicative relationship of height and width, we also took the logarithm of the attribute values to place them on the correct scale. For MDFT, the LCA, and the AAM, the time step used to simulate the models was set at 20 ms,

Fig. 3 Quantile–quantile plots for the empirical response times (x-axis of each panel) and the model predicted response times (y-axis of each panel) of experiments E1 and E2. The rows of panels give the different experiments, being attraction only (ATT; E1a), compromise only (COM; E1b), similarity only (SIM; E1c), and all effects together (ALL; E2). The columns of panels give the different models, being

the MLBA, MDFT, the LCA, and the AAM. Larger points indicate a greater probability of a response in the empirical data, meaning that the quantile estimates are based upon a greater amount of data. Different symbols represent different experimental conditions. The colors of the symbols denote different types of alternatives: green represents the target, blue represents the competitor, and red represents the decoy

though we also re-fit MDFT and the LCA with a time step of 5 ms and found no qualitative difference in results. It also should be noted that we chose to exclude trials with extremely long response times from our analysis, with our

criterion being 7 s. We also fit these data with the LCA that contained the degenerate leakage term used previously in the literature (Usher & McClelland, 2004; Tsetsos et al., 2010) and found no qualitative difference in results. These

additional analyses can be found in the [Supplementary Materials](#).

A summary of the results for fits of the models to these experiments can be seen in Figs. 2 and 3 for the response proportions and the response-time distribution quantiles, respectively. As can be seen in Fig. 3, all models appear to provide a good account of the response-time quantiles for all of the experiments. Although there appears to be some minor misfit for each of the models, they largely provide predictions that are very close to the empirical data, especially for the data points with the higher response proportions. The largest sources of misfit appear to be in some of the quantiles where the associated response had an extremely low response proportion in the empirical data, especially in the MLBA, which over-predicts the response times in these quantiles. However, as these quantiles make up a very small portion of the data, this provides little separation between the models.

As can be seen in Fig. 2, there appears to be some clear distinctions between the models in their ability to successfully predict the empirical response proportions. For the attraction experiment, all models appear to fit the empirical response proportions quite well, with very little misfit. However, in the compromise and similarity experiments, this is no longer the case. The MLBA is able to provide a good fit to both of these experiments, accurately accounting for the full range of response proportions of each participant over conditions. MDFT, LCA, and AAM, on the other hand appear to have major issues capturing the data in these experiments, with significant clustering on the horizontal axis. This suggests an inability of the models to correctly predict the change in response proportions across conditions for a majority of subjects, resulting in the models predicting all responses to have very similar response proportions centered on the average response proportion.

Trueblood et al. 2015 (E2)

The attribute values used to constrain the models were calculated in the same way as E1, as were the response-time exclusions applied to trials, as well as the time step used for MDFT, the LCA, and the AAM. A summary of the results for fits of the models to these experiments can be seen in Figs. 2 and 3, for the response proportions and the response-time distribution quantiles, respectively. As this experiment is a within-subjects version of E1, the results unsurprisingly appear to be a combination of those seen within E1, though the fits in general are slightly worse due to the increased constraint applied to the models. Again, all models appear to provide a good account of the response-time quantiles, and do so to a very similar extent. However, as with E1, the MLBA provides a good fit to all of the response proportions, whereas the MDFT, LCA, and AAM

predict some data well (the attraction conditions), but have a large amount of clustering on the horizontal axis, again suggesting an inability to correctly predict the change in response proportions over conditions and individuals.

Trueblood et al. 2012 (E3)

Similar to the previous experiments, we used the average attribute values for each condition as the inputs to the models. Although the attribute values for this experiment already contained an additive relationship, we transformed them by dividing them by 10, in order to put them on a similar scale to the values from the last experiments, and as was done in the original paper. For MDFT, the LCA and the AAM, the time step used to simulate the models was set at 100 ms. Although this may seem like a long time step, its increase from E1 and E2 is proportional to the increase in the standard deviation of response times. It also should be noted that we chose to exclude trials with extremely long response times from our analysis, with our criterion being 40 s.

A summary of the results can be seen in Figs. 4 (response-time quantiles) and 5 (response proportions), which take the same format as Figs. 2 and 3, respectively. As in the perceptual decision-making experiments of E1, all models appear to provide a good account of the response-time quantiles for all of the experiments, with only minor misfit, which mostly occurs on quantiles with a very small proportion of responses associated with them. The results also seem to be consistent between these experiments and E1 for the response proportions: all models appear to do a good job of accounting for the empirical response proportions in the attraction experiment, but only the MLBA is able to accurately predict the response proportions of the compromise and similarity experiment. Again, MDFT, the LCA, and the AAM have major clustering on the horizontal axis, suggesting an ability of the models to predict the change in response proportions over conditions. In general, the results of this experiment appear to be nearly identical to that of E1, with the only difference being that the overall level of misfit appears to be greater, likely due to the much lower number of trials per condition.

Trueblood et al. 2014 (E4)

The attribute values used to constrain the models were calculated in the same way as E3, as were the response-time exclusions applied to trials, as well as the time step used for MDFT, the LCA, and the AAM. A summary of the results for fits of the models to these experiments can be seen in Figs. 4 and 5, for the response proportions and the response-time distribution quantiles, respectively. As this experiment is merely a within-subjects version of E3, the results unsurprisingly appear to be a combination of

Fig. 4 Probability–probability plots for the empirical response proportions (x-axis of each panel) and the model predicted response proportions (y-axis of each panel) of experiments E3 and E4. The rows of panels give the different experiments, being attraction only (ATT; E3a), compromise only (COM; E3b), similarity only (SIM; E3c), and

all effects together (ALL; E4). The columns of panels give the different models, being the MLBA, MDFT, the LCA, and the AAM. Different symbols represent different experimental conditions. The colors of the symbols denote different types of alternatives: green represents the target, blue represents the competitor, and red represents the decoy

those seen within E3, though the fits in general are slightly worse due to the increased constraint applied to the models. Again, all models appear to provide a good account of the response-time quantiles, and do so to a very similar extent. However, as with E3, the MLBA provides a good fit to all of the response proportions, whereas the MDFT, LCA, and

AAM predict some data well (the attraction conditions), but have a large amount of clustering on the horizontal axis, again suggesting an inability to correctly predict the change in response proportions over conditions. As with E3, this experiment appears to reflect near identical results to those of E2, with a slightly worse overall fit.

Fig. 5 Quantile–quantile plots for the empirical response times (x-axis of each panel) and the model predicted response times (y-axis of each panel) of experiments E3 and E4. The rows of panels give the different experiments, being attraction only (ATT; E3a), compromise only (COM; E3b), similarity only (SIM; E3c), and all effects together (ALL; E4). The columns of panels give the different models, being

the MLBA, MDFT, the LCA, and the AAM. *Larger points* indicate a greater probability of a response in the empirical data, meaning that the quantile estimates are based upon a greater amount of data. Different *symbols* represent different experimental conditions. The colors of the symbols denote different types of alternatives: green represents the target, blue represents the competitor, and red represents the decoy

Farmer et al. 2016 (E5)

Again, we used the average attribute values for each condition as the inputs to the models. For the risky decision-making experiments, we had to normalize the potential

rewards, in order to put them on the same scale as the probability of obtaining the reward. Lastly, in order to give the attribute values an additive relationship and give them the same order of magnitude as the previous studies, we took the logarithm (this was done for both the risky and

perceptual tasks). For MDFT, the LCA, and the AAM, the time step used to simulate the models was set at 70 ms for E5a and E5b, and 20 ms for E5c. It should also be noted that we chose to exclude trials with extremely long response times from our analysis, with our criterion being 11 s for E5a and E5b, and 6 s for E5c. This difference in time steps and exclusion criteria between experiments was due to their different time scales, with E5a and E5b being slower, risky decision-making experiments, and E5c being a faster, perceptual decision-making experiment. Additionally, as noted in the method, we used maximum likelihood estimation to fit the models to these experiments rather than the Bayesian hierarchical framework used within the other experiments, due to computational tractability.

A summary of the results can be seen in the top three rows of Figs. 6 (response proportions) and 7 (response-time quantiles), which take the same format as Figs. 2 and 3, respectively. However, it should be noted that for this category of experiments, the group-averaged data points are shown instead of the individual-level data points. Our reason for this is that each condition only contained approximately eight trials, resulting in trends that were more difficult to interpret. However, assessing the individual data did not change the conclusion about the differences in the models' abilities to account for the data. In general, the predictions of all models appear to be somewhat worse than the attraction experiments in the perceptual and criminal inference studies, which is likely due to the data being much sparser in this experiment. However, the models still appear to account for the data well, considering that much larger numbers of trials are typically recommended for modeling response-time data (Lerche, Voss, & Nagler, 2017).

As in the previous attraction experiments, MLBA, MDFT, and LCA all appear to be about equally good at explaining both the response-time quantiles and the response proportions, with these models capturing the trends of the data. This holds for each of the experimental manipulations used, with all of these models fitting each of the data sets roughly as well as one another. Interestingly, these results seem to indicate that the models may not be easily distinguished, even at a quantitative level, by attraction experiments, as all three models are able to explain the data trends extremely well. However, one interesting finding is the difficulty AAM has in accounting for the data in these experiments.

Parrish et al. 2015 (E6)

As before, we used the average attributes for each condition as the model inputs, with the logarithm taken to make the relationship between the attributes is additive. For MDFT, the LCA, and the AAM, the time step used to simulate the models was set at 40 ms. It should also be noted that we

chose to exclude trials with extremely long response times from our analysis, with our criterion being 5 s.

A summary of the results can be seen in the bottom row of Figs. 6 (response proportion) and 7 (response-time quantiles). As with the Farmer et al. (2016) experiments, the predictions of all models appear to be slightly worse than those in the attraction experiment of E1, which again is likely to be due to the sparser amount of data per condition. However, the fits of the models still appear to be quite good to both the response-time and response-proportion data in an overall sense. In terms of the relative ability of the models to fit the data, as with the other attraction experiments, all models appear to do an equally good job of accounting for the trends in both the response-time quantiles and the response proportions. This finding seems somewhat unsurprising, seeing as this trend was present in all other attraction experiments in the other studies.

Generalization criterion analysis

Although the results have shown clear and convincing evidence that the MLBA provides a superior account of the compromise and similarity effects (for these data) compared to MDFT, the LCA, and the AAM, one could potentially argue that this superior fit is a result of its additional complexity. As noted in the model descriptions, the MLBA and the AAM contain nine parameters in comparison to the MDFT and LCA's eight parameters, which by classic standards suggests that the MLBA may be a more complex model (though this does not necessitate that it is more functionally complex, see Myung, 2000; Myung & Pitt, 1997; Evans, Howard, Heathcote, & Brown, 2017b; Evans and Brown, 2018). In order to address this potential issue, we assessed the predictive ability of the models to unseen data, using a version of the generalization criterion (Busemeyer & Wang, 2000). This process involved taking the estimated group-level mean values for each parameter for each model from the fits of each of the "combined" experiments (E2 and E4), and using these parameters to generate predicted data for the respective "separate" experiments (E1 and E3, respectively), which was then compared to the empirical data.

The probability–probability (P-P) and quantile–quantile (Q-Q) plots for both generalizations are shown in Fig. 8. All models do a fairly poor job of generalizing. However, it should be noted that a generalization task such as this is extremely difficult for models to perform in general, let alone when the models are ones whose drift rates are constrained to be a transformation of the attribute values of the stimuli. MLBA does appear to perform a little better, or at least no worse than MDFT, LCA, and AAM. These results indicate the underlying hypotheses, rather than increased complexity, are the source of MLBA's improved performance.

Fig. 6 Probability–probability plots for the empirical response proportions (x-axis of each panel) and the model predicted response proportions (y-axis of each panel) of experiments E5 and E6. The rows of panels give the different experiments, being Farmer 2A (E5a), Farmer 2B (E5b), Farmer 2C (E5c), and Parrish et al. (E6). The

columns of panels give the different models, being the MLBA, MDFT, the LCA, and the AAM. Different *symbols* represent different experimental conditions. The colors of the symbols denote different types of alternatives: green represents the target, blue represents the competitor, and red represents the decoy

Investigating the necessary components of MLBA in accounting for three context effects

In order to gain a better understanding of *why* the MLBA was able to qualitatively explain each of the three different context effects, we fit different versions of the MLBA to the data from E1a, E1b, E1c, and E2, with different parts of

the “front end” process removed. These experiments were chosen as E1, E2, E3, and E4 showed consistent findings, and E1 and E2 contained the greatest number of trials. Specifically, we defined three reduced MLBA models: one where different attributes received equal weight (i.e., no β parameter), one without the subjective value function (i.e., no m parameter), and one without asymmetry in the

Fig. 7 Quantile–quantile plots for the empirical response times (x -axis of each panel) and the model predicted response times (y -axis of each panel) of experiments E5 and E6. The rows of panels give the different experiments, being Farmer 2A (E5a), Farmer 2B (E5b), Farmer 2C (E5c), and Parrish et al. (E6). The columns of panels give the different models, being the MLBA, MDFT, the LCA, and the AAM.

Larger points indicate a greater probability of a response in the empirical data, meaning that the quantile estimates are based upon a greater amount of data. Different *symbols* represent different experimental conditions. The colors of the symbols denote different types of alternatives: green represents the target, blue represents the competitor, and red represents the decoy

attention weights (i.e., λ_+ and λ_- constrained to have the same value).

The results of these analyses can be seen in Table 1, with the full model fits also included for a point of reference. Interestingly, based on these analyses, the different components appear to have very different roles in the ability of the

MLBA to explain the context effects. When the β parameter is removed the MLBA becomes substantially worse in its ability to explain the attraction effect, as well as slightly worse in its ability to explain the similarity effect, suggesting that the ability to place different weights on attributes is key to explaining these attraction data. When

Fig. 8 Probability–probability (rows 1 and 3) and quantile–quantile (rows 2 and 4) plots for the generalization analysis. Each individual panel takes the same format as those of the previous figures. The *top two rows* display the generalization for the perceptual decision-making experiments (E2 generalized to E1), and the *bottom two rows*

display the generalization for the criminal inference experiments (E4 generalized to E3). Different *symbols* represent different experimental conditions. The colors of the symbols denote different models: green represents MLBA, blue represents MDFT, red represents LCA, and orange represents AAM

Table 1 Fits of several variants of MLBA, MDFT, LCA, and AAM to the data from experiments E1a, E1b, E1c, and E2

Model	PP				QQ				DIC			
	E1a	E1b	E1c	E2	E1a	E1b	E1c	E2	E1a	E1b	E1c	E2
MLBA	0.96	0.92	0.8	0.72	0.68	0.42	0.9	0.7	91283	73327	99886	126125
MDFT	0.95	0.66	0.23	0.52	0.72	0.65	0.86	0.72	94514	81029	110865	136481
LCA	0.92	0.36	0.22	0.49	0.74	0.63	0.89	0.71	89818	83572	110173	133698
AAM	0.87	0.06	0.09	0.37	0.72	0.64	0.89	0.72	92257	86119	109701	135348
MLBA _{β}	0.66	0.88	0.69	0.54	0.71	0.5	0.89	0.7	97891	92403	103046	133529
MLBA _{m}	0.96	0.54	0.29	0.53	0.7	0.5	0.9	0.7	94747	97741	105693	134084
MLBA _{λ}	0.95	0.92	0.69	0.71	0.7	0.55	0.9	0.69	94801	91929	103361	132118
MDFT _{sw}		0.66	0.14			0.65	0.86			81748	111742	
LCA _{sw}		0.38	0.13			0.64	0.89			83458	110867	
AAM _{sw}		0.12	0.12			0.63	0.89			86226	109612	
MDFT _{$sval$}		0.69	0.31			0.65	0.89			81171	109761	
LCA _{$sval$}		0.44	0.3			0.64	0.89			82993	109280	

The first four columns show the R^2 for the probability–probability correlation for response proportions between the observed data and the model simulated data, the second four columns show the R^2 for the quantile–quantile correlation for response-time quantiles between the observed data and the model simulated data, and the final four columns show the DIC values. For the MLBA models, the subscripts indicate the parameter that was removed from the model. For the MDFT/LCA/AAM models, the subscript sw refers to the models that had the switching mechanism removed and replaced with a single weighting parameter, and the subscript $sval$ refers to the models with a subjective value function added

the m parameter is removed, the MLBA becomes substantially worse in its ability to explain both the compromise and similarity effects, suggesting that the ability to place a subjective value on the attribute values is key to explaining the compromise and similarity data, and may be a key to the advantage of the MLBA over MDFT and the LCA in these data (note the AAM already contains a subjective value function). Lastly, when the λ parameters are constrained to be the same, regardless of whether the comparison of attributes is positive or negative, the MLBA only becomes slightly worse in its ability to explain the similarity effect, suggesting that a more restricted version of the model is sufficient for these data.

Investigating the deficiencies of MDFT, LCA, and AAM in explaining the compromise and similarity data

In order to gain a better understanding of *why* the MLBA outperformed MDFT, LCA, and AAM on the compromise and similarity data, we fit altered versions of these three models to the data of E1b and E1c. Although several differences exist between all of the models, there appears to be one major difference between the MLBA and these three models: the attribute weighting mechanism. Within the MLBA, all attributes are attended to simultaneously, with different weights assigned to each attribute based on the β parameter. In MDFT, LCA, and AAM, attributes are attended to one at a time, with a switching mechanism that results in the models “flickering” between attributes. The

justification for this attention switching mechanism was to allow the models to explain the similarity effect; however, given that these models all have difficulty explaining the similarity data, the mechanism appears to add little value. In order to assess whether the MLBA’s treatment of attributes was key to its superiority over MDFT, LCA, and/or AAM in fitting the compromise and/or similarity data, we created versions of these latter three models where attributes were attended to simultaneously (i.e., no switching mechanism), and the amount of attention paid to each attribute was controlled by a single parameter (i.e., a relative weight for the information from each attribute).

The results of these analyses can be seen in Table 1, with the full model fits and the MLBA fits also included for a point of reference. Interestingly, changing the attribute weighting mechanism in these three models (sw) does not appear to influence their ability to account for the compromise and similarity data of experiments E1b and E1c. Again, all three models perform extremely poorly in explaining the similarity and compromise data, with MDFT being better than the others, but about equally as good as in the original fits. These results suggest that differences in the attribute weighting mechanism do not explain any differences in the ability of the models to account for the empirical data.

Another key difference between the MLBA, and both MDFT and LCA, is the subjective value function used in the MLBA. Such a function allows for the possibility that an individual’s psychological interpretation of the options might differ from the way the options were defined

experimentally. For example, in the perceptual experiments, a pair of options with attributes P and Q, denoted by (P1, Q1) and (P2, Q2), were defined as indifferent if they had equal area, that is, if $\log(P1) + \log(Q1) = \log(P2) + \log(Q2)$. However, this simple and rational definition of indifference might not correspond to human perceptions of indifference. By introducing a subjective value function, we allow for differences between experimenter defined indifference and an observer's perceived indifference.

Interestingly, in our analyses with reduced (knocked-out) versions of the MLBA, we found that the MLBA was unable to explain the compromise and similarity effects (i.e., the fit was about equally as good as MDFT) when the subjective value function was removed. In order to assess whether the lack of a subjective value function was the reason for MDFT and LCA being unable to explain the compromise and similarity effects, we added a subjective value function to these models and fit them to the data from E1b and E1c. Our choice of subjective value function was the same as that used in the AAM (see Appendix E for details), which was also used in an application of MDFT to similarity data in Cataldo and Cohen (2018). The results of these analyses can be seen in Table 1. Interestingly, the addition of this subjective value function appears to do little to improve the ability of MDFT and LCA (*sval*) to account for the compromise and similarity data of experiments E1b and E1c.

Taken together, these results suggest that removing certain components of MLBA leads to poorer performance; however, borrowing aspects of MLBA (attention weighting and subjective valuation) and integrating them into the other models does not improve the ability of those models to account for observations. Thus it appears that MDFT, AAM, and LCA cannot jointly account for choice and RT data for the similarity and compromise experiments we had access to.

Fitting to only the response proportions

As discussed above, our results contrasting the models of multi-attribute choice provided two key findings: (1) all models provided a good fit the data from the attraction effect, and (2) although all models provided a good fit to the response-time quantiles for the compromise and similarity data, only the MLBA was able to provide a good fit to both the response times and proportions. In addition, our attempts to alter the functional form of MDFT, LCA, and AAM to be more like the MLBA did little to improve their ability to account for the full response-time distributions. This suggests one of two possibilities: (1) MDFT, LCA, and AAM are unable to account for the response proportions in these data, and the constraint of the response-time distributions had little impact, or (2) the models are able

to account for the response proportions in these data when they are purely assessed against them, but unable to jointly account for choice proportions and response times.

To test between these two possibilities, we fit all four models to just response proportions alone (i.e., ignoring the response-time distributions) for the compromise and similarity data for E1 (i.e., E1b and E1c). We did not consider the attraction effect here since all models provide a roughly equivalent accounting of both choice and RT data. Experiment E1 was chosen since E1, E2, E3, and E4 were all consistent, and E1 contained the greatest number of trials; however, the general pattern of results also held for E3 (see the [Supplementary Materials](#)). We fit to the response proportions of individual subjects by minimizing the root mean squared error (RMSE), which was optimized through differential evolution, using 500 iterations of 10k particles where *k* is the number of free parameters. We also fit the classic “random-walk” version of the LCA, which contained no decision threshold and terminated after 500 steps, but found no qualitative improvement in fit.

The results of these fits can be seen in the top two columns of Fig. 9. All four models provide at least a good account of the pure response proportions for the compromise effect, with the MLBA and MDFT providing the (near equally) best fit, and both the LCA and the AAM having some greater misfit. Thus, when choice proportion data alone is used, all models perform similarly. However, when choice proportions and RT data are fit together, MDFT, the LCA, and the AAM cannot jointly account for both. As a result, MDFT, LCA, and AAM fit the RT data at the expense of the choice proportions for the compromise effect.

Fits to the similarity effect appear to be relatively similar to those of the full response-time distributions. The MLBA provides the best fit even when choice proportions alone are used, and MDFT, LCA, and AAM are substantially worse with a great deal of misfit. However, it should be noted that MDFT, the LCA, and the AAM do fit the similarity response proportions better in this case than when they were constrained by the full response-time distributions. Thus, as with the compromise case, challenging MDFT, the LCA, and AAM with RT data weakens their ability to accurately account for choice proportions. Importantly, this is not the case for MLBA; fits of MLBA to choice proportions (for similarity and compromise data) are qualitatively similar (and good) when fit to purely proportion data or proportion/RT data. As with the fits to the response-time distributions, we also fit MDFT and LCA to the response proportions with an added subjective value function. These results can be seen in the bottom two columns of Fig. 9. Interestingly, these fits are greatly improved, with both models becoming about as good as the MLBA in explaining both the compromise and similarity data of E1b and E1c.

Fig. 9 Probability–probability plots for the empirical response proportions (x-axis of each panel) and the model predicted response proportions (y-axis of each panel) of experiments E1b and E1c, when the models were purely fit to the response proportions via RMSE. The rows of panels give the different experiments, being compromise only (COM; E1b) and similarity only (SIM; E1c). The *top two rows* show the fits to choices only for the versions of the models used in the fitting of the full choice response-time distributions. The *bottom*

two rows show the fits of MDFT and LCA with an added parameter to create a subjective function for the attribute values. The columns give the different models, being the MLBA, MDFT, the LCA, and the AAM. Different *symbols* represent different experimental conditions. The colors of the symbols denote different types of alternatives: green represents the target, blue represents the competitor, and red represents the decoy

These results provide some insights as to why the MLBA provides a superior accounting of the data when compared to MDFT, the LCA, and the AAM. First and foremost, these three models simply cannot account for joint choice

and RT data (while MLBA can). When you challenge these models with weaker data (choice alone), they appear to perform better. Furthermore, when you include subjective value in MDFT and LCA *and* challenge them with only

choice data, then and only then do you see rough parity between the models ability to account for data. We make two broad conclusions based on this analysis. First, the theoretical elements of MLBA more accurately capture the trends in data from these 12 data sets. Second, it is critically important to utilize response-time data when assessing these models. That is, one should test dynamic models with the kind of data they are designed to predict.

Discussion

Our study aimed to provide a comprehensive comparison between the prominent models of multi-attribute choice (MDFT, LCA, MLBA, and AAM) based on empirical choice response-time distributions of the key “context” effects that motivated their construction: the attraction effect, the compromise effect, and the similarity effect. In order to achieve this, our study used state-of-the-art hierarchical Bayesian methods to fit all models to full RT distributions. In order to ensure the robustness of our assessment of the performance of the models, we analyzed 12 different experiments appearing in six different studies. Based on this systematic approach, we come to two conclusions. First, the theoretical elements of MLBA provide a much better accounting of the similarity and compromise data used here than MDFT, LCA, and AAM (all are roughly similar for the attraction effect by comparison). Second, utilizing response times to test these models is critically important and in the case of this data, not doing so leads to spurious conclusions.

The importance of response-time distributions

Overall, our study showed consistent results across the experimental data sets. All models are able to fit choice proportions and RT’s for the attraction effect, except for AAM in E5. However, for similarity and compromise effects, only the MLBA was able to capture the choice response-time distributions. Specifically, MDFT, LCA, and AAM fit the response-time quantiles well, but produced poor predictions of choice proportions, resulting in an overall poorer fit to the choice response-time distributions.

Although these findings may intuitively seem to indicate that the choice proportions were the critical factor in distinguishing between the models, further analyses suggested that this was not the case. To test this possibility, we fit all models to only the choice proportions for E1b and E1c: the perceptual compromise and similarity effects, respectively (also see the [Supplementary Materials](#) for fits to the choice proportions of E3b and E3c). These results indicated that all models provided a good account of the response proportions for the compromise data set.

Furthermore, when a subjective value function was added to MDFT and the LCA, both models provided a good account of the choice proportions for the compromise and similarity data sets, but continued to provide a poor fit to the choice response-time distributions. These findings indicate that although all models are able to fit the response-time quantiles, constraining the models to account for the entire choice response-time distribution provides a greater ability to distinguish between them than fitting to choice proportions alone. This should not be surprising, as these models, by design, are dynamic models of decision-making. It thus stands to reason that dynamic (e.g., response-time) data is needed to evaluate their performance.

Which model components are important for explaining the context effects?

In addition to assessing each of the standard models, we also assessed altered versions of the models (using a knock-out/add-in approach) to try and better understand *why* specific models were able to explain specific context effects, and others were not. Firstly, we attempted to see what components of the MLBA were important in explaining the context effects by removing three key parts of the model’s “front end”: the ability to assign different weights to different attributes (β), the presence of a subjective value function (m), and asymmetry in the attention weights (different values for λ_+ and λ_-). Interestingly, each of these parameters appeared to contribute to some part of the MLBA’s ability to explain the three context effects. Specifically, the β and m parameters appeared to have the greatest impact, with the removal of the β parameter greatly reducing the model’s ability to fit the attraction data and slightly reducing the model’s ability to fit the similarity data, and the removal of the m parameter greatly reducing the model’s ability to fit both the compromise and similarity data. Restricting the λ parameters to take on the same value had less of an impact, though it did slightly reduce the model’s ability to fit the similarity data. Taken alone, these findings appear to suggest that the inclusion of a subjective value function may have resulted in the MLBA’s superiority over MDFT and LCA.

However, when we modified the MDFT and LCA to incorporate a subjective value function, these changes did little to improve the fit of these models to the choice response-time distributions. Specifically, when altering the form of these two models (as well as AAM) to remove the attention switching mechanism and replace it with a constant attention weight (i.e., similar to the MLBA’s β), we found little improvement in the ability of the models to account for the compromise and similarity effects. In addition, when incorporating a subjective value function into MDFT and LCA—the component of the MLBA that

was key to explaining both the compromise and similarity data—we also found little improvement in fit. Interestingly, these findings demonstrate that it is not specific components of the MLBA that allow it to explain certain effects, but rather how these specific components are integrated into the MLBA's specific functional form.

Why do our results differ from those of Turner et al. 2018?

Interestingly, the findings of our study fall in contrast to those of some previous studies, and particularly the recent comparison between these four models in Turner et al. (2018). Specifically, Turner et al. (2018) mostly found that AAM and the LCA provided a superior account to the MLBA and MDFT when only considering choice proportions. While we cannot state with certainty why results herein differ from those previously, we briefly discuss some of the key differences between these two studies.

First and foremost, the most obvious difference between our study and Turner et al. (2018) is the type of data being fit: choice response-time distributions vs. only choice proportions. As discussed above, the additional constraint of response-time distributions provided a greater ability to distinguish between the models in our data, suggesting that some of the difference in results may have been due to the increased ability of response-time data to distinguish the models.

That said, when we fit these models to choice proportions alone for the data sets considered here, we still find results inconsistent with those in Turner et al. (2018). There are two general possible explanations for this. First, they utilized a different experimental design that involved both binary and ternary choice whereas all of the experiments we considered involved ternary choices. It is possible that different models better match data from different experimental designs. It would, of course, raise serious concerns regarding the value of these models for understanding the associated cognitive phenomena if different experimental designs lead to different conclusions. However, even if that is the case here (which is not clear), we again raise the point that this is an issue when modeling choice proportions alone and the choice/RT data used here likely provides a stronger test of these theories. Testing this possibility can only be performed by carrying out a parallel version of the study in Turner et al. (2018) to capture RT data and fit that data using the methods demonstrated here.

The second possibility is that differences in the specific mathematical formulations of these models between the two studies are responsible for differences in outcomes. In this regard, the biggest issue we see with the study in Turner et al. (2018) is that there is a mismatch between the

experimental design and the models they fit. Specifically, Turner et al. (2018) utilized models that encoded the assumption that participants made a decision at a fixed time (i.e., an external stopping rule), while the experimental procedure allowed participants to make a decision at any time (i.e., an internal stopping rule). Again, however, the only way to test whether this is a source of disagreement would be to fit the data in Turner et al. (2018) using versions of all models incorporating internal stopping rules.

Overall, we believe that both our study and Turner et al. (2018) provide unique and important pieces of evidence in comparing the models of multi-attribute choice, each with their own strengths and weaknesses. At the moment, as is often the case in science, those studies come to different conclusions. We believe that future research should aim to understand and resolve the differences between these studies, which will hopefully provide an even more complete picture in the comparison of models of multi-attribute choice. Further, based on our results here, we think the best way forward to do so would be to utilize the methods discussed here to test dynamic models with dynamic data, namely the very response-time data they were designed to predict.

Future directions

In addition to the theoretical results, our study provides a new methodological standard for studies within the field of multi-attribute decision-making. As mentioned above, previous studies have mostly used basic simulations that attempt to see whether the models can produce trends that are qualitatively consistent with the concepts of the effects. Our study greatly advances on these previous methods, fitting the full choice response-time distributions at the level of individual subjects, using PDA to obtain a likelihood for the MDFT, LCA, and AAM, and Bayesian hierarchical modeling to gain the benefit of group inference.

It should also be noted that there are several practical limitations to using the MDFT, LCA, and AAM models. Although we were able to fit these models through advanced methods, this implementation involves a large practical cost. Firstly, the implementation of these advanced methods can be quite difficult, as customized code must be used to implement these state-of-the-art methods. Although our code is openly available on OSF (<https://osf.io/h7e6v/>), the implementation process is still more complex than that of the simpler MLBA. Secondly, although we managed to fit these models, the computational cost was significant. All fits took between 18 and 96 h of computer time to finish, using six or more cores simultaneously per fit and highly optimized C code to perform the model simulations. In contrast, the fits for the MLBA only used a single core per

fit and took half of the computer time (roughly 12 times quicker on the whole), using code written purely in R.

One important future direction is examining these models on choice and response-time data from preferential choice tasks (e.g., choices among consumer products). The current paper only includes data from perceptual, inference, and risky decision-making domains. At the time of writing, we were unable to locate any preferential choice experiments that included response-time data. This is in part due to the standard in this domain of only collecting choice data, assuming response times are less important. We hope that the current work sheds light on the importance of including response-time data in the evaluation of different theories and we hope future empirical studies will gather this type of data. We note that it is possible that our results might not hold in preferential choice tasks because there are fundamental differences between preferential choice and the domains we examined. In particular, preferential choice tasks involve individuals trying to satisfy personal goals rather than an externally imposed criterion as in perceptual and inference domains (such as selecting the largest rectangle).

Lastly, although our paper compared four prominent models of context effects in multi-attribute choice, several other models exist that could be used in future comparisons: the range-normalization model (Soltani, De Martino, & Camerer, 2012), the 2N-ary choice tree model (Wollschlaeger & Diederich, 2012), and the model of Howes et al. (2016). The range-normalization model is a neurally inspired model, which proposes that neural representations contain trial-to-trial variability, which allow it to account for the attraction and similarity effects, though it is yet to be assessed in its ability to account for the compromise effect. The 2N-ary choice tree model contains some theoretical similarities with the MLBA, proposing that the process consists of pairwise comparisons between alternatives on weighted attribute values, and having no requirement for inhibition or loss-aversion to explain the context effects. However, unlike the models assessed within our study, this model requires an inequality in attribute weights to explain the attraction effect. Howes et al. (2016) developed a model of expected value maximization, which showed that when attribute value estimation is noisy, context effects such as those explored within this paper can be expected to be observed under a rational framework. All three of these models provide interesting avenues for future research in comparisons of models of context effects.

Conclusions

Multi-alternative, multi-attribute choice, and the models that attempt to explain the processes that underlie these

decisions, have been studied for several decades across the fields of psychology, neuroscience, and economics. These models have greatly increased our understanding of what influences human decision-making. In particular, MDFT, the LCA, the MLBA, and the AAM have provided insights to violations of independence as seen in three context effects: the attraction, similarity, and compromise effects. Our study is one of the first to perform a comprehensive comparison of these four different theoretical explanations of multi-alternative, multi-attribute choice, comparing them on their ability to account for the choice response-time distributions of 12 experiments across six published studies in perceptual, inference, and risky decision-making domains. Results show that only MLBA provides a good fit to both choice proportion and RT data, significantly out-performing the other models.

Beyond this specific conclusion, our results point to two more general conclusions. First, none of these models are suitable for parameter-based inference. That is, all these models are “sloppy” (weakly identifiable) in the sense that a range of parameter values for each can produce very similar data distributions. Thus, it is not possible to draw conclusions from the specific values of parameters. This could of course potentially be ameliorated in the future by considering appropriate model simplifications or alternative experimental designs. Second, our results suggest it is critical to utilize choice/response-time data rather than choice proportions alone when working with these models. When only choice proportions are considered (RT’s are excluded for fitting), the models provide reasonable fits to data and capture the three basic context effects. Thus, challenging these models with the very response-time data they were designed to predict, rather than more limited choice proportion data, leads to fundamentally different conclusions. This last point underscores the need to use dynamic response-time data to assess the validity of dynamic theories of choice as neglecting it leads to under-constrained models and reduces the ability to discriminate between those theories.

Acknowledgements The authors would like to thank Audrey Parrish, Michael Beran, and George Farmer for sharing their data. The authors would also like to thank Jerome Busemeyer for his comments on extending MDFT to SDE formalism. All authors were supported by NSF grant SES-1556415. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the funding agency.

Appendix A: Converting random walks to stochastic differential equations

Here we provide a very brief general description of how we converted the discrete time, random walk models versions

of MDFT and LCA to stochastic differential equations (SDE's). We are not attempting to re-derive the link between random walks and SDE's. Rather our intent is to provide the interested reader with a short primer. We thus use formal rather than rigorous arguments and language.

Discrete time random walks, which are commonly used to describe the accumulation of evidence or time integration of stimuli, are typically described by specifying the form of a stochastic increment (size of a step) that advances a system from one “time step” to the next

$$A_{n+1} - A_n = rf(A_n) + Noise, \quad (7)$$

where “ r ” is the size of the step taken and “Noise” represents a generic stochastic term. We highlight the phrase “time step” here since this really refers to an indexed counter rather than a true time increment. While there are a number of ways to associate real time with random walks, the most common way of doing so when dealing with random walks is to convert them to a stochastic differential equation framework.

To do so, one has to associate each time increment “ n ” with a real time t_n . This introduces a time increment into the problem, namely $\Delta t = t_{n+1} - t_n$. In order to ensure that the statistical properties of the random walk at a fixed future point in time do not depend on the size of the time increment chosen, it is commonly assumed that $r = k\Delta t$, so that halving the size of the time step halves the size of the step. In other words, taking two steps of size $\Delta t/2$ yields similar results as taking one step of size Δt . With this assumption, the resulting stochastic differential equation obtained by taking $\Delta t \rightarrow 0$ is

$$dA = kf(A)dt + \sigma dw. \quad (8)$$

Since the random walk function f typically has scaling parameters of its own, k can often be folded into that function and omitted for brevity. The random walks formulations of MDFT and LCA can each be reformulated in this way.

Appendix B: MDFT

Multi-alternative decision field theory (MDFT) models the preferences for different alternatives of a multi-attribute choice over decision time through three key components: the valence for each alternative, the lateral inhibition between alternatives, and the evidence leakage within each alternative, and some stochastic noise within the process. Formally, the preference for the vector of alternatives P (e.g., for the choice alternatives X, Y, and Z, $P = [P_X, P_Y,$

$P_Z]$ ’) can be written as a stochastic differential equation, with the form:

$$dP = \mu \times dT + \sigma \times dW \quad (9)$$

$$\mu = [S \times P(t) + V(t)] \quad (10)$$

where W is the Wiener process. In terms of the model components, S is an $n \times n$ matrix (where n is the number of alternatives) that contains both the amount of lateral inhibition between the alternatives based upon their psychological distances, and the rate of evidence leakage for each alternative. $P(t)$ is the preference for the vector of alternatives at the time t , and $V(t)$ is the vector of valences for each alternative at time t based upon the attribute being attended to. σ is the standard deviation of the noise process, which is a free parameter of the model. This can be written as a discrete process of the form:

$$P(t + \Delta t) = P(t) + S \times P(t) \times \Delta t + V(t) \times \Delta t + \sigma \times \sqrt{\Delta t} \times \eta \quad (11)$$

where Δt is the time step, η is the standard normal distribution, and $P(t + \Delta t)$ is the preference for the vector of alternatives at the next point in time. The preferences for each alternative continue to evolve, until one of the preferences passes some threshold level (a).

In terms of the three key components, the first, S , is the matrix of the lateral inhibition between the alternatives, and the evidence leakage within alternatives. S is an $n \times n$ matrix (where n is the number of alternatives):

$$S_{ij} = -\phi_2 \times \exp(-\phi_1 \times Dist_{ij}^4) \quad (12)$$

where i and j index alternatives (1, 2, ... n) and ϕ_1 and ϕ_2 are parameters. More specifically, ϕ_2 is a free parameter that gives the amount of evidence leakage, and ϕ_1 is a free parameter that controls the overall strength of the lateral inhibition. $Dist_{ij}$ is the psychological distance between the attributes of alternatives i and j (which is 0 when $i = j$). To obtain the psychological distance between the attributes of the two alternatives, one obtains the Euclidean distance between the attributes based on the dimensions of indifference (i.e., where alternatives are equally as good) and dominance (i.e., where one alternative dominates another):

$$Dist_{ij} = \sqrt{(\Delta I_{ij})^2 + \beta \times (\Delta D_{ij})^2} \quad (13)$$

where ΔI_{ij} gives the distance between the attributes of the alternatives on the dimension of indifference, ΔD_{ij} gives the distance between the attributes of the alternatives on the dimension of dominance. β is a free parameter that allows the dominance and indifference dimensions to be weighted differently, where $\beta > 1$ has a stronger weight to the dominance dimension, $\beta < 1$ has a stronger weight to the indifference dimension, and $\beta = 1$ gives both dimensions

equal weighting. Also note that this formulation of distance is only possible when only two attributes are present in each stimulus. Although Berkowitsch et al. (2014, 2015) developed a formulation that works for more than two alternatives, that version reduces to this one in the specific case of two alternatives. Therefore, we choose to write the mathematically simpler version here.

To define the distance on each of the dimensions, one takes the two subjective attributes values for alternatives i and j , with the first attributed denoted as A and the second denoted as B , and calculates their distance on the unit plane:

$$\Delta I_{ij} = \frac{1}{\sqrt{2}} \times ((B_i - B_j) - (A_i - A_j)) \quad (14)$$

$$\Delta D_{ij} = \frac{1}{\sqrt{2}} \times ((B_i - B_j) + (A_i - A_j)) \quad (15)$$

Note that in our application, we constrain the subjective values of the stimulus to equal the objective values, meaning that these values for attributes A and B for alternatives i and j are merely the actual stimulus attributes for these alternatives.

The second key component, $V(t)$, is the vector of valences for each alternative at time t . $V(t)$ is defined as:

$$V(t) = C \times [M \times W(t)] \times \gamma \quad (16)$$

where γ is an addition that we made to the model, which is a free parameter that scales the valences. This parameter allows the model to create drift rates of different magnitude, which is critical when modeling the choice response-time distributions. C is an $n \times n$ matrix that transforms the absolute valence values into relative advantages/disadvantages, given by:

$$C = \begin{bmatrix} 1 & \frac{-1}{n-1} & \frac{-1}{n-1} \\ \frac{-1}{n-1} & 1 & \frac{-1}{n-1} \\ \frac{-1}{n-1} & \frac{-1}{n-1} & 1 \end{bmatrix} \quad (17)$$

M is an $n \times m$ matrix, where m is the number of attributes, which contains the subjective values for each attribute of each alternative, and in the case of our two-attribute design would be:

$$M = \begin{bmatrix} A_1 & B_1 \\ A_2 & B_2 \\ A_3 & B_3 \end{bmatrix} \quad (18)$$

$W(t)$ is a binary vector of length m , where all elements are set to 0, except for the attribute of the matrix that is being attended to at time t (either A or B), which is set to 1. For example, if attribute A were being attended to at time t , then $W(t) = W_A = [1, 0]'$, and if attribute B were being attended to, then $W(t) = W_B = [0, 1]'$. To determine which attribute would be attended to at time $t + \Delta t$, we implemented an exponential switching function, where the

probability of switching between attended attributes at any point in time was given by:

$$P(W_A \mapsto W_B, \Delta t) = 1 - \exp(-k_A \times \Delta t) \quad (19)$$

$$P(W_B \mapsto W_A, \Delta t) = 1 - \exp(-k_B \times \Delta t) \quad (20)$$

where k_A and k_B are free parameters, separate for attributes A and B to allow bias towards a single attribute, which control the switching probability, with higher values giving larger switching probabilities. The exponential switching function is another addition that we made to the model. Previous iterations of MDFT operated as a random walk where attention would switch on every step, serving as a “flicker” between attributes. However, this flicker process does not fit the stochastic differential equation version of the model for two key reasons. Firstly, from a psychological perspective, it seems unrealistic that participants would be changing their attentional focus on every time step, when the time step is in the order of several milliseconds. Secondly, the flicker process used in the previous iterations of MDFT would result in a stochastic differential equation where the number of switches occurring within a specified time period is dependent on the time step used, meaning that different results would potentially be obtained from the model on the basis of the time step used. Our exponential switching function serves as a solution to both of these issues, possessing a constant hazard function that results in the probability of switching after a specified time period to be independent of the time step used, and more generally allowing the switching time frame to take on a value more realistic for human attention processes.

Overall, our definition of MDFT gives the model eight free parameters: a , k_A , k_B , ϕ_1 , ϕ_2 , β , σ , and γ .

Appendix C: LCA

We first briefly describe the original formulation of LCA provided in Usher and McClelland (2004). Due to scaling issues, however, this is not the model used in the paper here. We thus subsequently describe our augmentation of the version used.

Random walk LCA

The original version of LCA posited that the activation level associated with alternative i is governed by the following random walk

$$A_{t+1}^i - A_t^i = (\lambda - 1)A_t^i + (1 - \lambda) \left(I_i(t) - \beta \sum_{j \neq i} A_t^j + \text{Noise} \right). \quad (21)$$

When converted to a SDE in the standard way, this model would take the form

$$dA^i = (1 - \lambda) \left[-A^i + I_i(t) - \beta \sum_{j \neq i} A_j^j \right] + \sigma dW \quad (22)$$

This introduces two issues. The first is that $(1 - \lambda)$ scales the entire drift component of this SDE. Thus it will trade off with response thresholds and yield a parameter degeneracy. We attempted data/parameter recovery and found that this parameter degeneracy caused serious issues and the model could not recover its own data. The second issue is the interpretation of the parameter λ . In Usher and McClelland (2004), this was described as the leakage rate parameter. However, if you look at either the original random walk (21) or its associated SDE (22), it is plain to see that λ does not act as leakage rate. Due to these two issues, we have constructed a slightly adjusted version of the original LCA model (in SDE form), that maintains all the inhibition and loss aversion features of the original in their exact mathematical form, but more appropriately describes leakage (in a manner that reflects the definition of the LCA in Usher & McClelland, 2001).

Adjusted SDE version of LCA

The leaky computing accumulator (LCA) models the activation for different alternates of multi-attribute choice over decision time through three key components: the activation input for each alternative, the global amount lateral inhibition between alternatives and leakage within each alternative, and some stochastic noise within the process. Formally, the amount of activation for the vector of alternatives A (e.g., for the choice alternatives X , Y , and Z , $A = [A_X, A_Y, A_Z]^T$) can be written as a stochastic differential equation, with the form:

$$dA = \mu \times dT + \sigma \times dW \quad (23)$$

$$\mu = [S \times A(t) + I(t)] \quad (24)$$

where W is the Wiener process. In terms of the model components, S is an $n \times n$ matrix (where n is the number of alternatives) that contains both the amount of lateral inhibition between the alternatives based upon their psychological distances, and the rate of evidence leakage for each alternative. $A(t)$ is the activation for the vector of alternatives at the time t , and $I(t)$ is the vector of inputs for each alternative at time t based upon the attribute being attended to. σ is the standard deviation of the noise process, which is a free parameter of the model. This can be written as a discrete process of the form:

$$A(t + \Delta t) = A(t) + S \times A(t) \times \Delta t + I(t) \times \Delta t + \sigma \times \sqrt{\Delta t} \times \eta \quad (25)$$

where Δt is the time step, η is the standard normal distribution, and $A(t + \Delta t)$ is the activation for the vector of alternatives at the next point in time. Additionally, the level of activation is truncated at 0, meaning that:

$$[A_i(t + \Delta t) \leq 0] = 0 \quad (26)$$

where i indexes an individual alternative. The activation for each alternative continues to evolve, until one of the alternatives' activation passes some threshold level (a).

In terms of the three key components, the first, the matrix of the lateral inhibition between the alternatives, and the evidence leakage within alternatives, S , is an $n \times n$ matrix (where n is the number of alternatives), defined as:

$$S = \begin{bmatrix} -\lambda & -\beta & -\beta \\ -\beta & -\lambda & -\beta \\ -\beta & -\beta & -\lambda \end{bmatrix} \quad (27)$$

where β is a free parameter of the model that controls the amount of global inhibition, and λ is a free parameter of the model that controls the amount of leakage.

The second key component, the vector of activation inputs for each alternative at time t , $I(t)$, is defined as:

$$I_i(t) = I_0 + \gamma \times \sum_{j \neq i}^n V(d_{ij}) \quad (28)$$

where I_0 is a free parameter of the model that is the baseline amount of input, V is a function where:

$$V(x) = \log(1 + x), x \geq 0 \quad (29)$$

$$V(x) = -[\log(1 + |x|) + \log(1 + |x|)^2], x < 0 \quad (30)$$

d_{ij} is given by:

$$d_{ij} = (M \times W(t))_i - (M \times W(t))_j \quad (31)$$

where M is an $n \times m$ matrix, m being the number of attributes, which contains the values for each attribute of each alternative. In the case of our two-attribute design, where the first attribute is denoted F and second attribute is denoted G , M would be:

$$M = \begin{bmatrix} F_1 & G_1 \\ F_2 & G_2 \\ F_3 & G_3 \end{bmatrix} \quad (32)$$

$W(t)$ is a binary vector of length m , where all elements are set to 0, except for the attribute of matrix that is being attended to at time t (either F or G), which is set to 1. For example, if attribute F were being attended to at time t , then $W(t) = W_F = [1, 0]^T$, and if attribute G were being attended to, then $W(t) = W_G = [0, 1]^T$. To determine which attribute would be attended to at time $t + \Delta t$, we implemented an exponential switching function, where the probability of

switching between attended attributes at any point in time was given by:

$$P(W_F \mapsto W_G, \Delta t) = 1 - \exp(-k_F \times \Delta t) \quad (33)$$

$$P(W_G \mapsto W_F, \Delta t) = 1 - \exp(-k_G \times \Delta t) \quad (34)$$

where k_F and k_G are free parameters, separate for attributes F and G to allow bias towards a single attribute, which control the switching probability, with higher values giving larger switching probabilities. The exponential switching function is another addition that we made to the model. Previous iterations of LCA operated as a random walk where attention would switch on every step, serving as a “flicker” between attributes. However, this flicker process does not fit the stochastic differential equation version of the model for two key reasons. Firstly, from a psychological perspective, it seems unrealistic that participants would be changing their attentional focus on every time step, when the time step is in the order of several milliseconds. Secondly, the flicker process used in the previous iterations of LCA would result in a stochastic differential equation where the number of switches occurring within a specified time period is dependent on the time step used, meaning that different results would potentially be obtained from the model on the basis of the time step used. Our exponential switching function serves as a solution to both of these issues, possessing a constant hazard function that results in the probability of switching after a specified time period to be independent of the time step used, and more generally allowing the switching time frame to take on a value more realistic for human attention processes.

Overall, our definition of the LCA gives the model eight free parameters: a , k_F , k_G , I_0 , λ , β , σ , and γ .

Appendix D: MLBA

The multi-attribute linear ballistic accumulator (MLBA) models the activation for different alternates of multi-attribute choice over decision time through three key components: the pairwise alternative comparison, the decaying comparison weights over increasing attribute differences (i.e., similarity-based attention), and the subjective attribute values. Specifically, the MLBA serves as a front-end extension of the linear ballistic accumulator (LBA; Brown & Heathcote, 2008), where the front-end serves to determine the drift rates based on some function of the attribute values. The linear, ballistic format of the LBA framework makes the MLBA more tractable than the MDFT and LCA, having an analytically solvable probability density function. The MLBA can be expressed as:

$$dA = \mu \times dT \quad (35)$$

$$\mu \sim TN(d, s, 0, Inf) \quad (36)$$

where TN is the normal distribution truncated to values between 0 and positive infinite, d is the mean drift rate between trials, and s is the standard deviation in drift rate between trials. With the MLBA, the standard deviation in drift rate is fixed to 1, to satisfy a scaling property of the model (Donkin, Brown, Heathcote, & Wagenmakers, 2011). Alternatives accumulate evidence independently from one another, with the starting evidence distributed as $U[0, A]$, until the evidence for one alternative reaches a threshold, b , with A and b being free parameters of the model.

The mean drift rate (commonly referred to as just the “drift rate”) for each alternative was determined by pairwise comparisons between the attributes of that alternative and the other alternatives:

$$d_i = I_0 + \gamma \times \sum_{j \neq i}^n V_{i,j} \quad (37)$$

where I_0 is the baseline drift rate (a free parameter of the model), n is the number of alternatives, and:

$$V_{i,j} = W_{P,i,j} \times (u_{P,i} - u_{P,j}) + W_{Q,i,j} \times (u_{Q,i} - u_{Q,j}), \quad (38)$$

where $W_{P,i,j}$ refers to the weight given to the difference between alternatives i and j for attribute P , given by:

$$W_{P,i,j} = \exp(-\lambda \times |u_{P,i} - u_{P,j}|) \quad (39)$$

$$\lambda = \lambda_1, u_{P,i} - u_{P,j} \geq 0 \quad (40)$$

$$\lambda = \lambda_2, u_{P,i} - u_{P,j} < 0 \quad (41)$$

and $u_{P,i}$ refers to the subjective attribute value for alternative i on attribute P , given by:

$$a = b = P_i + Q_i \quad (42)$$

$$\text{angle} = \tan^{-1} \left(\frac{Q_i}{P_i} \right) \quad (43)$$

$$u_{P,i} = \frac{b}{((\tan(\text{angle}))^m + \left(\frac{b}{a}\right)^m)^{\frac{1}{m}}} \quad (44)$$

$$u_{Q,i} = b \times \left(1 - \left(\frac{u_{P,i}}{a} \right)^m \right)^{\frac{1}{m}} \quad (45)$$

Overall, this gives the MLBA nine free parameters: A , b , t_0 , γ , β , λ_1 , λ_2 , I_0 , and m .

Appendix E: AAM

The associations and accumulation model (AAM) models the activation for different alternates of multi-attribute choice over decision time through three key components: the activation input for each alternative, the global amount

lateral inhibition between alternatives and leakage within each alternative, and some stochastic noise within the process. Formally, the amount of activation for the vector of alternatives A (e.g., for the choice alternatives X , Y , and Z , $A = [A_X, A_Y, A_Z]^T$) can be written as a stochastic differential equation, with the form:

$$dA = \mu \times dT + \sigma \times dW \quad (46)$$

$$\mu = [S \times A(t) + I(t)] \quad (47)$$

where W is the Wiener process. In terms of the model components, S is an $n \times n$ matrix (where n is the number of alternatives) that contains both the amount of lateral inhibition between the alternatives, and the rate of evidence leakage for each alternative. $A(t)$ is the activation for the vector of alternatives at the time t , and $I(t)$ is the vector of inputs for each alternative at time t based upon the attribute being attended to. σ is the standard deviation of the noise process, which is a free parameter of the model. This can be written as a discrete process of the form:

$$A(t + \Delta t) = A(t) + S \times A(t) \times \Delta t + I(t) \times \Delta t + \sigma \times \sqrt{\Delta t} \times \eta \quad (48)$$

where Δt is the time step, η is the standard normal distribution, and $A(t + \Delta t)$ is the activation for the vector of alternatives at the next point in time. The activation for each alternative continues to evolve, until one of the alternatives' activation passes some threshold level (a).

In terms of the three key components, the first, the matrix of the lateral inhibition between the alternatives, and the evidence leakage within alternatives, S , is an $n \times n$ matrix (where n is the number of alternatives), defined as:

$$S = \begin{bmatrix} -\lambda & -\beta & -\beta \\ -\beta & -\lambda & -\beta \\ -\beta & -\beta & -\lambda \end{bmatrix} \quad (49)$$

where β is a free parameter of the model that controls the amount of global inhibition, and λ is a free parameter of the model that controls the amount of leakage.

The second key component, the vector of activation inputs for each alternative at time t , $I(t)$, is defined as:

$$I_i(t) = \gamma \times V(d_i) \quad (50)$$

where V is a function where:

$$V(x) = \text{sign}(x)|x|^\alpha \quad (51)$$

d_i is given by:

$$d_i = (M \times W(t))_i \quad (52)$$

where M is an $n \times m$ matrix, m being the number of attributes, which contains the values for each attribute of

each alternative. In the case of our two-attribute design, where the first attribute is denoted F and second attribute is denoted G , M would be:

$$M = \begin{bmatrix} F_1 & G_1 \\ F_2 & G_2 \\ F_3 & G_3 \end{bmatrix} \quad (53)$$

$W(t)$ is a binary vector of length m , where all elements are set to 0, except for the attribute of matrix that is being attended to at time t (either F or G), which is set to 1. For example, if attribute F were being attended to at time t , then $W(t) = W_F = [1, 0]^T$, and if attribute G were being attended to, then $W(t) = W_G = [0, 1]^T$. To determine which attribute would be attended to at time $t + \Delta t$, we implemented an exponential switching function, where the probability of switching between attended attributes at any point in time was given by:

$$P(W_F \mapsto W_G, \Delta t) = 1 - \exp\left(-\left(k_F + \sum_{j=1}^n G_j\right) \times k_{scale} \times \Delta t\right) \quad (54)$$

$$P(W_G \mapsto W_F, \Delta t) = 1 - \exp\left(-\left(k_G + \sum_{j=1}^n F_j\right) \times k_{scale} \times \Delta t\right) \quad (55)$$

where k_F and k_G are free parameters, separate for attributes F and G to allow bias towards a single attribute, which control the switching probability, with higher values giving larger switching probabilities. The sum of the attribute values over all alternatives for the attribute being switched to represent the attribute's activation, with attributes that have higher overall values being more highly activated, and therefore, more likely to be attended to. The final parameter, k_{scale} , is used to re-scale these activation values onto how frequently the switching should occur. The exponential switching function is another addition that we made to the model, as with MDFT and the LCA, to aide the extension to a stochastic differential equation.

Overall, our definition of the AAM gives the model nine free parameters: a , k_F , k_G , k_{scale} , α , λ , β , σ , and γ .

Appendix F: Recovery assessment

Here we discuss the efficacy of the aforementioned methods when applied to these models with the type of data available for this study. Our goal is to determine how well each of these models accounts for data in these six studies. Toward this end, we generate data out of each model with pre-determined parameters, apply the fitting methodologies, and determine how well each model recovers the RT distributions for each alternative. Importantly, based on the results of these recoveries we choose to *not* report the

estimated parameter values from fitting any of the models, as based on the recovery analysis, these values are of little meaning.

Importantly, to make robust inferences using a quantitative model, one should first establish its reliability. Without having established this reliability, any inferences that are made about the estimated parameter values, or its ability to fit the data, may be spurious. Two important reliability benchmarks to establish are whether a model is able to recover its own parameters, where fitting a model to data it generated results in estimated parameters that match those used to generate the data, and whether a model is able to recover its own data, where fitting a model to data it generated results in the quantitative predictions of the estimated parameters matched the generated data. In this section, we overview our testing of these reliability benchmarks for the three models, and the full analysis can be found in the [Supplementary Materials](#).

For the recoveries of each models, we generated eight sets of data: four types of effects (attraction only, compromise only, similarity only, all three effects together), each with two different numbers of trials (120, 1000). Data recovery was assessed for each type of effect for both numbers of trials, whereas parameter recovery was assessed for each type of effect on only the highest (1000) number of trials. Note that each data set is generated using a single set of parameters with a single set of stimuli. This is akin to a simple experiment with just one condition.

To briefly summarize, each of these models appears to have strengths and weaknesses in terms of recovery. In terms of data recovery, MDFT AAM, and MLBA showed perfect performance in the 1000 trials data sets for all effects. Although the performance was not quite perfect in the 120 trials data sets, this appeared to be due to the high amount of noise in the generating data, as the 120 trials data set differed substantially from the 1000 trials data set. Thus, the combination of method and model performs very well at recovering the data generated from these models. However, both models did have issues with parameter identifiability. MLBA exhibited significant correlations between parameters in the function that maps attribute information to drift rates. MDFT had similar issues, with large correlations between the attention switching parameters, and between the threshold, drift rate, and stochastic noise coefficient parameters, as well as containing other unclear problems in the recovery of the inhibition function. For AAM, parameter correlation again appeared to be the key issue in inability to recover parameters, though the precise problematic correlations were less clear.

The LCA had more significant issues that required a slight augmentation of the stochastic differential equation. In terms of data recovery, the standard LCA (i.e., with the

“(1 − λ)” term applied to every element of the equation) produced poor recovery for all effects at all numbers of trials, which would mean inferences about the LCA’s ability to fit to empirical data would not be reliable. To fix this issue, so that LCA could be part of the comparison, we altered the functional form to no longer include the “(1 − λ)” term, which is the version that we presented earlier in the [“Models and methods”](#) section. When using this altered version, the performance of the LCA was very similar to that of MDFT: data recovery was near perfect on the 1000 trials, and parameter recovery revealed identifiability issues that appeared to apply to every parameter within its functional form to at least some degree. Therefore, we will only discuss the altered version from here onwards (though fitting E1 with the “(1 − λ)” term applied made no qualitative differences to the fit).

Appendix G: Additional modeling details

This section provides the Bayesian hierarchical structure (E1, E2, E3, E4, and E6) and the maximum likelihood boundaries (E5) used for each model across the six analyzed studies.

Bayesian hierarchical structure

MDFT

Data level :

$$(RT_i, resp_i) \sim MDFT(a_i, k_{A,i}, k_{B,i}, \phi_{1,i}, \phi_{2,i}, \beta_i, \sigma_i, \gamma_i)$$

Group level :

$$a_i \sim TN(\mu_a, \sigma_a, 0, 100)$$

$$k_{A,i} \sim TN(\mu_{k_A}, \sigma_{k_A}, 0, 200)$$

$$k_{B,i} \sim TN(\mu_{k_B}, \sigma_{k_B}, 0, 200)$$

$$\phi_{1,i} \sim TN(\mu_{\phi_{1,i}}, \sigma_{\phi_{1,i}}, 0, 10)$$

$$\phi_{2,i} \sim TN(\mu_{\phi_{2,i}}, \sigma_{\phi_{2,i}}, 0, 30)$$

$$\beta_i \sim TN(\mu_{\beta}, \sigma_{\beta}, 0, 40)$$

$$\sigma_i \sim TN(\mu_{\sigma}, \sigma_{\sigma}, 0, 40)$$

$$\gamma_i \sim TN(\mu_{\gamma}, \sigma_{\gamma}, 0, 400)$$

Prior distributions :

$$\mu_a, \mu_{k_A}, \mu_{k_B} \sim N_+(20, 20)$$

$$\mu_{\phi_{1,i}} \sim N_+(0.3, 0.3)$$

$$\mu_{\phi_{2,i}} \sim N_+(0.5, 0.5)$$

$$\mu_{\beta} \sim N_+(10, 10)$$

$$\mu_{\sigma} \sim N_+(1, 3)$$

$$\mu_{\gamma} \sim N_+(50, 30)$$

$$\sigma_a, \sigma_{k_A}, \sigma_{k_B}, \sigma_{\phi_{1,i}}, \sigma_{\phi_{2,i}}, \sigma_{\beta}, \sigma_{\sigma}, \sigma_{\gamma} \sim \Gamma(0.001, 0.001)$$

LCA

Data level :

$$(RT_i, resp_i) \sim LCA(a_i, k_{F,i}, k_{G,i},$$

$$I_{0,i}, \lambda_i, \beta_i, \sigma_i, \gamma_i)$$

Group level :

$$\begin{aligned}a_i &\sim TN(\mu_a, \sigma_a, 0, 100) \\k_{F,i} &\sim TN(\mu_{k_F}, \sigma_{k_F}, 0, 200) \\k_{G,i} &\sim TN(\mu_{k_G}, \sigma_{k_G}, 0, 200) \\I_{0,i} &\sim TN(\mu_{I_0}, \sigma_{I_0}, 0, 50) \\\lambda_i &\sim TN(\mu_\lambda, \sigma_\lambda, 0, 50) \\\beta_i &\sim TN(\mu_\beta, \sigma_\beta, 0, 1) \\\sigma_i &\sim TN(\mu_\sigma, \sigma_\sigma, 0, 40) \\\gamma_i &\sim TN(\mu_\gamma, \sigma_\gamma, 0, 400)\end{aligned}$$

Prior distributions :

$$\begin{aligned}\mu_a, \mu_{k_F}, \mu_{k_G} &\sim N_+(20, 20) \\\mu_{I_0}, \mu_\sigma &\sim N_+(1, 3) \\\mu_\lambda &\sim N_+(0.8, 0.4) \\\mu_\beta &\sim N_+(0.5, 0.5) \\\mu_\gamma &\sim N_+(50, 30) \\\sigma_a, \sigma_{k_F}, \sigma_{k_G}, \sigma_{I_0}, \sigma_\lambda, \sigma_\beta, \sigma_\sigma, \sigma_\gamma &\sim \Gamma(0.001, 0.001)\end{aligned}$$

MLBA

Data level :

$$(RT_i, resp_i) \sim MLBA(A_i, b_i, t_{0,i}, I_{0,i}, \lambda_{1,i}, \lambda_{2,i}, \beta_i, m_i, \gamma_i)$$

Group level :

$$\begin{aligned}A_i &\sim TN(\mu_A, \sigma_A, 0, Inf) \\b_i - A_i &\sim TN(\mu_b, \sigma_b, 0, Inf) \\t_{0,i} &\sim TN(\mu_{t_0}, \sigma_{t_0}, 0, Inf) \\I_{0,i} &\sim TN(\mu_{I_0}, \sigma_{I_0}, 0, Inf) \\\lambda_{1,i} &\sim TN(\mu_{\lambda_1}, \sigma_{\lambda_1}, 0, Inf) \\\lambda_{2,i} &\sim TN(\mu_{\lambda_2}, \sigma_{\lambda_2}, 0, Inf) \\\beta_i &\sim TN(\mu_\beta, \sigma_\beta, 0, Inf) \\m_i &\sim TN(\mu_m, \sigma_m, 0.2, 5) \\\gamma_i &\sim TN(\mu_\gamma, \sigma_\gamma, 0, Inf)\end{aligned}$$

Prior distributions :

$$\begin{aligned}\mu_A, \mu_b &\sim N_+(2, 2) \\\mu_{I_0}, \mu_\beta &\sim N_+(1, 3) \\\mu_{\lambda_1}, \mu_{\lambda_2} &\sim N_+(0.3, 0.3) \\\mu_{t_0} &\sim N_+(0.5, 0.5) \\\mu_m &\sim N_+(1, 0.5) \\\mu_\gamma &\sim N_+(2, 3) \\\sigma_A, \sigma_b, \sigma_{t_0}, \sigma_{I_0}, \sigma_{\lambda_1}, \sigma_{\lambda_2}, \sigma_\beta, \sigma_m, \sigma_\gamma &\sim \Gamma(0.001, 0.001)\end{aligned}$$

AAM

Data level :

$$(RT_i, resp_i) \sim AAM(a_i, k_{F,i}, k_{G,i}, k_{scale,i}, \lambda_i, \beta_i, \sigma_i, \alpha_i, \gamma_i)$$

Group level :

$$\begin{aligned}a_i &\sim TN(\mu_a, \sigma_a, 0, 100) \\k_{F,i} &\sim TN(\mu_{k_F}, \sigma_{k_F}, 0, 200) \\k_{G,i} &\sim TN(\mu_{k_G}, \sigma_{k_G}, 0, 200) \\k_{scale,i} &\sim TN(\mu_{k_{scale}}, \sigma_{k_{scale}}, 0, 50) \\\lambda_i &\sim TN(\mu_\lambda, \sigma_\lambda, 0, 50) \\\beta_i &\sim TN(\mu_\beta, \sigma_\beta, 0, 1) \\\sigma_i &\sim TN(\mu_\sigma, \sigma_\sigma, 0, 40) \\\alpha_i &\sim TN(\mu_\alpha, \sigma_\alpha, 0, 50) \\\gamma_i &\sim TN(\mu_\gamma, \sigma_\gamma, 0, 400)\end{aligned}$$

Prior distributions :

$$\begin{aligned}\mu_a, \mu_{k_F}, \mu_{k_G} &\sim N_+(20, 20) \\\mu_{k_{scale}}, \mu_\sigma &\sim N_+(1, 3) \\\mu_\lambda &\sim N_+(4, 4) \\\mu_\beta, \mu_\alpha &\sim N_+(0.5, 0.5) \\\mu_\gamma &\sim N_+(50, 30) \\\sigma_a, \sigma_{k_F}, \sigma_{k_G}, \sigma_{k_{scale}}, \sigma_\lambda, \sigma_\beta, \sigma_\sigma, \sigma_\alpha, \sigma_\gamma &\sim \Gamma(0.001, 0.001)\end{aligned}$$

Note that the truncation used in MDFT, LCA and AAM was included to prevent extreme parameter scaling, which is a result of the models being unidentifiable.

Maximum likelihood boundaries

MDFT

LCA

$$\begin{aligned}a &= [0, 100] & a &= [0, 100] \\k_A &= [0, 200] & k_F &= [0, 200] \\k_B &= [0, 200] & k_G &= [0, 200] \\\phi_1 &= [0, 10] & I_0 &= [0, 50] \\\phi_2 &= [0, 30] & \lambda &= [0, 50] \\\beta &= [0, 40] & \beta &= [0, 1] \\\sigma &= [0, 40] & \sigma &= [0, 40] \\\gamma &= [0, 400] & \gamma &= [0, 400]\end{aligned}$$

MLBA

AAM

$$\begin{aligned}A &= [0, 10] & a &= [0, 100] \\b - A &= [0, 10] & k_F &= [0, 200] \\t_0 &= [0, 1.2] & k_G &= [0, 200] \\I_0 &= [0, 8] & k_{scale} &= [0, 50] \\\lambda_+ &= [0, 0.8] & \lambda &= [0, 50] \\\lambda_- &= [0, 0.8] & \beta &= [0, 1] \\\beta &= [0, 30] & \sigma &= [0, 40] \\m &= [0.2, 5] & \alpha &= [0, 50] \\\gamma &= [0, 30] & \gamma &= [0, 400]\end{aligned}$$

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Berkowitsch, N. A., Scheibehenne, B., & Rieskamp, J. (2014). Rigorously testing multialternative decision field theory against random utility models. *Journal of Experimental Psychology: General*, 143(3), 1331.
- Berkowitsch, N. A., Scheibehenne, B., Rieskamp, J., & Matthäus, M. (2015). A generalized distance function for preferential choices. *British Journal of Mathematical and Statistical Psychology*, 68(2), 310–325.
- Bhatia, S. (2013). Associations and the accumulation of preference. *Psychological Review*, 120(3), 522.
- Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, 57, 153–178.
- Bussemeyer, J. R., & Diederich, A. (2002). Survey of decision field theory. *Mathematical Social Sciences*, 43(3), 345–370.
- Bussemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences*, 23(3), 255–282.
- Bussemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision-making in an uncertain environment. *Psychological Review*, 100(3), 432.
- Bussemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, 44(1), 171–189.
- Cataldo, A. M., & Cohen, A. L. (2018). Reversing the similarity effect: The effect of presentation format. *Cognition*, 175, 141–156.
- Cohen, A. L., Kang, N., & Leise, T. L. (2017). Multi-attribute, multi-alternative models of choice: Choice, reaction time, and process tracing. *Cognitive Psychology*, 98, 45–72.
- Donkin, C., Brown, S., Heathcote, A. J., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: Different models for response time, same conclusions about psychological mechanisms? *Psychonomic Bulletin & Review*, 55, 140–151.
- Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P., & Donkin, C. (2018). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-017-1417-2>
- Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, 53(2), 134.
- Evans, N. J., & Brown, S. D. (2017). People adopt optimal policies in simple decision-making, after practice and guidance. *Psychonomic Bulletin & Review*, 24(2), 597–606.
- Evans, N. J., & Brown, S. D. (2018). Bayes factors for the linear ballistic accumulator model of decision-making. *Behavior Research Methods*, 50(2), 589–603.
- Evans, N. J., Hawkins, G. E., Boehm, U., Wagenmakers, E.-J., & Brown, S. D. (2017a). The computations that support simple decision-making: A comparison between the diffusion and urgency-gating models. *Scientific Reports*, 7, 16433.
- Evans, N. J., Howard, Z. L., Heathcote, A., & Brown, S. D. (2017b). Model flexibility analysis does not measure the persuasiveness of a fit. *Psychological Review*, 124(3), 339.
- Evans, N. J., Rae, B., Bushmakin, M., Rubin, M., & Brown, S. D. (2017c). Need for closure is associated with urgency in perceptual decision-making. *Memory & Cognition*, 45(7), 1193–1205.
- Evans, N. J., Brown, S. D., Mewhort, D. J., & Heathcote, A. (2018). Refining the law of practice. *Psychological Review*, 125(4), 592.
- Evans, N. J., Steyvers, M., & Brown, S. D. (2018). Modeling the covariance structure of complex datasets using cognitive models: An application to individual differences and the heritability of cognitive ability. *Cognitive Science*.
- Farmer, G. D., Warren, P. A., El-Deredy, W., & Howes, A. (2016). The effect of expected value on attraction effect preference reversals. *Journal of Behavioral Decision Making*.
- Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., & Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, 3(10), e189.
- Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, 7(2), 185–207.
- Ho, T. C., Yang, G., Wu, J., Cassey, P., Brown, S. D., Hoang, N., & Yang, T. T. (2014). Functional connectivity of negative emotional processing in adolescent depression. *Journal of Affective Disorders*, 155, 65–74. <https://doi.org/10.1016/j.jad.2013.10.025>
- Holmes, W. R. (2015). A practical guide to the probability density approximation (PDA) with improved implementation and error characterization. *Journal of Mathematical Psychology*, 68, 13–24.
- Holmes, W. R., & Trueblood, J. S. (2018). Bayesian analysis of the piecewise diffusion decision model. *Behavior Research Methods*, 50(2), 730–743.
- Holmes, W. R., Trueblood, J. S., & Heathcote, A. (2016). A new framework for modeling decisions about changing information: The piecewise linear ballistic accumulator model. *Cognitive Psychology*, 85, 1–29.
- Hotaling, J. M., Bussemeyer, J. R., & Li, J. (2010). Theoretical developments in decision field theory: A comment on K. Tsetsos, N. Chater, and M. Usher. *Psychological Review*, 117, 1294–1298.
- Howes, A., Warren, P. A., Farmer, G., El-Deredy, W., & Lewis, R. L. (2016). Why contextual preference reversals maximize expected value. *Psychological Review*, 123(4), 368.
- Huang, K., Sen, S., & Szidarovszky, F. (2012). Connections among decision field theory models of cognition. *Journal of Mathematical Psychology*, 56(5), 287–296.
- Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 9, 90–98.
- Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods*, 49(2), 513–537.
- Liew, S. X., Howe, P. D., & Little, D. R. (2016). The appropriacy of averaging in the study of context effects. *Psychonomic Bulletin & Review*, 23(5), 1639–1646.
- Miletić, S., Turner, B. M., Forstmann, B. U., & van Maanen, L. (2017). Parameter recovery for the leaky competing accumulator model. *Journal of Mathematical Psychology*, 76, 25–50.
- Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44(1), 190–204.
- Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4(1), 79–95.
- Nosofsky, R. M., & Palmeri, T. J. (2015). An exemplar-based random-walk model of categorization and recognition. In *The Oxford handbook of computational and mathematical psychology* (p. 142). Oxford University Press, USA.
- Parrish, A. E., Evans, T. A., & Beran, M. J. (2015). Rhesus macaques (*Macaca mulatta*) exhibit the decoy effect in a perceptual discrimination task. *Attention, Perception, & Psychophysics*, 77(5), 1715–1725.

- Pettibone, J. C. (2012). Testing the effect of time pressure on asymmetric dominance and compromise decoys in choice. *Judgment and Decision Making*, 7(4), 513.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108.
- 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.
- Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108, 370–392.
- Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, 16, 158–174.
- Soltani, A., De Martino, B., & Camerer, C. (2012). A range-normalization model of context-dependent choice: A new model and evidence. *PLoS Computational Biology*, 8(7), 1–15.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.
- Ter Braak, C. J. (2006). A Markov chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces. *Statistics and Computing*, 16(3), 239–249.
- Trueblood, J. S. (2012). Multi-alternative context effects obtained using an inference task. *Psychonomic Bulletin & Review*, 19(5), 962–968.
- Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, 121(2), 179.
- Trueblood, J. S., Brown, S. D., & Heathcote, A. (2015). The fragile nature of contextual preference reversals: Reply to Tsetsos, Chater, and Usher (2015). *Psychological Review*, 122(4), 848–853.
- Trueblood, J. S., Brown, S. D., Heathcote, A., & Busemeyer, J. R. (2013). Not just for consumers: Context effects are fundamental to decision-making. *Psychological Science*, 24, 901–908.
- Trueblood, J. S., Holmes, W. R., Seegmiller, A. C., Douds, J., Compton, M., Szentirmai, E., & Eichbaum, Q. (2018). The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making. *Cognitive Research: Principles and Implications*, 3(1), 28.
- Trueblood, J. S., & Pettibone, J. C. (2017). The phantom decoy effect in perceptual decision making. *Journal of Behavioral Decision Making*, 30(2), 157–167.
- Tsetsos, K., Chater, N., & Usher, M. (2015). Examining the mechanisms underlying contextual preference reversal: Comment on Trueblood, Brown, and Heathcote (2014). *Psychological Review*, 122(4), 838–847.
- Tsetsos, K., Usher, M., & Chater, N. (2010). Preference reversal in multi-attribute choice. *Psychological Review*, 117, 1275–1291.
- Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018). Competing models of multi-attribute, multi-alternative preferential choice. *Psychological Review*, 125, 329–362.
- Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review*, 21(2), 227–250.
- Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, 18(3), 368.
- Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, 79, 281–299.
- Usher, M., Elhalal, A., & McClelland, J. L. (2008). The neurodynamics of choice, value-based decisions, and preference reversal. In Chater, N., & Oaksford, M. (Eds.) *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 277–300). Oxford: Oxford University Press.
- Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3), 550.
- Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, 111, 757–769.
- van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2012). A diffusion model decomposition of the effects of alcohol on perceptual decision making. *Psychopharmacology*, 219(4), 1017–1025. <https://doi.org/10.1007/s00213-011-2435-9>
- Wollschlaeger, L. M., & Diederich, A. (2012). The 2n-ary choice tree model for n-alternative preferential choice. *Frontiers in Cognitive Science*, 3, 1–11.