Empirical Comparison of the Adjustable Spanner and the Adaptive Toolbox Models of Choice

Antonia Krefeld-Schwalb[1], Chris Donkin[2], Ben R. Newell[2], Benjamin Scheibehenne[1]

[1] University of Geneva, [2] School of Psychology, University of New South Wales

Author Note

Corresponding author's address: Antonia Krefeld-Schwalb, Geneva School of Economics and Management, Uni Mail, Bd du Pont-d'Arve 40, CH- 1112 Geneva 4

Phone: + 41 22 379 89 97

Email: antonia.krefeld-schwalb@unige.ch

Abstract

Past research indicates that individuals respond adaptively to contextual factors in multi-attribute choice tasks. Yet it remains unclear how this adaptation is cognitively governed. In this paper, empirically testable implementations of two prominent competing theoretical frameworks are developed and compared across two multi-attribute choice experiments: The Adaptive Toolbox framework assuming discrete choice strategies and the Adjustable Spanner framework assuming one comprehensive adaptive strategy. Results from two experiments indicate that in the environments we tested, in which all cue information was presented openly, the Toolbox makes better predictions than the Adjustable Spanner both in- and out-of-sample. Follow-up simulation studies indicate that it is difficult to discriminate the models based on choice outcomes alone but allowed the identification of a small sub-set of cases where the predictions of both models diverged. Our results suggest that people adapt their decision strategies by flexibly switching between using as little information as possible and use of all of the available information.

Keywords: Choice models; evidence accumulation; decision strategies.

Empirical Comparison of the Adjustable Spanner and the Adaptive Toolbox Models of Choice

Making every day judgments and decisions typically requires a trade-off between the time and effort one spends on the task and the quality of its outcome. Behavioral scientists commonly agree that this trade-off is adaptive such that people adjust their effort and the amount of information they process depending on the respective context and the goals they want to achieve (Payne, Bettman, & Johnson, 1988; Simon, 1956). However, there is an ongoing discussion regarding how this adaptation takes place (Chater, Oaksford, Nakisa, & Redington, 2003; Lee & Cummins, 2004; Marewski & Schooler, 2011; Söllner & Bröder, 2016).

Central to this debate is whether adaptation is reflected by distinct strategies (Gigerenzer & Todd, 1999), or by continuous differences in the evidence that is accumulated in a given environment (Newell, 2005). Understanding the mechanisms underlying this adaptive process could elucidate decision making in a wide range of contexts, including consumer behavior (Scheibehenne, Miesler & Todd, 2007; Scheibehenne, von Helversen, & Rieskamp, 2015) and managerial decisions (Artinger, Petersen, Gigerenzer & Weibler, 2015). Moreover, solutions for how to approach this specific debate in judgment and decision making research can inform related debates in other areas of cognition. For example, the debates between discrete-state versus continuous models of recognition memory (Batchelder & Alexander, 2013; Bröder & Schutz, 2009; Pazzaglia, Dube & Rotello, 2013) or between different models of working memory capacity (Zhang & Luck, 2008; Bays & Husain, 2008; Cowan 2005; Donkin, Kary, Tahir & Taylor, 2016) share features with that of the multi-attribute choice debate. Indeed, a key element of all of these debates is that the models' predictions often overlap due to the fact that competing models are generated to explain the

same behavior (e.g. Newell, 2005). We propose, as a solution to this problem, a method to identify experimental designs that will discriminate between the models in a first step, and test the models' predictions in a second step (see optimal experimental design as an alternative approach to the same problem; Myung & Pitt, 2009).

In our model comparison we focus on the two most prominent theoretical frameworks that have been proposed to elucidate the processes underlying multi-attribute choice (Söllner & Bröder, 2016; Busemeyer, 2017). One framework assumes that decision makers adapt their behavior by applying qualitatively different cognitive strategies. In analogy to a craftswoman who selects her tools depending on the requirement of the job she faces, this idea is sometimes referred to as an Adaptive Toolbox (Gigerenzer & Todd, 1999). The alternative framework adopts a single, albeit more comprehensive cognitive process to model adaptation. This alternative class of models predict that evidence or information about alternative options is sequentially sampled and accumulated until a certain threshold is reached (Busemeyer & Townsend, 1993; Lee & Cummins, 2004). Thus, the threshold defines how much information is integrated and by adjusting it, sequential sampling models can account for adaptive changes in behavior. This idea is sometimes, in analogy to the Toolbox metaphor, referred to as an Adjustable Spanner (Newell, 2005; Söllner & Bröder, 2016).

Both frameworks aim at explaining adaptive decision making as it has been observed empirically. For example, the affect richness of the choices (Suter, Pachur & Hertwig, 2016), incidental emotions (Scheibehenne & von Helversen, 2015), the type of learning task (Pachur & Olsson, 2012), whether the task involves information search in memory (Bröder & Schiffer, 2003a) or the distribution of the cues' validities (Mata, Schooler & Rieskamp, 2007) can all influence the application of qualitatively different strategies. Likewise, contextual factors have been shown to influence the threshold of the sequential sampling process (Lee,

Newell & Vandekerckhove, 2014; Newell & Lee, 2011; Simen, Cohen & Holmes, 2006). However, it remains unclear which of the two frameworks provides a better description of human choice in complex situations across different environments.

This lack of clarity is partly due to the fact that both models have not yet been compared on empirical choice data, while both being implemented as computational models that account for *inter*- and *intra*-individual differences. We aim to overcome this gap in the literature by comparing specific computational implementations of the Adaptive Toolbox and the Adjustable Spanner on empirical choice data. Moreover, we will implement the models in a highly comparable manner, so that they differ only with regard to whether distinct strategies or a continuous threshold of evidence accumulation governs how much evidence is considered in a given choice task. In other words, we maximize the similarity of the models in order to rigorously focus the model comparison on the most relevant aspect of the debate.

Two competing frameworks: the Adaptive Toolbox and the Adjustable Spanner

Various versions of the Adaptive Toolbox have been proposed in the literature (Gigerenzer & Todd, 1999). For example, a typical Toolbox includes one simple non-compensatory (take-the-best, TTB) strategy, and a more complex compensatory (weighted additive, WADD) strategy (Scheibehenne, Rieskamp, & Wagenmakers, 2013). The simple TTB strategy predicts that people search for the best (i.e. most valid) cue that discriminates between the options and do not take further information into account. The option that scores highest on the most valid, discriminating cue is chosen. In other words, further information cannot compensate for (or over-ride) the initial discriminating cue. In contrast, the more complex compensatory (WADD) strategy predicts that cues lower in predictive validity can – when combined and weighted appropriately – overcome (compensate for) the information provided by a single more valid cue.

Consider a hypothetical scenario where a manager wants to predict which of two upcoming Hollywood movies will be more successful at the box office. As a basis for this prediction, she can refer to the recommendations of six movie critics (= cues) who differ in their predictive ability (= cue validity). According to TTB, the manager would only rely on the one critic with the highest predictive validity. If the best critic does not discriminate between the two movies, the manager would consider the recommendations of the second-best critic, and so on until a decision is made.

In stark contrast to TTB, the compensatory WADD strategy predicts that all available cue values are taken into account and weighted by their respective validities. The option with the highest weighted sum is then chosen. Faced with the same movie selection, a manager following WADD would consider the recommendations of all the critics available and weight each critic's recommendation with their respective validity. The Adaptive Toolbox, incorporating both TTB and WADD, further assumes that decision makers select the respective strategy depending on situational and personal characteristics (e.g. Bröder, 2003; Marewski & Schooler, 2011; Rieskamp & Otto, 2006; Newell & Lee, 2011), rather than consistently relying on one strategy (Bröder, 2000; Newell & Shanks, 2003; Newell, Weston, & Shanks, 2003; van Ravenzwaaij, Moore, Lee, & Newell, 2014).

While there is an ongoing discussion on the appropriateness of Toolbox models for explaining and predicting behavior in multi-attribute choice tasks (Bröder & Newell, 2008), simple Toolboxes containing only TTB and WADD have been shown to predict behavior more accurately in such tasks than more complex toolboxes, such as a Toolbox that also includes a tallying strategy (Scheibehenne, Rieskamp, & Wagenmakers, 2013). For example, Scheibehenne et al. (2013) compared a Bayesian implementation of a Toolbox, containing only TTB and WADD to single decision strategies and a more complex Toolbox. Those

comparisons provided support for the superiority of this rather simple Toolbox model over both more complex and single decision strategies (Scheibehenne et al., 2013).

The Adjustable Spanner assumes that the decision maker samples evidence from the environment until a threshold level of evidence is reached (Hausmann & Läge, 2008; Lee & Cummins, 2004; Newell & Lee, 2011). In the movie critics' example, the Adjustable Spanner would assume that the manager would consider each of the critics' recommendations, weighted by their validities, until there was enough evidence for one of the two movies, as indicated by an individual threshold of evidence accumulation.

The placement of the threshold for stopping evidence accumulation and making a decision has a critical influence on the model's behavior. A very low threshold makes the Adjustable Spanner mimic the TTB strategy, since people would integrate only a minimal amount of information. On the other hand, a sufficiently high threshold would mean that all of the available information is used when making a decision, thus mimicking WADD. This flexibility in threshold setting allows the Adjustable Spanner to mimic both of the strategies in the Adaptive Toolbox we consider here (Newell, 2005; Newell & Lee, 2011) as well as permitting a process that can capture choices that deviate from the predictions of the TTB and WADD strategies. For example, if both TTB and WADD predict the same choice, an evidence threshold that governs evidence accumulation in between accumulating all or only the best discriminating information can lead to a different choice.

Earlier comparisons of the two competing frameworks

There have been a number of attempts to contrast the predictions of Toolbox and evidence accumulator frameworks (see Newell & Bröder, 2008). However, we know of no attempt, to date, to compare the two types of models in their complete forms on their ability to quantitatively capture and predict empirical choice data. This lack of direct comparisons in

part reflects the difficulty of generating divergent predictions from the two types of models. This difficulty in turn stems from insufficiently specified instantiations of models loosely grouped under the ‘toolbox’ and ‘spanner’ metaphors.

Earlier approaches tried to solve these issues by restricting the testing environments such that the differences between the model classes are more readily observable (Lee & Cummins, 2004; Newell, Collins, & Lee, 2007; Newell & Lee, 2011). These comparisons typically revealed superiority of their respective implementations of evidence accumulation models. According to the authors of these comparison studies, this superiority arose due to two main factors. First, the evidence accumulation models more readily accommodated the presence of intra-individual consistency but inter-individual differences in strategy use in the same decision environment (e.g. Lee & Cummins, 2004; Newell & Lee, 2011; van Ravenzwaaij et al., 2014). The authors argue that such a pattern is difficult to reconcile with a toolbox approach in which environmental (rather than individual-level) constraints are thought to be the primary drivers of strategy selection/adaptation (e.g. Gigerenzer & Todd, 1999). In contrast, a sequential sampling model which views TTB and WADD as extremes on a continuum of evidence accumulation can explain such a pattern by assuming that individuals would either choose exclusively in line with a compensatory or a non-compensatory decision rule, depending on a preferred evidence threshold.

The second aspect favoring the evidence accumulation framework proposed in earlier studies is the presence of behavior that falls outside the deterministic prescriptions of the TTB and WADD strategies. Specifically, the Adjustable Spanner metaphor can accommodate stopping rules (for information acquisition) that are intermediate between ‘one-cue’ (TTB) and ‘all-information’ (WADD) (e.g. Lee et al., 2014; Söllner, Bröder, Glöckner, & Betsch,

2014; Söllner & Bröder, 2016), or cue-search orders that do not rely solely on cue-validity (e.g. van Ravenzwaaij et al., 2014).

There are however, some limitations to these previous comparison attempts. One is that despite using methods to make the comparisons between the models 'fair', this has not always been satisfactorily achieved. For example, the naive strategy selection model used as comparator to an accumulator model by Newell and Lee (2011) can be criticized on the grounds that it was too complex for the task they used and thus unfairly punished by the measure of model fit they adopted (i.e. minimum description length).

A second limitation is that previous attempts have tended to focus on a single decision environment – that is one in which cue validities remain the same (e.g. Hausmann & Läge, 2008; Lee & Cummins, 2004; Newell et al., 2007; Newell & Lee, 2011; though see Lee, Newell & Vandekerckhove, 2014 for an exception). Conducting comparisons across environments with different cue structures is important, given earlier findings indicating that the environment does influence the decision strategies applied (Bröder 2000, 2003; Bröder & Schiffer, 2003b; Rieskamp & Hoffrage, 1999; Rieskamp, 2006; Rieskamp & Otto, 2006) as well as thresholds in an accumulation process (Lee, Newell & Vandekerckhove, 2014; Simen, Cohen & Holmes, 2013). In particular, recent studies show that the dispersion of the cue validities influences the amount of information search, with high dispersion favoring less information search and low dispersion favoring more information search (Mata, Schooler, Rieskamp, 2011).

A final key reason is that a complete quantitative comparison has not been possible because of the lack of precise, directly comparable formalizations of models inspired by the Spanner and Toolbox metaphors. While a probabilistic model of the Adaptive Toolbox was recently implemented by Scheibehenne et al. (2013), there has up to now not been a

comparable computational implementation of the Adjustable Spanner. We provide this instantiation here by building on earlier approaches (Lee & Cummins, 2004; Newell, Collins, & Lee, 2007; Newell & Lee, 2011). Armed with formal versions of both choice frameworks, we can make a direct comparison between Spanner and Toolbox models on empirical grounds (Farrell & Lewandowsky, 2010).

In the following, we present a rigorous comparison of these models in several decision environments differing in the dispersion of cue validities, in order to examine the adaptability of the models to different environments and their ability to capture individual differences in decision making. First, we propose a formal version of the Adjustable Spanner to the literature. We then note that the predictions of the models are largely overlapping in most choice environments. To deal with this problem, we apply a methodology that is similar to, but distinct from, optimal experimental design. In our experiment, participants completed two sessions of data collection. The first session yielded data to which we fitted both models, allowing us to identify a set of decisions for each participant that would minimize the mimicry between the models. In the second session, the same participants returned to the lab to make choices in the set of trials we had selected for them. This second session served as a generalization test to assess each models' ability to predict a second set of choices (see Scheibehenne, Rieskamp, & González-Vallejo, 2009, for a similar approach).

Model Specification

In the following we formally specify an Adaptive Toolbox and an Adjustable Spanner (i.e. a sequential sampling model) for multi-attribute binary choices between two options, labelled A and B[1].

[1] The R function for both models and all data can be found online: https://osf.io/e8h3e/

The Adaptive Toolbox

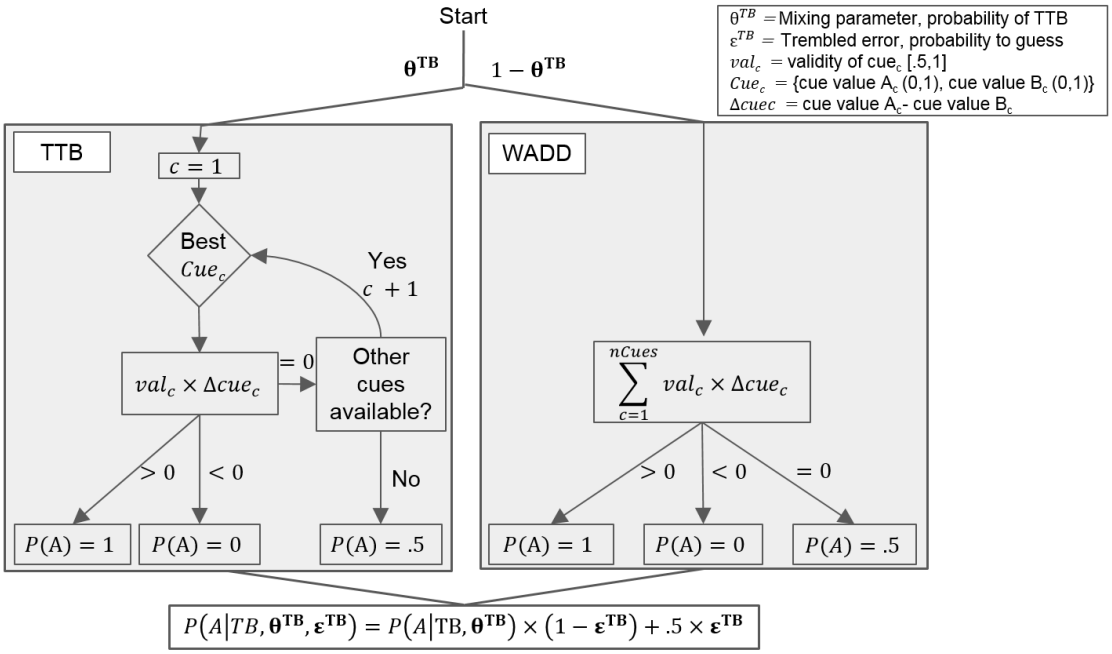

Figure 1. Flow diagram of the Adaptive Toolbox consisting of TTB (left box) and WADD (right box). The probabilistic prediction ranging from 0 to 1 is expressed as the probability of choosing option A out of a set of two options A and B, (i.e. the likelihood). It is determined by the mixing parameter θ^{TB}, the trembling hand error ε^{TB} and the predictions of both strategies being either 0, 1 or .5; c represents the index of the cues.

In order to implement a Toolbox consisting of WADD and TTB, we relied on an earlier implementation by Scheibehenne, Rieskamp, and Wagenmakers (2013) where the probability of an individual decision maker to select TTB over WADD is governed by a free "mixing" parameter θ^{TB}. Accordingly, the probability of selecting WADD is 1-θ^{TB}. To allow for the possibility of inconsistent choices or application errors when using a particular strategy, an explicit error term ε^{TB} was included indicating the probability that a decision is made at random (i.e. a so-called "trembling hand" error, Loomes, Moffatt, & Sugden, 2002). Here, an error of $\varepsilon^{TB} = 1$ indicates pure guessing, whereas $\varepsilon^{TB} = 0$ indicates perfect consistency with the predictions of the respective strategies. Figure 1 illustrates the model

comprising both strategies as a flow diagram, illustrating TTB on the left side and WADD on the right side of the diagram.

As a first step in the diagram, θ^{TB} indicates the probability by which a strategy is chosen. If the TTB strategy is used, a choice is made if the best available cue differentiates between the options. If the cue does not discriminate (i.e. if the difference between the cue values is zero), the second best cue is considered, and so on, until a decision can be made. If no cue discriminates between the options (i.e. if the options have an identical cue pattern), TTB reverts to guessing, thus random choice between the options. If WADD is used, all available evidence is considered, and all cue values are weighted with their validities and summed. The option with the highest weighted sum is then chosen. If the weighted sum of both options is equal, the model selects either option with equal probability. Both strategies make predictions for the probability to choose the option A that are either 0, .5 or 1. However in combination with the mixing parameter and the trembled hand error, the likelihood as indicated in the white box at the bottom of the flow diagram in Figure 1 can take on any value between 0 and 1. To sum up, the toolbox model has two free parameters the mixing parameter θ^{TB} and the trembling hand error ε^{TB}.

The Adjustable Spanner

To implement the Adjustable Spanner proposed by Newell (2005) and Newell and Lee (2009), we defined a threshold δ^{ACC} that determines the proportion of information a decision maker considers relative to the maximum possible evidence in a given environment. Hence, the threshold δ^{ACC} is scaled between 0, indicating that only the evidence of the first cue is accumulated and 1 indicating that the evidence of all cues is accumulated. The maximum information in a trial is thus given by the sum of the cue validities, and the threshold indicates how much of this is encountered in any trial. This corresponds to a fixed

number of cues in one environment but can correspond to a different number of cues in different environments. Thus, the notion "threshold" δ^{ACC} in this implementation, diverges from the evidence threshold as defined in many other accumulation models. After the threshold is crossed, the accumulated values of the attended cues weighted with their respective validities determine the choice. The option with the most accumulated evidence is then chosen. If the accumulated evidence is indecisive, the next cue is accumulated until either the accumulated evidence is decisive, or all information is accumulated. In the latter case, the model reverses to guessing.

Several alternative implementations of an Adjustable Spanner are possible. The model implementation at hand was chosen because it captures the main assumptions of the metaphor and provided a suitable account of the observed choice data. The online supplementary material contains an alternative implementation of the Adjustable Spanner[2]. It is worth noting that the strategies of the Toolbox, TTB and WADD, are special cases of this implementation of the Adjustable Spanner. The highest threshold of the Adjustable Spanner $\delta^{ACC} = 1$ corresponds to the mixing probability of the Toolbox $\theta^{TB} = 0$ and $\delta^{ACC} = 0$ corresponds to $\theta^{TB} = 1$. Figure 2 illustrates the model as a flow diagram, similar to that of the Toolbox outlined above. The threshold δ^{ACC} defines how many cues are searched in a given environment, correspondingly information search is stopped if the summed validities of the cues reach a critical value of accumulated validities. The critical value is a function of δ^{ACC} and the distribution of validities in the trial. After crossing this critical value a choice is made if the accumulated evidence (the cue values weighted with their validities) up to this point

[2] Although the core assumption of the model is captured, the notion of a two-step 'accumulate-then-weight' model is perhaps not the most direct interpretation of evidence accumulation. Previous implementations (e.g. Lee and Cummins, 2004) propose that the evidence accumulated after each cue is acquired reflects the weight of that cue (cue value multiplied by its validity) and that the threshold is on the weight not the number of cues. As outlined in the supplementary material, this version of the spanner model did not provide an adequate fit for the current data; speculations as to why this was the case are included in the General Discussion.

indicates a preference for one of the options. If no preference is indicated more evidence is accumulated. In resemblance to the implementation of the Toolbox, a trembling hand error (Loomes et al., 2002) was implemented, assuming that participants guess with probability ε^{ACC}. This addition means that the model has also two free parameters, the threshold parameter δ^{ACC} and the trembling hand error ε^{ACC}.

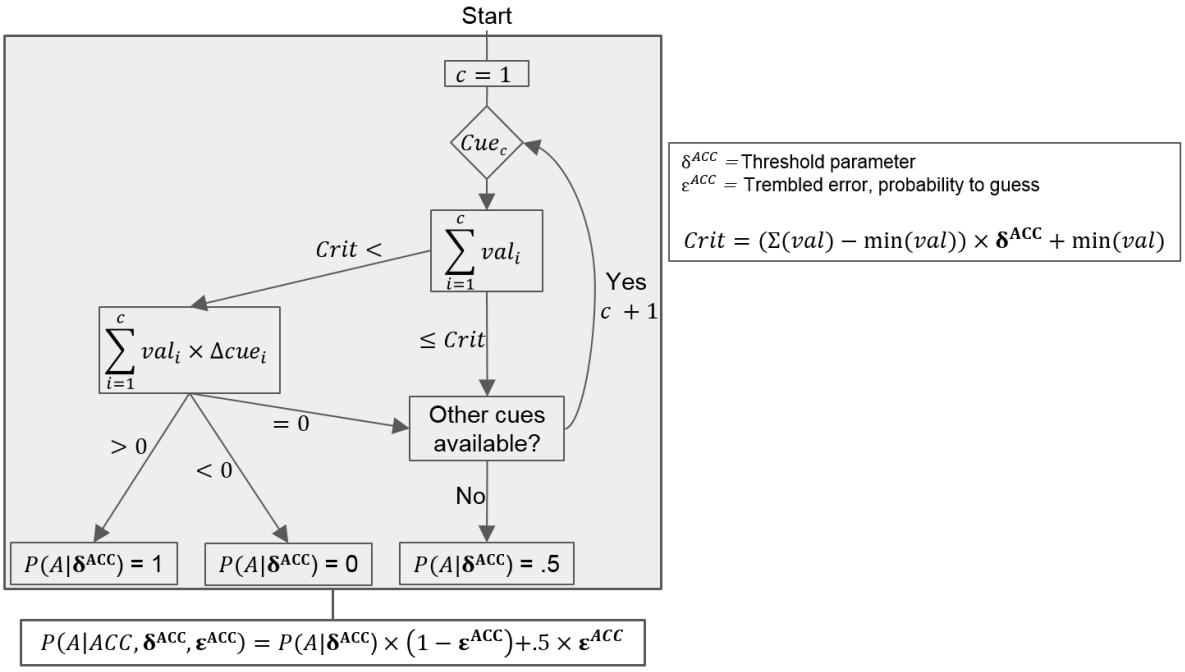

Figure 2. Flow diagram of the Adjustable Spanner. The plate indicates the flow of the accumulation of evidence up to the critical value *Crit*, determined by the individual threshold δ^{ACC} as defined in the white plate on the right. It illustrates the prediction of the choice of option A, P(A| δ^{ACC}) and in combination with the trembling hand error ε^{ACC}, P(A|ACC, δ^{ACC}, ε^{ACC}) shown on the bottom row.

It is worth noting that the strategies of the Toolbox, TTB and WADD, are special cases of this implementation of the Adjustable Spanner. The highest threshold of the Adjustable Spanner corresponds to the mixing probability of the Toolbox $\theta^{TB}= 0$ and $\delta^{ACC} = 0$ corresponds to $\theta^{TB}= 1$.

General Overview of the Experiments

In order to compare the models on empirical grounds, we conducted an experimental procedure that consisted of two experimental sessions, a and b. The first session was used to fit the competing models to choice data and estimate best-fitting parameters for every individual across three different environments defined by the distribution of cue validities. The latter manipulation aims at causing variability in decision making and subsequently, comparing the models' ability to accommodate this variability. By having different decision environments, we are able to compare the models' adaptability, which is one of the core features of the two frameworks being compared. Further, fitting the models to each participant's data allows us to evaluate the models' ability to account for individual differences. It is one of the crucial features of cognitive models that they can be applied to describe individual differences with latent variables (Riefer et al., 2002). Moreover, fitting the models to different conditions and participants, and subsequently evaluating the models' fit on the participant and the average level allows us to evaluate the models' flexibility. The second session served as a generalization test to compare the predictions of those best-fitting models (Busemeyer & Myung, 1992).

In both sessions of the experiment, participants performed a binary multi-attribute choice task. In this task participants chose between two options, described with six binary cues and their corresponding validities for each cue. As a cover story, we asked participants to choose which of two movies they think will be more successful at the box office. Their decision was to be guided by recommendations from six movie critics. This task was presented across three within-subject conditions, with each condition having a different environment, i.e. the distribution of cue validities being either uniform, linear or j-shaped. We expected that the lower dispersion of cue validities in the uniform condition would be

associated with an increased use of compensatory strategies and more evidence accumulation, compared to the linear and j-shaped conditions which have a higher dispersion of cue validities. There was no cost to obtaining the information provided by any given cue – all information was available on screen from the beginning of the trial. In the first session, all participants received the same set of trials under three within-subjects conditions.

In the second, generalization test, session, participants were presented with their own set of trials again under three within- subject conditions that were especially tailored to increase the number of trials in which the Toolbox and Spanner models predicted different choices. These trials were identified based on a simulation study using the best-fitting individual parameters from the first session. The procedure was repeated in a second experiment with a different group of participants. Both sessions of Experiment 1 are described in corresponding sections, namely Experiment 1a and Experiment 1b.

Experiment 1a: Model Fitting

Method

Participants. A total of 129 (F = 48%, M_{age}= 21, $Range_{age}$= [19, 37]) Business bachelor students participated in the experiment in exchange for course credit and a small chocolate bar. Seven participants were excluded due to misunderstanding the task. These participants consistently chose movies that our hypothetical critics unambiguously predicted would not succeed.[3] We aimed at a sample size of 120 valid observations in order to achieve sufficient variety in the parameter values across individuals.

Design and Material. Participants were asked to predict which of two hypothetical movies, movie *A* or movie *B*, would be more successful at the box office. To aid their choice,

[3] Excluding or including these participants does not change any of the following conclusions, since both models fail equally to predict this worse than chance performance.

six hypothetical movie critics provided recommendations (see Scheibehenne & von Helversen, 2015, for a similar design). The six movie critics differed in their predictive validities, and each critic could either recommend one, both, or none of the movies.

All information, including the cues and the cue validities were openly presented from the beginning of the trial. As shown in Figure 3, an image was used to represent each critic, and the validity of their recommendations were expressed as percentages (i.e. how often a critic had previously recommended the more successful movie). Animal heads were chosen as icons for the critics in order to avoid any gender or ethnic bias when using illustrations of human faces. The images were randomly attributed to the cues and conditions, but were constant within all trials in a condition. The validities were additionally visualized by the size of the percentages' font and the critics' icons. Cue values were illustrated with asterisks (recommended) and hyphens (not recommended), respectively.

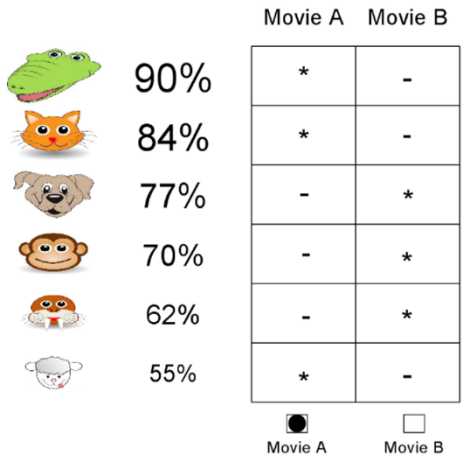

Figure 3. An example trial given to a participant. Here, the set of cue validities follows a linear distribution. The hypothetical participant chose Movie A, the option predicted by both TTB and WADD since the critic with the highest validity recommended only Movie A, and because the weighted average of the critics' choices prefer this option.

After 5 practice trials, each participant made 120 decisions, composed of 40 trials in each of three conditions. Each condition differed with regard to the distribution of the critics' validities, thus every condition represents a different environment. The validities were either approximately uniform (65%, 63%, 63%, 62%, 60%, 58%), j-shaped (90%, 69%, 68%, 66%,

63%, 60%), or linearly distributed (90%, 84%, 77%, 70%, 62%, 55%). Within each condition, the cue validities remained constant throughout both sessions of the experiment. The order of the three conditions was randomized between participants.

In each of the cue validity distribution conditions, the cue values were chosen such that in 50% of the trials TTB and WADD made the same predictions and in 50% of the trials TTB and WADD made opposite predictions. To avoid pairs of movies in which one option dominated, we used only pairs where the sum of the cue values for one option minus the sum of cue values for the other option did not exceed three. After the experiment participants were asked whether they agreed to be contacted about a possible follow-up experiment in the future (i.e. the generalization test).

Results and Discussion

Participants made sensible choices. For the trials in which both the compensatory (WADD) and non-compensatory (TTB) strategies predicted the same option, participants chose that option on 94% of trials, on average (standard deviation, *SD*, of 7%), across the three distribution conditions. For those trials in which the two strategies made different predictions, participants chose in line with WADD in 65% of the trials (*SD* = 18%).

To test which of the models provided a better description of the observed choice data, the parameters of both models were fitted with a grid search algorithm that minimized the negative log likelihood (-LL) separately for each participant and cue validity distribution condition. Minimizing –LL is equivalent to maximizing likelihood, however the –LL values are positive, making it easier to determine which values are smaller. For δ^{ACC} and θ^{TB}, the search grid spans across values between 0 and 1 in steps of 0.01. For ε^{ACC} and ε^{TB}, the grid ranged from .001 to .999, to avoid extreme values of the likelihood being zero or one, respectively. To additionally test both models, we fitted a pure guessing model to the data.

Both models provided a better fit than a pure guessing model[4] across all participants and conditions, the Toolbox provided a better fit than the Adjustable Spanner. The best-fitting -LL summed across all participants and conditions were $-LL_{TB}$ = 4,458 and $-LL_{ACC}$ = 4,797 (smaller –LL values indicate a relatively better fit). If we take the ratio of these likelihoods to form a likelihood ratio LR of the Adjustable Spanner over the Toolbox ($LR_{ACC/TB}$), the odds strongly favor the Toolbox ($LR_{ACC/TB} < .001$). Summed across the conditions the individuals -LL of the Toolbox was smaller for 73% of the participants. Comparing each model to a guessing model, using BIC to correct for differences in model complexity, the Toolbox provided a better fit for 97% of participants across all three conditions. The Adjustable Spanner performed equally well compared to the guessing model, providing a better fit than the guessing model for 97 % of participants.

In the condition with uniform distribution of validities, both models fit the data about equally well. Here, the –LL of the Toolbox was smaller than the –LL of the Adjustable Spanner for 48% of the participants. The differences between the models were larger in the other environments. The Toolbox outperformed the Spanner model for 77% of the participants in the conditions with j-shaped distribution of validities, and for 61% of the participants in the conditions with linear distribution of validities.

Figure 4illustrates the best-fitting parameters for each participant for the Toolbox (upper panel) and the Adjustable Spanner (lower panel) respectively. The Figure shows that the parameters are widely scattered, indicating a lot of individual differences, with an overall tendency to set relatively high thresholds in the Adjustable Spanner (δ^{ACC}: mean= .59, sd = .27), and for using WADD over TTB (θ^{TB}: mean= .29, sd = .26). This result is in line with

[4] The pure guessing model predicts equal choice probabilities P(A) =.5 for both options in all trials.

previous results that indicated a preference for WADD in environments with no search costs (Rieskamp & Otto, 2006).

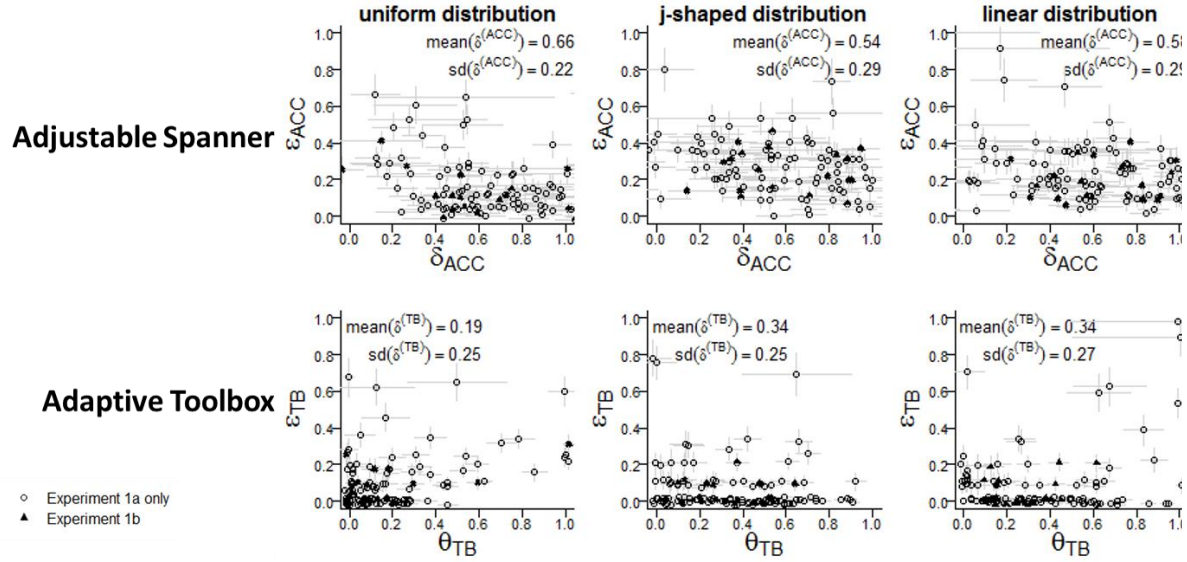

Figure 4. Illustration of the best-fitting parameters for each participant across both models and the three validity distributions (jitter added) in Experiment 1a. The mean and standard deviations of δ^{ACC} and θ^{TB} are indicated in the left upper corner of the respective plot. The error bars represent 95% confidence intervals around both best-fitting parameters. The parameter values of those participants who participated in Experiment 1b are shown as filled triangles.

As expected, the mixing parameter of the Toolbox, as well as the threshold of the Spanner differed between the three conditions: θ^{TB}, $F(2, 242) = 11.21$, $p<.001$, $BF_{10} = 707$; $\delta^{ACC,}$ $F(2, 242) = 26.05$, $p<.001$, $\log(BF_{10}) = 19$. Participants used most information in the condition with uniform distribution of validities and thus smallest dispersion of validities, as indicated by less frequent use of TTB (Toolbox) and higher thresholds (Adjustable Spanner) respectively, compared to the conditions with linear and j-shaped distributions of validities and thus more dispersion of validities. This finding further illustrates how the parameter values of θ^{TB} are mirroring the parameter values of δ^{ACC} around .5. Plotting the best-fitting parameters in Figure 4 shows, that while most best-fitting δ^{ACC} are bigger than .5 most best-

fitting θ^{TB} are smaller than .5, underlining that larger thresholds correspond to smaller mixing parameters and vice versa.

Simulation Study

Based on the estimated parameters for every individual in Experiment 1a, we ran a simulation study to identify situations in which the models made qualitatively different predictions. The simulation study had two goals: first, to identify regularities in discriminating trials, in which one model predicts one option, while the other model predicts the other option; and second, to identify those participants for which the models make most opposing predictions.

Method

For each subject tested in Experiment 1 we used the individually best-fitting parameter values θ^{TB} and δ^{ACC} and the corresponding error parameters ε^{TB} and ε^{ACC}, to simulate choice probabilities from the Toolbox and the Adjustable Spanner for all possible combinations of cue values in each of the three cue validity environments. We then searched for cue value combinations for which the Toolbox and the Adjustable Spanner made opposing predictions, as defined by one model's predicted probability of either movie being bigger than .55 and the other model's predicted probability for the same movie being smaller than .45.

Results and Discussion

Figure 5 plots the differences of the predicted choice probabilities of the Adjustable Spanner and the Toolbox. Comparing the horizontal dashed lines in the two panels of Figure 5 illustrates that both models make similar predictions in most of the trials. Only in a small subset, about 3% of trials, illustrated in the right panel of Figure 5 the models make different predictions.

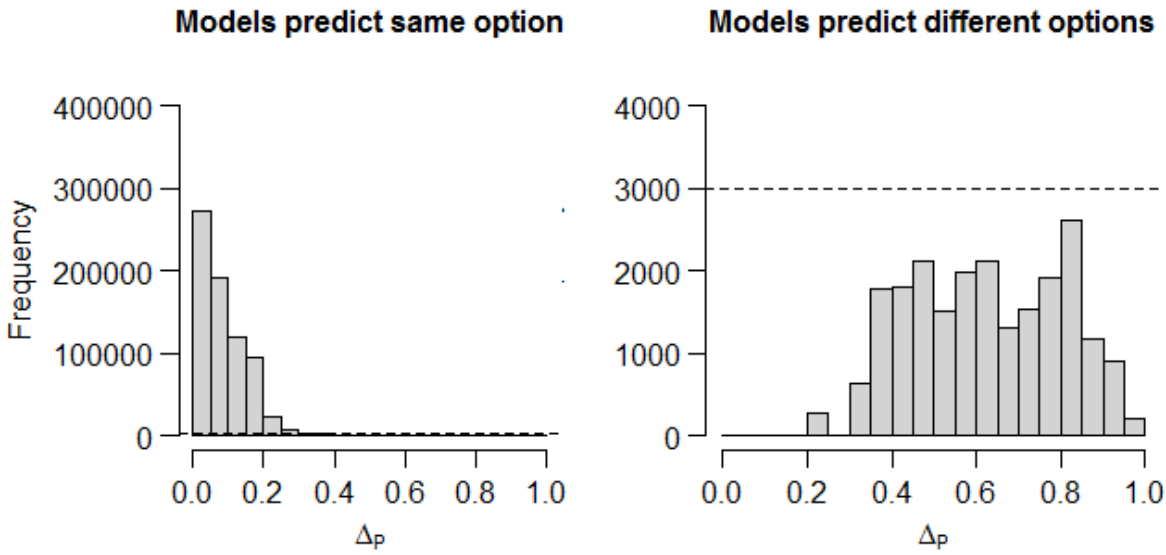

Figure 5. Histogram of the absolute differences of the predictions of choices probabilities of the two models, i.e. Δ_p = |P(A|ACC) - P(A|TB)|. The left panel illustrates the distribution of Δ_p if the models make the same prediction and the right panel illustrates the distributions if both models make different predictions, P(A|ACC) < .45 & P(A|TB) >.55 or P(A|ACC) > .55 & P(A|TB) >.45. Note that the scale on the y-axis differs between the two panels, as illustrated by the horizontal dashed line at 3000 (which sits very close to 0 on the Y-axis in the left panel due to the scale used).

Figure 5 further illustrates that the trials in which the predictions differed, defined by P(A|ACC) < .45 & P(A|TB) >.55 or P(A|ACC) > .55 & P(A|TB) >.45, were indeed more informative than the remaining trials on the model comparison. Because the average absolute difference of the models predictions Δ_p was much bigger in the discriminating (mean(Δ_p) = .56, sd(Δ_p)= .16) compared to those trials that did not discriminate between the models (mean(Δ_p) = .08, sd(Δ_p)=.07).

The trials that discriminated between the models were different from those that did not discriminate between the models. Trials that discriminated between the models were characterized by a higher proportion of critics making recommendations for only one of the two movies, thus yielding a higher proportion of discriminating cues (3.7 out of 6 vs. 3 out of

6 for the discriminating and non-discriminating trials, respectively). However, the average sum of the cue values was smaller (0.93 vs. 1.39) for the discriminating trials. Together, these characteristics result in the sequentially accumulated evidence switching more often between favoring one option or the other option in discriminating (1.63 switches on average) than the non-discriminating trials (0.67 switches). This difference in evidence-switching explains why those trials have a higher probability to discriminate the models, because only if the evidence switches twice between the two choice options, the Adjustable Spanner can in principal predict a choice that is not in line with either of the strategies in the Toolbox. Figure 6 illustrates this situation with an example of a discriminating trial. After the evidence of the first cue is accumulated, option A is favored over option B. Upon collecting more evidence (up to the fifth cue), option B is favored over option A. Finally, if the evidence of all cues is accumulated, option A is again favored over option B. Both TTB and WADD predict the same choice in this trial, while an accumulator model with moderate threshold settings (e.g. accumulating the information from five cues) would make the opposite choice.

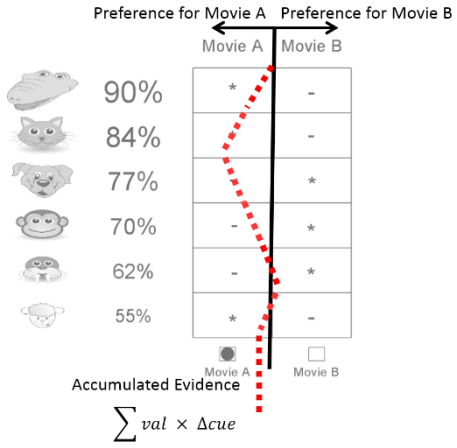

Figure 6. Illustration of accumulated evidence ($\sum val \times \Delta cue$ as red dashed line in a typical discriminating trial. The screenshot of the same trial is illustrated in Figure 3.

Alternatively, this difference can also be explained as follows: A toolbox with TTB and WADD represent two extremes of a continuum (using the least and most amount of

available information, respectively). In contrast, to this, a corresponding accumulator model with a given threshold can also collect an intermediate amount of information. If both, TTB and WADD make the same prediction while the accumulated evidence in-between indicates the opposite choice there must have been a switch of preference in-between. However, only if the individual threshold estimated in the Adjustable Spanner actually falls within this intermediate area. For example, if an individual integrates five cues in the environment of the trial illustrated in Figure 6, the Adjustable Spanner makes a different prediction than the Adaptive Toolbox in this specific trial. On the other hand, if a participant either relies consistently on the best cue, or always integrates all information in any given trial in this environment, both models make the same predictions on every possible trial. In principle, the Adjustable Spanner should provide a better fit to these trials than the Toolbox, because the Toolbox can only predict one of the choices in these trials, whereas the Adjustable Spanner can predict both choices. But, across all trials this does not lead to a superior fit of the Adjustable Spanner due to mainly two reasons. First, trials as depicted in Figure 6 are rare, and only specific configurations of the Adjustable Spanner lead to a choice prediction that is different from the prediction of the Toolbox. Second, these specific configurations of the Adjustable Spanner would lead to an inferior fit to the choices in the remaining trials.

We identified a set of participants from Experiment 1a for whom the models made different predictions across all three cue validity distribution conditions. The model parameters estimated for this subset of participants differed from the parameters of the other participants, showing a higher probability of using TTB (mean =.38, sd =.21, vs. mean =.27, sd =.27, t(107.82)=3.58, p<.001, BF_{10}=12.19), and likewise a smaller threshold (mean =.52, sd =.15, vs. mean =.61, sd =.29, t(175.68)=-3.58, p<.001, BF_{10}= 2.19). The parameter values

of these participants (who were subsequently invited to participate in Experiment 1b) are emphasized with triangles in Figure 4.

Experiment 1b: Generalization

Experiment 1b aimed to compare both models based on their ability to predict new data out-of-sample. For that purpose, individual trial lists were created, based on the previous simulation study. We intended to re-invite those participants from Experiment 1a for whom we could identify 15 trials in which the models make opposite predictions and present these participants only those discriminating trials. However, unfortunately, we invited participants based on an erroneous simulation that overestimated the number of discriminating trials. As a result, the actual percentage of discriminating trials was smaller and differed between the participants (range 33% to 100%; mean = 64%).

Method

Participants. Thirty-six of the participants from Experiment 1a were re-invited for Experiment 1b in exchange for a 20 Sfr (~ 20 USD) show-up fee. Out of these, 20 individuals agreed and eventually participated. Additionally, we confirmed that the model fit of the Toolbox and the Adjustable Spanner was about equal for this specific subset of participants, with the –LL of the Toolbox being smaller in 50% of the cases. Thus, the subset did not favor one model over the other a-priori.

Design and Material. The design and procedure of the experiment was similar to Experiment 1a except that this time, each participant saw an idiosyncratic set of 15 new pairs, plus 10 re-test trials from Experiment 1a in each of the three distribution conditions, thus 75 trials in total. Within each condition and participant, the order of the trials was randomized.

Results and Discussion

In order to replicate the findings of Experiment 1a, we first refitted the models to the complete datasets of Experiment 1b. The results confirmed the superior fit of the Toolbox. Furthermore, the trials that had been also previously presented in Experiment 1a, i.e. the old trials, indicated substantial (Landis & Koch, 1977) retest-reliability (mean $\kappa = .61$, sd = .1, between subjects). Participants chose the same option on 81% of the trials.

For the generalization test we investigated how well the models could predict the choices in the new trials, assuming that participants had the same parameters as estimated from Experiment 1a. Thus, we compared the simulated choices, based on the best-fitting parameters in Experiment 1a, with the observed choices in Experiment 1b. Because the models' parameters were estimated via minimizing the negative log-likelihood, we used similar methods for comparing the models' predictions (Elliott, Ghanem, & Kruger, 2014; Wulff & van den Bos, in press). We calculated –LL as well as the mean squared error (MSE) as an indicator of the predictive accuracy.

Table 1
Comparison of the accuracy of the models' predictions in the new trials in Experiment 1b

	Adjustable Spanner	Toolbox
-LL	1,173	730
MSE	.40	.25

Note. -LL = summed negative logarithmic likelihood; MSE = mean squared error. For both measures, smaller values indicate better fit.

As shown in Table 1, the Toolbox made more accurate predictions than the Adjustable Spanner, according to both –LL and MSE. The predictions from the Toolbox provided a smaller -LL than the Adjustable Spanner across all conditions combined but also separately in the uniform condition (236 vs. 445), the j-shaped condition (243 vs. 409), and

the linear condition (251 vs. 319). Across all conditions, the choices of 16 of the 20 participants were better predicted with the Toolbox than with the Adjustable Spanner. The superiority of the Toolbox is also apparent if only the –LL summed across those trials that discriminate the models is considered (474 vs. 877). In all previous comparisons of the models' predictions: $LR_{ACC/TB} < .001$.

To conclude, the Toolbox more accurately predicted the choices out of sample than the Adjustable Spanner. However, by design, we did not invite back all of the participants from Experiment 1a to participant in Experiment 1b and as mentioned above, the number of discriminating trials was smaller than intended. Therefore, we tested the robustness of our results by conducting another experiment where we invited all participants to participate in both sessions of the experiment.

Experiment 2

Method

The method of this second experiment was similar to Experiment 1. Thirty participants were recruited on the university's campus and social network groups. As all participants were invited for the second experimental session, we aimed at an average sample size of 20 participants for both sessions, similar to the sample size of the generalization session of Experiment 1. Participants received 20 Sfr (~ 20 USD) for completing both sessions of the experiment. All but two participants completed both sessions.

All participants completed the same choice set in the first session, and we estimated their individual model parameters. Following this, a second choice set including 15 new and 10 old trials was constructed based on a simulation study with the best-fitting parameters from the first session. We included as many discriminating trials as possible into each individuals' choice sets. If we could not identify 15 discriminating trials in a given condition,

we randomly chose the remaining trials from all possible new trials. Participants' choices in the second experimental session were used to compare the predictive accuracy of the models.

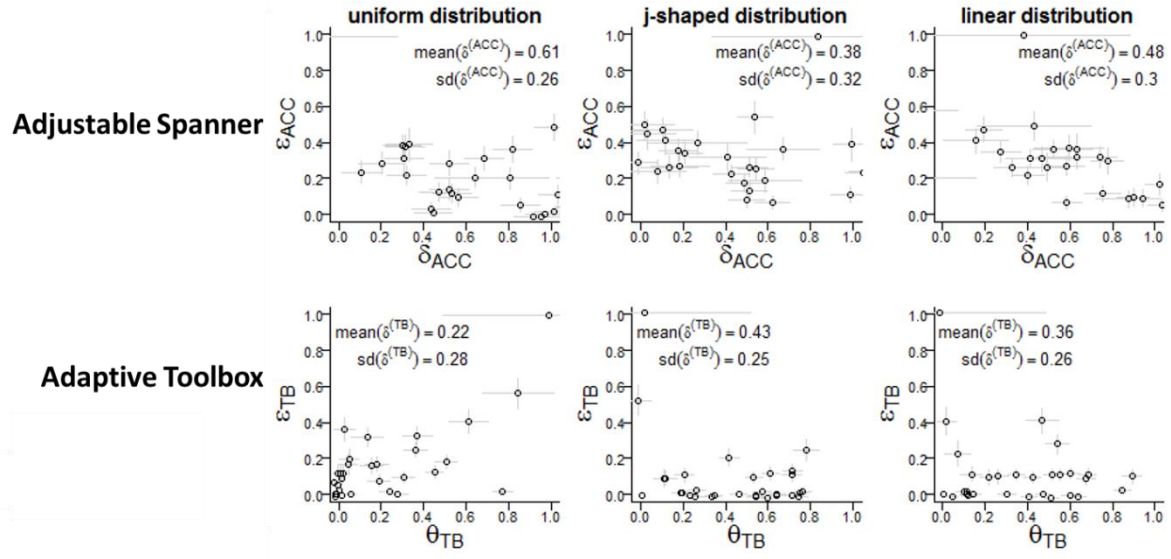

Figure 7. Illustration of the best-fitting parameters for each participant across both models and the three validity distributions (jitter added) in Experiment 2. The mean and standard deviations of δ^{ACC} and θ^{TB} are indicated in the left upper corner of the respective plot. The error bars represent 95% confidence intervals around the best-fitting parameters.

Results and Discussion

As an indicator of participants' choice consistency, we found that on the trials on which the compensatory and non-compensatory strategy made the same prediction, participants chose that dominant option on, on average 91% of the trials (*SD*= 16%).

We again observed that the Toolbox provided a superior fit to the observed data in the first session of the experiment. On average, the Toolbox fitted the data better, with a smaller –LL summed across participants and conditions ($-LL_{TB}$ = 1,226 and $-LL_{ACC}$ = 1,341, $LR_{ACC/TB}$ <.001). For the linear- and j-shaped environments, the Toolbox led to a smaller –LL for 76% and 67% of participants respectively. In the uniform condition, the Toolbox provided a better fit for only 40% of the participants.

Turning to the best-fitting parameters, participants preferred WADD over TTB within the toolbox (θ^{TB}: mean= .34, sd = .28). Likewise, the threshold of the Adjustable Spanner was estimated at intermediate values (δ^{ACC}: mean= .49, sd = .31). Figure 7 shows that parameters across participants were well distributed across the parameter space. Further analyses indicate that participants used most information in the condition with uniform distribution of validities (θ^{TB} = .22 and δ^{ACC} = .61), compared to the linear (θ^{TB} = .35 and δ^{ACC} = .48), and j-shaped condition (θ^{TB} = .43 and δ^{ACC} = .38). There was strong evidence for the difference in parameter values between the conditions, for θ^{TB}, $F(2, 58) = 8.11$, $p<.001$, $BF_{10} = 45$, and δ^{ACC}, $F(2, 58) = 7.76$, $p=.001$, $BF_{10} = 36$.

Similar to Experiment 1, in order to identify discriminating trials we again simulated choices for all possible combinations of choices values in the different conditions, with the best-fitting parameters from the first session of Experiment 2. The simulation revealed that the models predict opposite choices on only 3% of trials across all participants and possible combinations of cue values and validities (according to our criterion specified in the earlier section *Simulation Study*). Accordingly, the models made different predictions on 59% of the new trials used in the second session of this experiment.

Table 2
Comparison of the accuracy of the models' predictions in the new trials in second session of Experiment 2

	Adjustable Spanner	Toolbox
-LL	1,023	821
MSE	.26	.20

Note. -LL = summed negative logarithmic likelihood; MSE = mean squared error;

Results for the second experimental session indicate that participants' choices were quite reliable. In 83% of the old trials they made the same choice as in the first session,

leading to $\kappa = .65$. The results of the generalization test are summarized in Table 2. As can be seen in Table 2, the Toolbox made better out-of-sample predictions. Overall, the Toolbox more accurately predicted the data in terms of –LL (821 vs. 1022, $LR_{ACC/TB}<.001$) and MSE (.20 vs. .26) and also within the three conditions (all $LR_{ACC/TB}<.001$). Across all three conditions, the majority of participants (17 out of 28) were better described by the Toolbox than by the Adjustable Spanner. Focusing only those trials that discriminated the models, the Toolbox predicted the choices more accurately than the Adjustable Spanner, in terms of -LL (665 vs. 847, $LR_{ACC/TB}<.001$). To conclude, the superiority of the Toolbox in fitting and predicting the observed choice data was replicated in Experiment 2.

General Discussion

In order to model and predict behavior, it is important to understand how individual decision makers adapt to changes in the environment and how they trade-off accuracy against time and effort spent on a task. It is an ongoing discussion whether this process is better described by evidence accumulation models that assume an adaptive threshold or by Toolbox models consisting of multiple discrete strategies that are selected adaptively. Here, we focused on a comparison of both types of models on empirical grounds. For this purpose, empirically testable versions of the Adaptive Toolbox and the Adjustable Spanner were implemented and compared based on experimental choice data in multi-attribute choice tasks. The results indicated that, despite a great overlap of the models' predictions, the models could be distinguished based on choice data. Comparisons based on model fit and out-of-sample predictions indicated that in the task at hand, the Adaptive Toolbox described and predicted choices better than the Adjustable Spanner.

The present approach adds to earlier attempts comparing evidence accumulation models to Toolbox models based on discrete choices (e.g. Lee & Cummins, 2004; Newell et

al., 2007; Newell & Lee, 2011) as well as process data (e.g. Söllner et al. 2014; Söllner & Bröder, 2016). In an extension to these previous approaches, here we developed and tested a general mathematical implementation of the Adjustable Spanner (Newell, 2005) and rigorously tested it against a corresponding implementation of the Adaptive Toolbox (Scheibehenne et al., 2013) across multiple choice environments with different distributions of cue validities.

Computational Implementation of the Choice Models

Our implementation of the Adjustable Spanner represents the core theoretical ideas of accumulator models, by providing an individual threshold of evidence accumulation that is adaptive to the environment (Newell, 2005). The implementation extends earlier approaches (Hausmann & Läge, 2008; Newell & Lee, 2011), and facilitates a direct comparison of the Adjustable Spanner and the Toolbox. A recent study further supports the plausibility of the current implementation (Oh et al. 2016). The authors selected the best-fitting model in a multi-attribute choice task among models representing all possible combinations of cue use. They showed that individuals likely consider a certain number of cues ordered by their validity, e.g. the three or two best out of four cues, for making their choice. Though this formulation of a threshold emulates our current implementation it is only one of the possible ways to instantiate the theoretical idea of an adaptive threshold. Future research should systematically compare different versions of the spanner, for example versions in which stopping is determined by the sequential accumulation of individual cue values weighted by their validities, rather than the two-step ‘accumulate-then-weight’ process implemented here.

Likewise, the Toolbox used for the present comparison comprised only two distinct strategies, although larger and more complex Toolboxes have also been proposed (Gigerenzer & Todd, 1999). Using a Toolbox with only one non-compensatory and one compensatory

strategy is common practice (e.g. Rieskamp & Otto, 2006; Marewski & Schooler, 2011; Newell & Lee, 2011; Scheibehenne et al., 2013) and has also been used in earlier comparisons of Toolbox and accumulator models of choice (Lee & Newell, 2011). The principal advantage of this reduced-form toolbox is that it allowed us to test the core difference between the models while holding other factors constant. Specifically, we asked whether the intermediate steps of accumulation predicted by the Adjustable Spanner are observed or not. This approach necessarily sacrifices generalizability of our results to other larger toolboxes, spanners with different rules for sequential accumulation (see Footnote 2), and alternative ways to present information. But this trade-off of external validity against the internal validity required for a rigorous model comparison is unavoidable.

The Dissimilarity of the Models and the Superiority of the Toolbox

Only a small number of discriminating trials, about 3%, were identified in the simulations. However, the trials systematically differed from the non-discriminating trials. Most importantly we illustrated that whether the trial discriminates the models equally relates to features of the specific choice task as well as to the individual. Our two-stage design enabled the empirical identification and subsequent testing of specifically those trials. This adds to earlier approaches by identifying discriminating trials separately for every individual, rather than designing a generic set of discriminating trials for all participants (Lee & Cummins, 2004; Newell & Lee, 2011).

Nonetheless, across all trials and also only the discriminating trials, the Toolbox consistently provided a better fit than the Adjustable Spanner for on average 61.5 % of the participants in the conditions in the model fitting sessions, and 70% of the participants in the generalization settings. Although this indicates consistent superiority of the Toolbox over the

Adjustable Spanner not only on the average but also on the individual level, it also shows that several individuals were better described by the competing Adjustable Spanner.

Adaptation of the Models to Different Environments

The best-fitting model parameters varied systematically between the conditions. The individual threshold decreased and correspondingly the probability of using TTB increased as the validity of one cue was increasingly higher than the other cues. This indicates that individuals adapted to the environment, and that this adaption was successfully captured by the models. Comparing the models fit separately in the different environments showed that the Toolbox did not only perform better on average, e.g. due to worse fit in one condition and better fit in another, instead it fitted the data more accurately than the Adjustable Spanner in two out of three conditions in both experiments and for most of the participants.

Adaptive application of strategies to the environment has also been observed when participants receive feedback in line with one best performing strategy (Rieskamp, 2006; Rieskamp & Otto, 2006; Payne, Bettman, & Johnson, 1993) or simply on their predictive accuracy (Bröder & Schiffer, 2003a). In particular, Mata, Schooler and Rieskamp (2011) showed a similar adaptation of the use of strategies in environments with low versus high dispersion of cue validities, providing feedback after every trial. Going beyond these findings, our results show empirically that even without feedback, participants seem to have expectancies about the effectiveness of different strategies in specific environments such as the distribution of cues validities. Echoing this finding, Parpart, Jones and Love (2017) likewise showed that the performance of strategies depends on the structure of the environment. The authors implemented a Bayesian model with varying prior strength that allows to simulate decisions strategies varying in the amount of information that is considered for a given choice. The model compromises TTB and tallying on the one end of

the continuum and a linear regression model on the other extreme of the continuum. Across a large range of datasets, they show that the model with an intermediate prior strength that considers an intermediate amount of information was superior to more extreme versions in which either lots of information was ignored or all information was integrated. Our results dovetail with these finding as both the best-fitting Spanner and the Toolbox also settled at an intermediate degree of information integration.

How Did the Testing Environment Influence the Comparison of the Two Models?

Participants in both experiments saw all available information openly and freely on the computer screen. This may well have boosted the performance of the Toolbox. The assumption of an individual adaptive threshold made by the Adjustable Spanner is psychologically more plausible in situations where searching for information is costly (Newell, 2005). In order to see a larger number of choices corresponding to the current implementation of the Adjustable Spanner, participants would have to stop accumulating evidence at *intermediate stages* between the first discriminating and the last cue – a stopping point that adheres to neither TTB nor WADD. There is presumably a trade-off between the effort of monitoring the accumulated evidence at intermediate stages and the costs of uncovering and accumulating additional evidence. Because every monitoring step is costly, it is presumably less effortful to monitor only once, after all information is integrated. When all information is openly presented, as in the current experiments, such monitoring at the completion of integration is only marginally costlier than monitoring the accumulated evidence at intermediate stages. To the contrary, if the evidence at the intermediate stages is indecisive, additional evidence accumulation and monitoring is needed, making monitoring at intermediate stages even costlier than monitoring only at the completion of integration. Similarly, in paradigms involving effortful search from memory (e.g. Bröder & Schiffer,

2003), the Adjustable Spanner might be superior.

Furthermore, both models assume that after the threshold is crossed, or TTB is applied, respectively, the rest of the presented information is ignored. In contrast, to this assumption, Söllner and colleagues (2014) showed that individuals do not ignore additional information which is inconsistent with their current preference. If this is interpretation is correct then it has implications for participants' behavior in the discriminating trials in the second session of our experiments. In these trials the evidence after the first cue was most likely inconsistent with the preference indicated by the first cue. This kind of contradictory evidence is readily accounted for by the present implementation of the Toolbox due to the probabilistic allocation of the strategy by the mixing parameter. In contrast, the 'accumulate-then-weight' implementation of the Adjustable Spanner does not allow for a corresponding shift of the threshold, because a fixed decision threshold for a certain environment is assumed. In consequence, in the context at hand, the Toolbox seems to be the more flexible model and this may contribute to its superiority over the Adjustable Spanner in fitting the choice data.

In order to investigate the flexibility of the two models more closely we ran a parameter recovery study. Here we will only report the major findings of this study. The details are in the Supplementary Material to this article. We used both models in different configurations to predict choices in all possible trials of the multi-attribute choice task used in the empirical comparison of the models. Subsequently, we fitted both models to the predicted choices. This procedure supported the assumption that the Toolbox is the more flexible model. The Adjustable Spanner could not reproduce all configurations of the Toolbox, whereas the Toolbox reproduced all configurations of the Adjustable Spanner. This is due to the fact that every threshold δ^{ACC} manifests itself in a specific proportion of TTB and WADD

choices. Due to θ^{TB} indicating a probabilistic allocation of the strategies, every possible proportion of TTB and WADD choices can be predicted. However, vice versa not every proportion of TTB choices corresponds to a specific threshold.

This interpretation supports previous criticism of the Adaptive Toolbox emphasizing difficulties to falsify the model because of its flexibility (Glöckner & Betsch, 2011; Newell, 2005). The implementation we used addresses many of these concerns, especially with regard to the model's testability. Nonetheless, the model's mixing parameter allows the Toolbox considerable flexibility that arises from its functional form rather than the mere number of included strategies or free parameters. Future implementations of the Toolbox should attempt to predict strategy use on the trial level, based on trial and environmental features, in order to reduce the unexplained flexibility of the model.

What have we learned about the nature of adaptive decision making?

Decision making is adaptive. Individuals do not only adapt their decision making following feedback (Bröder & Schiffer, 2006; Lee et al., 2014; Rieskamp & Otto, 2006), but they also seem to hold expectations on the performance of certain strategies in specific environments. Beyond the adaptation on a larger scale, (i.e. different environments), the better fit of the current implementation of the Toolbox over the Adjustable Spanner further suggests sensitivity on a smaller scale, that is at the trial-specific level of particular cue-patterns. It appears that these subtle differences between trials can lead individuals to change their decision making behavior.

Following this line of reasoning, it is important for an in-depth understanding of human rationality not only to model the average decision making style across a large range of trials, but also to understand the driving forces of adaptation on a trial to trial basis. A probabilistic allocation as implemented here, can only serve as a proxy for modeling this

flexibility, given the lack of further knowledge on the small scale adaptation such a specific pattern of cue values.

Our approach to dealing with model mimicry

As noted earlier, the Adjustable Spanner and the Toolbox models show a large degree of mimicry. Here, we developed an approach to deal with the overlap in predictions made by the two model classes. Our method was to use an initial testing session to estimate model parameters, and then use them to develop a task environment to use in a second session that would allow for discrimination between the two models. The second session also provides a way of testing the ability of the models to predict new data, and thus avoid problems with over-fitting (Myung, 2000). However, the approach does come with limitations.

One issue with our method is that the particular experimental design developed for each individual will be based on the parameters estimated in a first session, and so our approach relies on these quantities being relatively stable. For example, in our experiment, it may be that participants set their response thresholds, or choose strategies, based on the particular set of cues and outcomes with which they are presented. Methods such as adaptive design optimization (Myung & Pitt, 2009) may be more robust in cases where the parameters change over time.

Another potential issue is that the way in which we design the choice sets makes it unclear how informative the data yielded will be for testing other theories. A number of alternative theories for multiple-cue judgment tasks exist (e.g. Busemeyer & Townsend, 1993; Glöckner, Hilbig & Jekel, 2014) and there is a long list of strategies that might be included in a more general Toolbox. For example, the Tallying strategy proposes that participants simply count up the number of cues in favor of each alternative, and choose the option with the most positive outcomes (Dawes, 1979). In an online supplementary, we show

in more detail that the Tallying model performs worse than both the Adjustable Spanner and the Toolbox. However, our experiment was not designed to distinguish between such models, and so while the result suggests that the Tallying model provides a relatively poor account of our data, it is unclear whether a quantitative assessment of the models' performance is compromised by the way that our choice sets were developed. In particular, testing a larger Toolbox that includes both Tallying and WADD as alternative strategies is not feasible in the context at hand because the way we constructed the choice options leads Tallying and WADD to make the same predictions in almost all trials.

Future Research and Limitations

Future studies could extend our findings in three ways. First, the models should be compared in environments with search costs imposed and/or information search in memory. Second, future studies could expand the existing implementation of both the Adaptive Toolbox as well as the Adjustable Spanner by integrating theoretically driven modeling of situational influences. Such an extension is especially important given that the assumptions that all remaining information is ignored after the threshold is crossed, or that one simple strategy is applied, do not seem to be valid in many cases (e.g. Söllner et al., 2014), whereby competing hypotheses on the influence of this additional evidence (Khader et al., 2013) could be tested empirically. Third, different implementations of the threshold within the Adjustable Spanner (which might capture the original conception of the model somewhat more directly) should be investigated.

Conclusion

Our results on the comparison of two particular versions of a Toolbox and a Spanner model, representing the core conflicting theories of adaptive decision making, indicate the

importance of situational and contextual factors on the flexible adaptation of decision processes, even without feedback and training.

References

Artinger, F., Petersen, M., Gigerenzer, G., & Weibler, J. (2015). Heuristics as adaptive decision strategies in management. *Journal of Organizational Behavior*, *36*(S1), S33–S52. doi: 10.1002/job.1950

Batchelder, W. H., & Alexander, G. E. (2013). Discrete-state models: Comment on Pazzaglia, Dube, and Rotello (2013). *Psychological Bulletin*, *139*(6), 1204–1212. doi: 10.1037/a0033894

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science (New York, N.Y.)*, *321*(5890), 851–854. doi: 10.1126/science.1158023

Bröder, A. (2000). Assessing the empirical validity of the "take-the-best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(Experiment 1), 1332–1346. doi: 10.1037/0278-7393.26.5.1332

Bröder, A. (2003). Decision making with the "Adaptive Toolbox": influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 611–625. doi: 10.1037/0278-7393.29.4.611

Bröder, A., & Newell, B. R. (2008). Challenging some common beliefs: Empirical work within the Adaptive Toolbox metaphor. *Judgment and Decision Making*, *3*(3), 205–214.

Bröder, A., & Schiffer, S. (2003a). Take The Best versus simultaneous feature matching: Probabilistic inferences from memory and effects of reprensentation format. *Journal of Experimental Psychology: General*, *132*(2), 277–293. doi: 10.1037/0096-3445.132.2.277

Bröder, A., & Schiffer, S. (2003b). Bayesian Strategy Assessment in Multi-attribute Decision Making. *Journal of Behavioral Decision Making*, *16*(3), 193–213. doi: 10.1002/bdm.442

Bröder, A., & Schiffer, S. (2006). Adaptive flexibility and maladaptive routines in selecting fast and frugal decision strategies. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32(4), 904–918. https://doi.org/10.1037/0278-7393.32.4.904

Bröder, A., & Schutz, J. (2009). Recognition ROCs are curvilinear-or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(3), 587–606. doi: 10.1037/a0015279

Busemeyer, J. R. (2017). Old and New Directions in Strategy Selection. *Journal of Behavioral Decision Making*. doi: 10.1002/bdm.2005

Busemeyer, J. R., & Rieskamp, J. (2014). Psychological research and theories on preferential choice. In S. Hess & A. Daly (Eds.), *Handbook of Choice Modelling: The State of the Art and the State of Practice* (pp. 49–72). Cheltenham, UK: Edward Elgar.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*(3), 432–459.

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes, 90*(1), 63–86. doi: 10.1016/S0749-5978(02)00508-3

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), S0140525X01003922. doi: 10.1017/S0140525X01003922

Dawes, R. M. (1979). Robust beauty of improper linear models in decision making. *American Psychologist*, 34, 571–582. doi: 10.1037/0003-066X .34.7.571

Donkin, C., Kary, A., Tahir, F., & Taylor, R. (2016). Resources masquerading as slots: Flexible allocation of visual working memory. *Cognitive Psychology*, 85, 30-42.

Elliott, G., Ghanem, D., & Krüger, F. (2016). Forecasting Conditional Probabilities of Binary Outcomes under Misspecification. *Review of Economics and Statistics*, *98*(4), 742–755. doi: 10.1162/REST_a_00564

Farrell, S., & Lewandowsky, S. (2010). Computational Models as Aids to Better Reasoning in Psychology. *Current Directions in Psychological Science*, *19*(5), 329–335. doi: 10.1177/0963721410386677

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: the Adaptive Toolbox. *Simple Heuristics That Make Us Smart*. doi: 10.1177/1354067X0171006

Glöckner, A. & Betsch, T. (2011). The Empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making,* 6(8), 712–722.

Hausmann, D., & Läge, D. (2008). Sequential evidence accumulation in decision making : The individual desired level of confidence can explain the extent of information acquisition. *Judgment and Decision Making*, *3*(3), 229–243.

Khader, P. H., Pachur, T., & Jost, K. (2013). Automatic activation of attribute knowledge in heuristic inference from memory. *Psychonomic Bulletin & Review*, *20*(2), 372–377. doi: 10.3758/s13423-012-0334-7

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics. **33** (1): 159–174. doi: 10.2307/2529310.*

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: unifying the "take the best" and the "rational" models. *Psychonomic Bulletin & Review*, *11*(2), 343–352. doi: 10.3758/BF03196581

Lee, M. D., & Newell, B. R. (2011). Using hierarchical Bayesian methods to examine the tools of decision-making. *Judgment and Decision Making*, *6*(8), 63–86. doi: 10.1016/j.cognition.2008.05.007

Lee, M.D., Newell, B.R., & Vandekerckhove, J. (2014). Modeling the adaptation of search termination in human decision making. *Decision, 4, 223-251*

Loomes, G., Moffatt, P. G., & Sugden, R. (2002). A microeconometric test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty*, *24*(2), 103–130. doi: 10.1023/A:1014094209265

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: an ecological model of strategy selection. *Psychological Review*, *118*(3), 393–437. doi: 10.1037/a0024143

Mata, R., Schooler, L. J., & Rieskamp, J. (2011). The Aging Decision Maker: Cognitive Aging and the Adaptive Selection of Decision Strategies. *Heuristics: The Foundations of Adaptive Behavior*, *22*(4), 796–810. doi: 10.1093/acprof:oso/9780199744282.003.0022

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*. 116, 499-518.

Newell, B. R. (2005). Re-visions of rationality? *Trends in Cognitive Sciences*, *9*(1), 11–15. doi: 10.1016/j.tics.2004.11.005

Newell, B. R., & Bröder, A. (2008). Cognitive processes, models and metaphors in decision research. *Judgment and Decision Making*, *3*(3), 195–204.

Newell, B. R., Collins, P., & Lee, M. D. (2007). Adjusting the spanner: Testing an evidence accumulation model of decision making. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, (2004), 533–538.

Newell, B. R., & Lee, M. D. (2009). Learning to adapt evidence thresholds in decision making. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 473–478. Retrieved from http://csjarchive.cogsci.rpi.edu/proceedings/2009/papers/85/paper85.pdf

Newell, B. R., & Lee, M. D. (2011). The right tool for the job? Comparing an evidence accumulation and a naive strategy selection model of decision making. *Journal of Behavioral Decision Making*, *24*(5), 456–481. doi: 10.1002/bdm.703

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 53–65. doi: 10.1037/0278-7393.29.1.53

Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best." *Organizational Behavior and Human Decision Processes*, *91*(1), 82–96. doi: 10.1016/S0749-5978(02)00525-3

Oh, H., Beck, J. M., Zhu, P., Sommer, M. A., Ferrari, S., & Egner, T. (2016). Satisficing in split-second decision making is characterized by strategic cue discounting. *Journal of Experimental Psychology: Learning Memory and Cognition*, *42*(12), 1937–1956. https://doi.org/10.1037/xlm0000284

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*(2), 207–240. doi: 10.1016/j.cogpsych.2012.03.003

Parpart, P., Jones, M., & Love, B. C. (2018). Strategies as Bayesian inference under extreme priors. *Cognitive Psychology*, *102*, 127–144. https://doi.org/10.1016/j.cogpsych.2017.11.006

Payne, J., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14*(3), 522–534.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge University Press.

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, *139*(6), 1173–203. doi: 10.1037/a0033044

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., &Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment,14*, 184–201.

Rieskamp, J. (2006). Perspectives of probabilistic inferences: Reinforcement learning and an adaptive network compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(6), 1355–1370. doi: 10.1037/0278-7393.32.6.1355

Rieskamp, J., & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell. *Simple Heuristics That Make Us Smart*, (October), 141–167.

Rieskamp, J., & Otto, P. E. (2006). SSL: A Theory of How People Learn to Select Strategies. *Heuristics: The Foundations of Adaptive Behavior*, *135*(2), 207–236. doi: 10.1093/acprof:oso/9780199744282.003.0011

Rouder, J. N., Morey, R. D., Nelson, C., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An Assessment of Fixed-Capacity Models of Visual Working Memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(16), 5975–5979. doi: 10.1073/pnas.0711295105

Scheibehenne, B., Miesler, L., & Todd, P. M. (2007). Fast and frugal food choices: Uncovering individual decision heuristics. *Appetite*, *49*(3), 578–589. doi: 10.1016/j.appet.2007.03.224

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing Adaptive Toolbox models: a Bayesian hierarchical approach. *Psychological Review*, *120*(1), 39–64. doi: 10.1037/a0030777

Scheibehenne, B., & von Helversen, B. (2015). Selecting decision strategies: The differential role of affect. *Cognition and Emotion*, *29*(1), 158–167. doi: 10.1080/02699931.2014.896318

Scheibehenne, B., von Helversen, B., & Rieskamp, J. (2015). Different Strategies for Evaluating Consumer Products: Attribute- and Exemplar-Based Approaches Compared. *Journal of Economic Psychology*, 46, 39–50. doi: 10.1016/j.joep.2014.11.006

Scheibehenne, B. Rieskamp, J. & González-Vallejo, C. (2009). Models of Preferential Choice: Comparing the Decision Field Theory with the Proportional Difference Model. *Cognitive Science*, 33, 911-939.

Simen, P., Cohen, J. D., & Holmes, P. (2006). Rapid decision threshold modulation by reward rate in a neural network. *Neural Networks*, *19*(8), 1013–1026.

Simon, H. A. (1956). Rational Choice and the Structure of the Environment. *Psychological Review*, *63*(2), 129–138.

Söllner, A., & Bröder, A. (2016). Toolbox or adjustable spanner? A critical comparison of two metaphors for adaptive decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 215–237. doi: 10.1037/xlm0000162

Söllner, A., Bröder, A., Glöckner, A., & Betsch, T. (2014). Single-process versus multiple-strategy models of decision making: Evidence from an information intrusion paradigm. *Acta Psychologica*, *146*(1), 84–96. doi: 10.1016/j.actpsy.2013.12.007

Suter, R. S., Pachur, T., & Hertwig, R. (2016). How Affect Shapes Risky Choice: Distorted Probability Weighting Versus Probability Neglect. *Journal of Behavioral Decision Making*, *29*(4), 437–449. doi : 10.1002/bdm.1888

van Ravenzwaaij, D., Moore, C. P., Lee, M. D., & Newell, B. R. (2014). A Hierarchical Bayesian Modeling Approach to Searching and Stopping in Multi-Attribute Judgment. *Cognitive Science*, *38*(7), 1384–1405. doi : 10.1111/cogs.12119

Venkatraman, V., Payne, J. W., & Huettel, S. A. (2014). An overall probability of winning heuristic for complex risky decisions: Choice and eye fixation evidence. *Organizational Behavior and Human Decision Processes*, *125*(2), 73–87. doi: 10.1016/j.obhdp.2014.06.003

Wulff, D. U., & van den Bos, W. (in press). Modeling Choices in Delay Discounting. *Psychological Science*.

Zhang, W., & Luck, S. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. doi: 10.1038/nature06860