

Do All Subjects Fit the Same Recognition Memory Model? Comparisons of Continuous, Discrete, and Hybrid Models Using Extended Multinomial Processing Trees

Anahí Gutkin¹, James F. Juola^{2, 3}, Manuel Suero⁴, and Juan Botella⁴

¹ Faculty of Medicine, Universidad Francisco de Vitoria

² Department of Psychology, University of Kansas

³ Department of Human-Technology Interaction, Eindhoven University of Technology

⁴ Faculty of Psychology, Universidad Autónoma de Madrid



In Gutkin et al. (2024), we studied traditional models of recognition memory, such as the two-high-threshold model and signal detection theory, using the benefits of multinomial processing tree models to quantify the effects of both ordinal and continuous variables. Simulations found that including relevant variables, such as response times (RTs) and confidence levels (CLs), significantly improved the estimation accuracy of model parameters and the power of model selection in fitting data from a recognition memory experiment. However, when applying these extended models, we found that no single model was suitable for all subjects. Thus, in the present study, we explored whether recognition memory is best represented as a continuous or discrete process by proposing a hybrid model that extends the original Atkinson–Juola (AJ) model (Atkinson & Juola, 1973, 1974; Juola et al., 1971). Data were sourced from a recognition memory experiment conducted by Juola et al. (2019), involving manipulations of relative target frequencies and measuring RTs and CLs. Our findings indicate that the extended AJ model provides a superior fit for most participants compared to purely discrete or continuous models. Additionally, the patterns of CLs and RTs align with the AJ model’s assumptions about their distributions. Despite these insights, there remains notable variability in cognitive processing and decision-making strategies across individuals. Therefore, there is still a need for a more general model, as suggested in the present article, that could accommodate the diverse strategies and individual differences influencing search, decision making, and response processes underlying recognition memory.

Keywords: recognition memory, signal detection theory vs. two-high-threshold theory, extended Atkinson and Juola model, response times and confidence levels, extended multinomial processing tree models

Supplemental materials: <https://doi.org/10.1037/xlm0001574.supp>

Multinomial processing tree (MPT) models are stochastic models for representing categorical data. They are useful for defining and measuring cognitive processes that underlie human behavior in various tasks and for testing the psychological assumptions on which these models are based. They can be used to assess the

validity of verbal cognitive theories that propose the existence of latent processes underlying cognitive states, decision making, and subsequent observed responses (Batchelder & Riefer, 1999; Erdfelder et al., 2009; Schmidt et al., 2025). One of the contexts in which MPTs have been successfully employed is in modeling the

This article was published Online First January 26, 2026.

Jeffrey Starns served as action editor.

Anahí Gutkin  <https://orcid.org/0000-0003-2693-0620>

Preliminary results of the present study were presented at the 64th meeting of the Psychonomic Society, November 16–19, 2023, in San Francisco, California, United States. This article has not been previously submitted to any journal. There are no conflicts of interest to disclose. This work was supported by a predoctoral contract “Formación de Personal Investigador” from the Universidad Autónoma de Madrid.

This work is licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International License (CC BY-NC-ND 4.0; <https://creativecommons.org/licenses/by-nc-nd/4.0/>). This license permits copying and redistributing the work in any medium or format for noncommercial use provided the original authors and source are credited and a link to the license is

included in attribution. No derivative works are permitted under this license.

Anahí Gutkin played a lead role in conceptualization, data curation, formal analysis, investigation, methodology, software, supervision, validation, visualization, writing—original draft, and writing—review and editing. James F. Juola played a lead role in writing—review and editing and a supporting role in conceptualization. Manuel Suero played a lead role in supervision and a supporting role in writing—review and editing. Juan Botella played a supporting role in writing—review and editing.

 The data are available at <https://osf.io/rbxw7/> and <https://osf.io/y78mk/>.

 The experimental materials are available at <https://osf.io/rbxw7/>.

Correspondence concerning this article should be addressed to Anahí Gutkin, Faculty of Medicine, Universidad Francisco de Vitoria, Ctra. Pozuelo-Majadahonda Km 1,800, Pozuelo de Alarcón, Madrid 28223, Spain. Email: anahi.gutkin@hotmail.com or anahi.gutkin@ufv.es

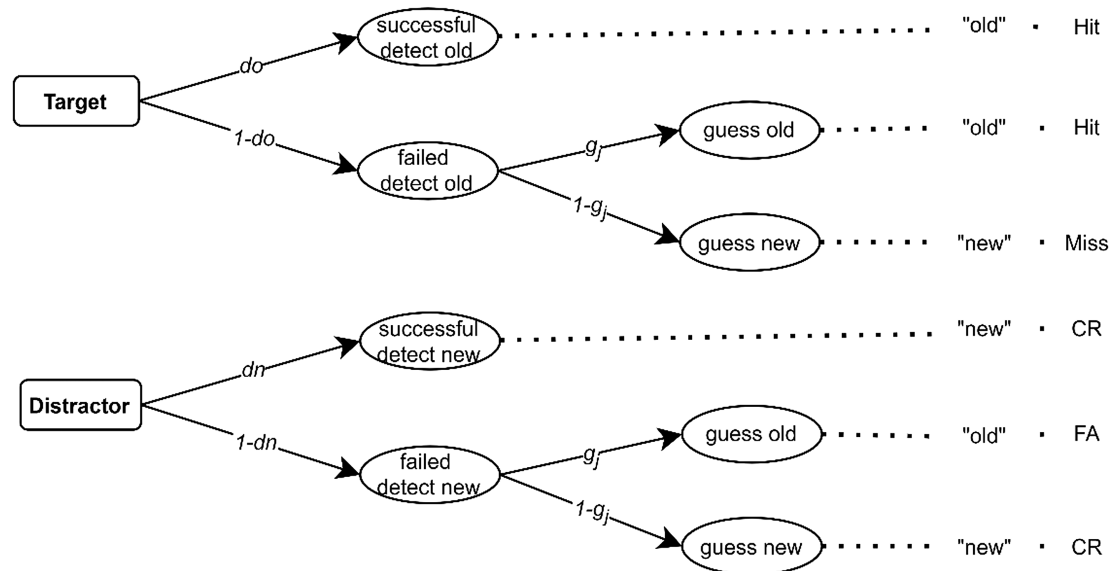
results from recognition memory experiments. In such research, participants typically study a list of words and then attempt to identify whether or not test words presented later had been previously studied. Words previously studied are to be categorized as “old,” whereas nonstudied words are to be called “new.” Depending on the type of word presented during the test phase and the participants’ responses, four response categories can be distinguished. Previously studied words (targets) can result in a hit (an “old” response to an old item) or a miss (a “new” response to an old item), while new words (distractors) can result in a false alarm (FA, an “old” response to a new item) or a correct rejection (CR, a “new” response to a new item).

MPT models define tree structures in which probability parameters are associated with unobserved discrete cognitive states or processes. We can then estimate the probabilities of entering certain latent states or executing latent processes through the frequencies of the observed response categories. One example of an MPT model commonly used to fit categorical data from recognition experiments is the two-high-threshold (2HT) model (Snodgrass & Corwin, 1988). As depicted in Figure 1, the 2HT model includes probabilities do and dn of successful old and new detection processes, respectively. Their complementary probabilities refer to detection failures, leading to guesses. Guess processes result in responding “old” with probability g and responding “new” with probability $1 - g$; for simplicity, we will refer to these processes as resulting from guess old and guess new states, respectively. Note that, in the 2HT model, cognitive states associated with detections (with do or dn probabilities) always lead to

a correct answer (hit or CR, respectively). In addition, it is possible for the subject to be correct even if he/she does not detect the stimulus. This would be possible if a guess old state occurs when a target is presented, or a guess new state occurs when a distractor is presented. According to the 2HT model, errors can only result from guessing new for targets (miss) and guessing old for distractors (FA). Experimental manipulations can influence each probability’s parameters. It is expected that when increasing the relative target frequency in a test block, the probability of guessing “old” should increase, meaning that g should increase, while do and dn should remain unchanged (e.g., Bröder & Schütz, 2009; Juola et al., 2019). Hence, when the relative target frequency is manipulated, as in Juola et al. (2019), g includes the subscript j (where $j: 1, \dots, J$), which refers to each condition of target frequency manipulation.

In contrast to models such as 2HT, which are based on discrete cognitive states, continuous memory strength models have been proposed, based on signal detection theory (SDT). According to SDT models, each stimulus elicits a latent value represented along a so-called memory strength or familiarity continuum. This familiarity value can vary for each presented stimulus and is often modeled as a random variable that follows a specific distribution depending on the stimulus presented. Unstudied items (distractors) have a baseline familiarity, but when items are studied (targets), their familiarity is assumed to increase, leading to two separate familiarity distributions. As depicted in Figure 2, in the context of recognition memory, these distributions are often assumed to follow normal (Gaussian)

Figure 1
2HT Model



Note. The 2HT multinomial processing tree model. The figure shows a separate tree for each type of stimulus. The boxes on the left indicate the presented stimuli (target or distractor). The ovals represent cognitive processes, respectively, being internal detection states and their subsequent response decisions. The probabilities of these processes are indicated along the arrows connecting them. In the first stage, there is a detection process, detect old or detect new (for the top and bottom tree, respectively), which can either succeed (i.e., do , dn , respectively) or fail (complementary probabilities). In cases of failed detection, there is a second phase in which a guess is made, with probability g_j for guess old and probability $1 - g_j$ for guess new. The subscript j allows the variation of the parameter g as a function of the relative target frequency manipulation. Depending on the presented stimulus (target or distractor) and the response given (“old”/“new”), we can distinguish four categories (hit, miss, CR, FA), which are shown in the right column. It is assumed that the probability of guessing does not vary depending on the stimulus presented; therefore, we find the same parameter g_j in both trees. 2HT = two-high-threshold; CR = correct rejection; FA = false alarm.

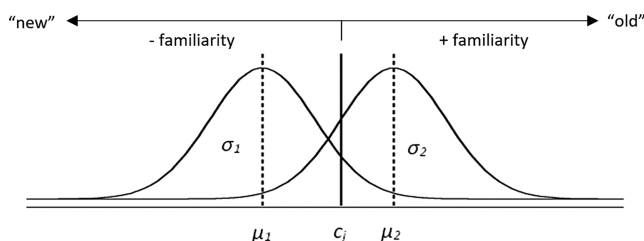
distributions (Macmillan & Creelman, 2004; Rotello, 2017), one for targets and the other for distractors. The model can be simplified by assuming that the mean of the familiarity distribution, μ_1 , for distractors is set to 0, and its standard deviation, σ_1 , is set to 1. Also, μ_2 and σ_2 are the mean and standard deviation¹ of the target distribution, which for simplicity will be denoted as μ and σ , respectively. By defining three parameters, μ , σ , and c , this model allows predictions of the four types of categorical responses. The parameter c represents an internal decision criterion that enables participants to divide the familiarity continuum into two regions: an “old” response region above c and a “new” response region below c . Depending on experimental conditions, these parameters can vary. For instance, when the proportion of target stimuli in a test block increases, c is expected to decrease, while μ and σ should remain constant (Juola et al., 2019; Swets et al., 2000; Wickens, 2001). Therefore, in this case, such as in Juola et al.’s (2019) data, the criterion can vary, and therefore it should be appended by the subscript j , such as in Figure 2.

It should be noted that although SDT models are not inherently MPT models, they can be translated into multinomial tree models while retaining the same parameters (see Singmann & Kellen, 2013, for reparameterization details). That is, we can express the SDT model, which assumes a continuous variable (familiarity), as a multinomial tree model by reparameterizing the probabilities associated with the nodes. Figure 3 shows the reparameterization of the probabilities of each category of the multinomial tree as a function of the parameters μ , σ , and c of the SDT model.

This reparameterization of SDT into a multinomial tree model is particularly useful as it enables us to include continuous variables (such as familiarity) in an MPT.

The SDT and the 2HT model are originally meant to be used for categorical variables. However, it is common in recognition experiments to measure categorical (e.g., hit/miss/FA/CR), ordinal (e.g., confidence ratings: high, medium, or low confidence), and continuous (response times [RTs]) variables. We now have statistical models that allow us to study these three types of dependent variables (DVs) simultaneously. One of the ways to do this is through

Figure 2
SDT Model



Note. The SDT model. The probability areas to the left of c_j , labeled as –familiarity, determine the probabilities of “new” responses, whereas to the right of c_j , the +familiarity area determines the probabilities of “old” responses. The +familiarity area for the distractor (left) and target (right) distributions represent false alarm and hit probabilities, respectively. The –familiarity areas in the target and distractor distributions represent the respective probabilities of a miss and a correct rejection. This figure shows homoscedastic distributions, but heteroscedastic distributions are allowed in the model. SDT = signal detection theory.

MPT for discrete and continuous variables, known as MPT-DC (Heck & Erdfelder, 2016; Heck et al., 2018; Klauer & Kellen, 2018).

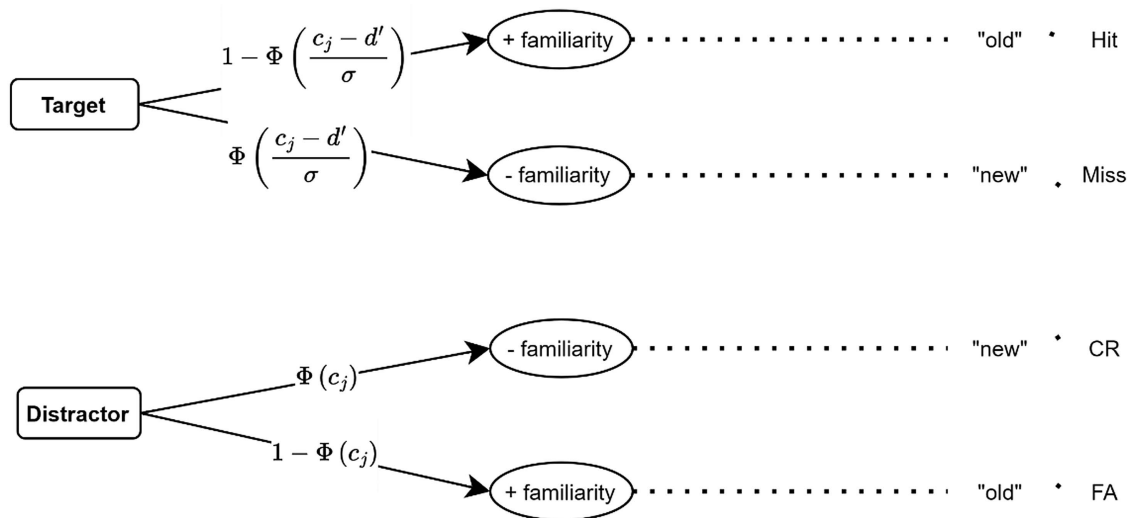
In previous work, it has been found that including relevant continuous variables as DVs has proven advantageous in reducing the standard estimation error and favoring model selection (Heck & Erdfelder, 2020; Heck et al., 2018; Laukenmann et al., 2023; Starns, 2021), and more importantly, it allows us to address new research questions. Studies that have included RTs in recognition memory experiments have found an inverse relationship between confidence and RT (e.g., Weidemann & Kahana, 2016), identified violations of the predictions that can only be detected if we take into account the interactions among accuracy, confidence, and RTs (Gutkin et al., 2024), or detected speed–accuracy effects as evidence of a misleading memory mechanism (Starns & Ma, 2018; Voormann et al., 2025). Simulations have shown that when comparing 2HT and SDT models using MPT-DC while including accuracy measures, confidence levels (CLs), and RTs simultaneously, we can be considerably better at recovering the data-generating model than when using only categorical variables or ignoring relevant DVs (Gutkin et al., 2024). This latter finding can be useful for addressing the ongoing debate between the relative validities of discrete and continuous recognition memory models.

Attempts to resolve debates between proponents of discrete and continuous models have been limited by persistent challenges in parameter estimation and model selection. Simulations of these models, as reported by Gutkin et al. (2024), shed some light on these issues and suggest that failures to fit the data may arise from the influence of unexamined factors, such as not jointly fitting CLs and RTs, which can distort the results. However, when fitting real data, the study could not provide a definitive conclusion due to substantial variability across subjects. It is also possible that individual differences hinder a deterministic resolution to the discrete/continuous dilemma. In other words, participants may employ different response strategies, or other unexamined factors may influence their engagement with the task, causing the data to align more closely with one model or the other. Consequently, no single model might be able to account for the data from all participants. A possible solution to these problems could be the development of a hybrid model, one that is more comprehensive in order to encompass both the SDT and 2HT frameworks, leading to better fits to the data. Whereas some types of hybrid models have been proposed, evidence has been presented against them (Rotello & Zeng, 2008; Wixted & Mickes, 2010). Here, we propose an extended hybrid model based on that developed by Atkinson and Juola (Atkinson & Juola, 1973, 1974; Juola et al., 1971), known as the Atkinson–Juola (AJ) model. It was derived from classical SDT but includes elements of continuous memory strength and discrete memory states. In Figure 4, we present a schematic of the original AJ model.

The AJ model, represented in Figure 4, describes how subjects decide whether a stimulus is a target or a distractor in a recognition memory task. The model presumes that subjects assign an internal familiarity value for each stimulus that represents the test item’s memory strength. As can be seen in Figure 4, the memory strengths for old and new stimuli are assumed to follow two normal distributions, respectively. The familiarity of the stimulus is evaluated against two internal criteria: a low criterion (c_L) and a high criterion

¹ When homoscedasticity is assumed, then $\sigma = 1$, and only $\mu = \delta$ and c will be estimated, where δ is the sensitivity parameter.

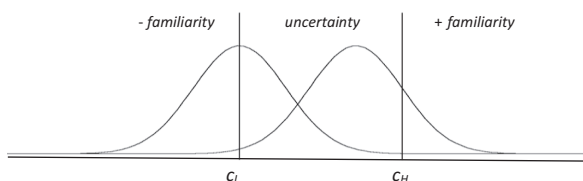
Figure 3
Reparametrized SDT Model



Note. The SDT model is a multinomial tree. The figure shows the probability areas of +familiarity (probability of saying “old”) and –familiarity (probability of saying “new”) for the respective targets and distractor distributions. Depending on the presented stimulus and the response given (“new”/“old”), we can distinguish four categories (hit, miss, CR, FA), which are shown in the right column. The Φ symbol represents a normal probability function with mean 0 and standard deviation 1. SDT = signal detection theory; CR = correct rejection; FA = false alarm.

(c_H). Depending on the results of those comparisons, subjects enter one of three possible states: –familiarity, uncertainty, and +familiarity. In the –familiarity state, subjects reject the stimulus as a target because its familiarity value is below c_L . In the +familiarity state, subjects identify the stimulus as a target because its familiarity is greater than c_H . If, on the other hand, the elicited familiarity falls within the uncertainty area, subjects are unsure whether the stimulus is a target or not and therefore engage in a subsequent long-term memory search to find further evidence. Because in this case there is a second search process, responses associated with the uncertainty area should be slower than those derived from extreme familiarity responses (– or +familiarity areas). The original AJ model was based on a situation in which the studied list was well-memorized. Therefore, the search was assumed to be successful for target items,

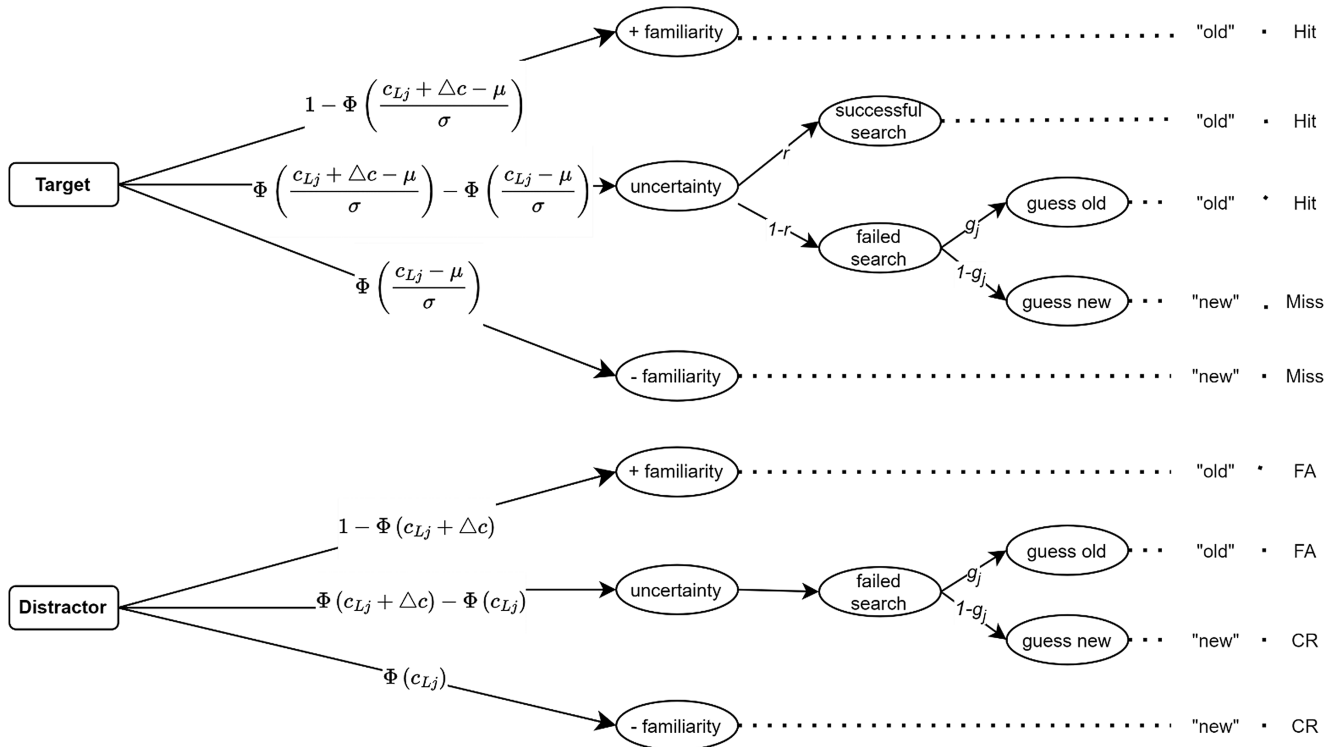
Figure 4
AJ Model



Note. Original AJ model. The areas to the left of c_L (low criterion) in the distractor (left) distribution and the target (right) distribution represent the –familiarity probability area, where a “new” response is given. The areas to the right of c_H (high criterion) in the target and distractor distributions represent the +familiarity area, where an “old” response is given. Between the c_L and the c_H criteria, there is the uncertainty probability area for the target and distractor distributions. AJ = Atkinson–Juola.

leading to an “old” response, and an unsuccessful search led directly to a “new” response. There was no guess parameter, since when the familiarity value falls into the uncertainty zone, it could only trigger a search determined to yield a correct response, with high confidence and of long duration. In a more typical situation, in which the studied materials are not assumed to be fully memorized, if the search fails, the subject may still make a correct guess, which would probably be of relatively low confidence. In the original AJ model, these different ways of arriving at the same answer within the state of uncertainty are not modeled by specific parameters. Therefore, we need to model the different sources of correct and incorrect answers within the area of uncertainty and the CL and RT distributions associated with each of them. To this end, in the present article, we propose an extension of the AJ model using an MPT-DC. Next, in Figure 5, we show the reparameterization of the AJ model as a multinomial model and its extension to the uncertainty area.

As mentioned previously, participants can enter into three cognitive states: –familiarity, +familiarity, and uncertainty. The extended AJ model is a hybrid model. Specifically, in the extreme ($\pm$) familiarity areas, it is a continuous model (such as the SDT model), whereas in the uncertainty areas, it functions as a threshold model that includes several discrete processes. The probability of entering each state (see Figure 5) depends on the type of stimulus presented (target or distractor), the target mean parameter μ , and the low and high decision criteria (c_L and c_H). It is assumed that distractor stimuli familiarities follow a normal probability density function with a mean of 0 and a standard deviation of 1, while target familiarities follow a normal probability density function with a mean of μ and a standard deviation of σ . Note that Φ represents the cumulative distribution function of the standard normal probability density function. In this manner, for example, when a distractor is presented, the probability of entering

Figure 5*Hybrid AJ Model Reparametrized as a Multinomial Model*

Note. Reparametrized and extended AJ model. The model includes the +familiarity, -familiarity, and uncertainty probability areas of the target and distractor distributions. The uncertainty area of the target distribution includes a search process for information stored in long-term memory, which may or may not be successful. When the search process for the target is unsuccessful, a guess is produced. The uncertainty area of the distractor distribution necessarily leads to a guess since it cannot be successful. Depending on the stimulus presented and the response given ("new"/"old"), we can distinguish four categories (hit, miss, CR, FA), which are shown in the right column. AJ = Atkinson–Juola; CR = correct rejection; FA = false alarm.

the -familiarity area (cumulative probability below c_L) would be represented by $\Phi(c_L)$, and the probability of entering the +familiarity area would be represented by $1 - \Phi(c_H)$. The model establishes that the decision criterion c_H is greater than c_L . To satisfy that $c_H > c_L$, we assume that $c_H = c_L + \Delta c$, where $\Delta c \in \mathbb{R}^+$. Thus, we can express the probability of entering the +familiarity area given a distractor as $1 - \Phi(c_H) = 1 - \Phi(c_L + \Delta c)$. For targets, the probabilities for +familiarity, -familiarity, and uncertainty are $1 - \Phi((c_L + \Delta c - \mu)/\sigma)$, $\Phi((c_L - \mu)/\sigma)$, and $\Phi((c_L + \Delta c - \mu)/\sigma) - \Phi((c_L - \mu)/\sigma)$, respectively. If the search yields evidence suggesting that the item is a target, which happens with probability r , a positive response should be made; otherwise, with probability $1 - r$, a guess should result. This guessing process can produce an "old" response, which happens with a probability of g , or a "new" response, which happens with a probability of $1 - g$. On the other hand, when a distractor is presented, the subject cannot have a successful search for the target. Therefore, one can only guess "old" with a probability of g or guess "new" with a probability of $1 - g$. The model does not assume a difference in the g parameter for targets and distractors, given that each results from a failed search when an ambiguous familiarity value is elicited.

In the present study, we will use the data from Juola et al. (2019) to address the question of which recognition memory model best fits the data from each individual and whether there can be a single

unique model for all subjects or if different models for different subjects can coexist. To do so, we will compare pure continuous or discrete models, such as SDT and 2HT, respectively, along with a specific hybrid model, our proposed extension of the AJ model. We will include accuracy, CLs, and RTs as simultaneous DVs by means of an MPT-DC, because doing so has proved to be beneficial in terms of model selection and improvement of parameter estimation precision (Gutkin et al., 2024).

Method

Data

Data for the study were collected from <https://osf.io/y78mk/>, and detailed information about the participants, materials, and procedures is available in Juola et al. (2019). The data set includes responses from 47 Universidad Autónoma de Madrid students and academic community members.

Materials consisted of 500 common Spanish nouns, presented one at a time on a computer monitor using E-Prime 3 software. Participants first studied 250 words, pronouncing them as they appeared. Afterward, they engaged in a short intervention task followed by five test blocks of 100 words each, with varying but blocked target proportions. In the test phase, participants were

informed of the relative percentage of targets in each block of 100 trials (being 15%, 35%, 50%, 65%, or 85%). Only conditions with target percentages of 35%, 50%, and 65% were analyzed in the present report because conditions with 15% and 85% led to some cells with relative frequencies equal to or close to 0 and, consequently, resulted in worse goodness-of-fit estimates of the models (Gutkin et al., 2024). Participants were encouraged to respond rapidly with “old” or “new” decisions, and RTs associated with the test phase were measured. Subsequent to their responses, participants were instructed to rate their confidence on a 3-point scale (“certain,” “relatively certain,” and “uncertain”). All “certain” responses were labeled as high-confidence responses, whereas low-confidence responses, due to the relative lack of “uncertain” responses, were a combination of “relatively certain” and “uncertain” responses.

Note the experiment generated three types of data: categorical (hit/miss/FA/CR), ordinal (CL; high/low), and continuous (RT) in all three conditions of relative target proportions.

Model Details

We applied the MPT-DC approach proposed by Heck and Erdfelder (2016) to extend the AJ model to simultaneously account for accuracy, CLs, and RTs. To achieve this, the continuous RT variable was discretized into fast and slow RT bins, and then the model was reparametrized, creating what is known as an MPT-extended model. In the extended version, the branches of the original MPT model were subdivided into the same number of RT bins, using as a bin boundary the log-normal approximation recommended in previous simulation studies (see log-normal approximation procedure and simulations in Heck & Erdfelder, 2016). The procedure consisted of log-transforming each subject's RTs, calculating the mean and variance, obtaining the quantiles, and then converting them back to the standard RT scale to obtain the necessary cut-offs for categorization.

Below, we define the parameters estimated by each model. The representation of the extended models can be found in Appendix A.

The extended 2HT model has 17 free parameters (see Figure A1). There are five parameters to assess discrete state probabilities: do , dn , and three g_j parameters, one for each relative frequency target condition. There are four L_a parameters to model RT bins: one for the probability of each cognitive state resulting in a fast response. Slow response probabilities are complementary (i.e., $1 - L_a$). The subscript “ a ” refers to the cognitive state, $a = \{do: \text{detect old}, dn: \text{detect new}, go: \text{guess old}, gn: \text{guess new}\}$. There are eight parameters, $H_{a,b}$, for the probability of high CLs: subscript “ a ” refers to each of the four mentioned cognitive states, and subscript “ b ” refers to the associated RT bins, $b = \{f: \text{fast}, s: \text{slow}\}$. There is a distinct high-confidence probability parameter for each state and RT-bin combination in the model. Again, the high- and low-confidence probabilities are complementary (i.e., $H_{a,b}$ and $1 - H_{a,b}$).

Regarding the parametric constraints² of the 2HT model, it has been established a priori that the parameter g_j increases with the manipulation of target frequency, such that $g_1 < g_2 < g_3$, where subscripts 1, 2, and 3 represent target percentages of 35%, 50%, and 65%, respectively. Also, because RTs are expected to be shorter for successful detections than for guesses, we established that $L_{do} > L_{go}$ and $L_{dn} > L_{gn}$. As for CLs, the higher confidence responses

result from detections, followed by those from guessing. Therefore, $H_{do,b} > H_{go,b}$ and $H_{dn,b} > H_{gn,b}$.

The extended SDT model has 11 free parameters (see Figure A2). Six are dedicated to modeling categorical data: μ , σ , and three parameters (c_j), one for each relative frequency target condition. Additionally, there are four parameters (L_a) for RT bins, each representing the probability of each category resulting in a fast response. The subscript “ a ” refers to each category or branch of the tree model, $a = \{hit: \text{hit}, miss: \text{miss}, fa: \text{FA}, rc: \text{CR}\}$. Since CLs in the context of SDT are linearly defined on the familiarity continuum, two increment criteria, referred to as c_2 and c_3 , have been established to estimate the distances between the different criteria for response confidence ratings.

For the SDT model, it is expected that the criterion c_j decreases with the increase in relative target frequency. Therefore, parametric constraints were established as follows: $c_{L1} > c_{L2} > c_{L3}$. Regarding RTs, it was established that responses for CRs and hits are expected to be faster than those for FAs and misses. This expectation is based on the inverse U-shaped relationship between latencies and decision-confidence responses (i.e., “new-high,” “new-low,” “old-low,” “old-high”), as previously analyzed in the data from Juola et al. (2019) and reported in Gutkin et al. (2024). Thus, for the familiarity distribution of targets, we have established the following parametric restrictions: $L_{hit} > L_{miss}$ and for distractors $L_{cr} > L_{fa}$.

Three versions of the extended AJ model (see Figure A3) were fitted to the data. Two of them had a total of 18 free parameters. In the first version, referred to as AJ_{cj} , it is assumed that the manipulation of the relative target frequency condition shifts both criteria simultaneously, whereas in the second version, referred to as AJ_{gj} , it is assumed that relative target frequency affects the probability of the guessing process. For the AJ_{cj} model, there were eight parameters to model categorical variables: μ , Δc , σ , three parameters c_{Lj} (one for each relative frequency target condition), and the state probabilities r and g (for searching and guessing old processes, respectively). Similarly, the AJ_{gj} model also consisted of eight categorical parameters: μ , Δc , σ , one parameter c_L , three guessing state probabilities g_j , and one successful search probability parameter r . The third version, named the AJ model, has a total of 20 parameters and allows both the criterion and guessing probabilities to vary depending on the target proportion conditions. It therefore includes the following 10 parameters to model categorical variables: μ , Δc , σ , three c_{Lj} parameters, three state probabilities g_j , and a search probability parameter r . In all AJ models, it was assumed that Δc does not vary between relative frequency target conditions, which ensures that the distance between c_L and c_H remains constant. Thus, Δc does not carry the subscript j . The other parameters characterizing RTs and CLs are identical across all model versions. For simplicity and identifiability, it was assumed that the probability of a high-confidence response is equal for positive and negative familiarity, which is reasonable given the inverse U-shaped relationship between familiarity and CLs (Baranski & Petrusic, 1994, 1998; Murdock & Dufty, 1972; Ratcliff & Starns, 2009;

² The ordinal constraints on the probability parameters have been implemented using shrinkage parameters, which entail a restructuring of the models. We have explained their use in Appendix B so that readers who wish to use the model files in the repository can understand them. However, because their use is not necessary for presenting the main results, in the main text we will use their simplified version, as explained in the Model Details section.

Starns, 2021; Weidemann & Kahana, 2016). Hence, there were only three L_a parameters for modeling fast RT bins, one for each level of $a = \{f: \text{familiarity}, s: \text{search}, g: \text{guess}\}$. There were six $H_{a,b}$ parameters for assessing high-confidence probabilities, where again, the a subindex refers to each state, and b refers to each RT bin, $b = \{f: \text{fast}, s: \text{slow}\}$.

For the AJ model, we assumed that g_j increases with relative target frequency (i.e., $g_1 < g_2 < g_3$) while the low criterion c_{Lj} decreases (i.e., $c_{L1} > c_{L2} > c_{L3}$). Note that $c_{Hj} = c_{Lj} + \Delta c$, so c_H decreases in parallel with c_{Lj} . In the constrained versions, AJ_{cj} and AJ_{gj} , either g or c , respectively, are fixed. In all versions of the AJ model, RTs associated with familiarity states are expected to be shorter than those from successful searches or guessing, as these require an additional process. Thus, the estimated probabilities of fast responses follow: $L_f > L_s$, $L_f > L_{go}$, and $L_f > L_{gn}$. As for confidence parameters, the highest confidence responses are expected from states following a successful search, followed by those with extreme familiarity, and finally those from guessing, as the subject has no sense of recognition or certainty about the answer in the latter case (Kang et al., 2025). Therefore, we established that $H_{s,b} > H_{f,b} > H_{go,b}$ and $H_{s,b} > H_{f,b} > H_{gn,b}$.

All models can be found on the online repository at <https://osf.io/rbxw7/>.

Model Fits and Analysis

The models have been fitted using the *MPTinR* package (Singmann & Kellen, 2013). The choice of this library is due to the fact that we are using models, such as SDT and AJ, that include parameters that are not probabilities, such as the parameter d' whose range is $[-\infty, +\infty]$. To our knowledge, *MPTinR* is the only package that allows fitting reparametrized multinomial models that are not pure MPTs and therefore does not restrict the range of parameters to $[0, 1]$. To account for individual differences, fits were performed individually for each participant. Nevertheless, individual, aggregated, and mean estimated parameters can be found in the online repository <https://osf.io/rbxw7/>.

The standard errors of the estimated parameters were obtained directly from the *MPTinR* package (Singmann & Kellen, 2013) and are based on the Hessian matrix. The quantification of model misfit is also automatically obtained by *MPTinR* through the estimation of G^2 statistics for the goodness-of-fit test. The results of individual goodness-of-fit tests³ and the justification for the decision to eliminate data from conditions with extreme target percentages (15% and 85%) are presented in Appendix B. Comparisons among the 2HT, SDT, and AJ models were conducted using the Akaike information criterion (AIC). We also performed Monte Carlo simulations to evaluate the performance of different model selection criteria. As the best model recovery was achieved with AIC, we chose to report only the results using AIC to shorten the main text and focus on its primary objectives. However, simulations using other criteria can be found in Appendix C.

To validate the established psychological theories proposed by Atkinson and Juola (1973, 1974) regarding the AJ model (AJ_{cj} and AJ_{gj}), we tested the following predictions: (a) the influence of manipulating the relative target frequencies with respect to the criteria for the AJ_{cj} model and with respect to the probability of guessing “old” in the AJ_{gj} model; (b) predictions regarding RTs

($L_f > L_s$, $L_f > L_{go}$, and $L_f > L_{gn}$); and (c) predictions regarding CLs ($H_{s,b} > H_{f,b} > H_{go,b}$ and $H_{s,b} > H_{f,b} > H_{gn,b}$).

For testing these hypotheses, we used the log-likelihood ratio test for comparing the model incorporating ordinal parametric relationships (full model), such as $L_f > L_s$, against a null model that assumes equality constraints (restricted model),⁴ such as $L_f = L_s$. The test statistics of the log-likelihood ratio test, G^2 , follow a chi-square distribution with degrees of freedom equal to the difference in degrees of freedom between the more complex model and the nested model (formulas and detailed logic can be found in Agresti, 2012). A p value of less than the nominal significance level of $\alpha = .05$ indicates a significant difference in the model fit, rejecting the null hypothesis and, therefore, accepting that the studied effect is significant.

Transparency and Openness

To ensure replicability and maintain transparency, this study fully discloses any exclusion of data, all experimental manipulations, and every measure used. All additional online materials (data sets, codes, and research materials) are available at <https://osf.io/rbxw7/>. The codes for fitting the experimental data and simulations were developed in R, Version 4.3.2 (R Core Team, 2020). Model fitting was performed using the *MPTinR* package (Singmann & Kellen, 2013), while visualizations were created with the *ggplot2* package, Version 3.5.1 (Wickham, 2016). For the evaluation of model comparison criteria included in the analyses in Appendix C, the *NLMulti* package in R (Kellen & Klauer, 2020) was used. This package processes files containing models written in C++, which are imported into R as packages to optimize computation time. The present study was not preregistered.

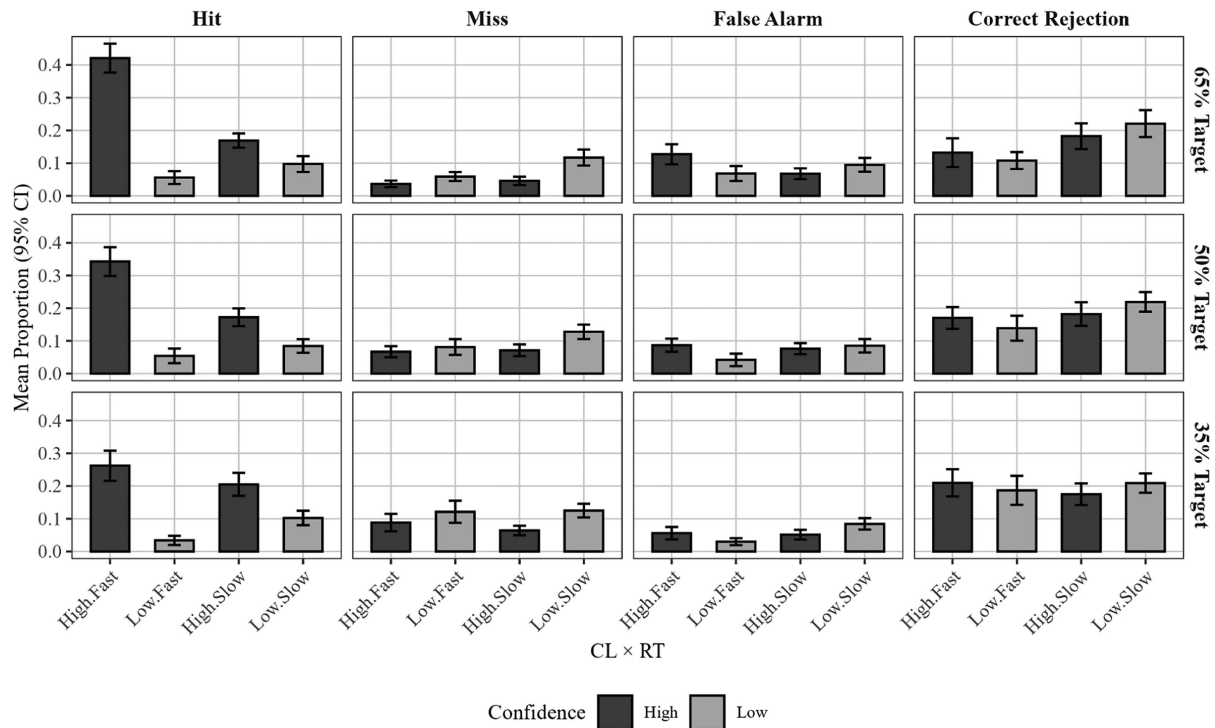
Results

Figure 6 shows the mean proportions of responses for each combination of response category and CL $\times$ RT bins across the relative target frequency manipulation conditions. As expected, increasing target frequency resulted in a marked increase in “old” responses and a corresponding decrease in “new” ones. Hits were mostly high-confidence and fast responses. Nevertheless, a substantial proportion of high-confidence hits were slower responses. Overall, the data align with the prediction that target prevalence increases hit rates while reducing CRs. The response patterns further indicate that high-confidence hits are not exclusively associated with fast responses, as longer but confident detections represent a nontrivial component of performance, a pattern that the AJ model seeks to explain through a successful secondary information search process (i.e., recollection) engaged in when the participant enters a state of uncertainty. Additional visual representations of these data,

³ The G^2 statistic summarizes the differences between the predicted and observed response frequencies, using the maximum-likelihood parameter estimates. This statistic follows a chi-square distribution, with the degrees of freedom calculated by subtracting the number of free parameters from the number of independent response categories. In Appendix B, we provide the G^2 values and corresponding p values for each model and participant.

⁴ Note that, since we use shrinking parameters (see Appendix A), the nested model shrinking factors are set to 1, which would indicate equality between parameters. For example, if $g_2 = g_1 \cdot s_{g2}$, we would be testing whether we can choose the model where $s_{g2} = 1$ without causing a significant loss of fit. The restricted model is the same as the one that assumes $g_2 = g_1$.

Figure 6
Mean Proportions of Responses



Note. Panels show the mean proportion of responses in each category (columns) across participants, with proportions computed separately for each CL x RT bin (x-axis) and target frequency condition (rows: 65%, 50%, 35%). Proportions are calculated relative to the number of trials of the corresponding stimulus type (target or lure) within each condition. Error bars represent $\pm 95\%$ CIs based on variability across participants. Dark bars = high confidence; light bars = low confidence. CI = confidence interval; CL = confidence level; RT = response time.

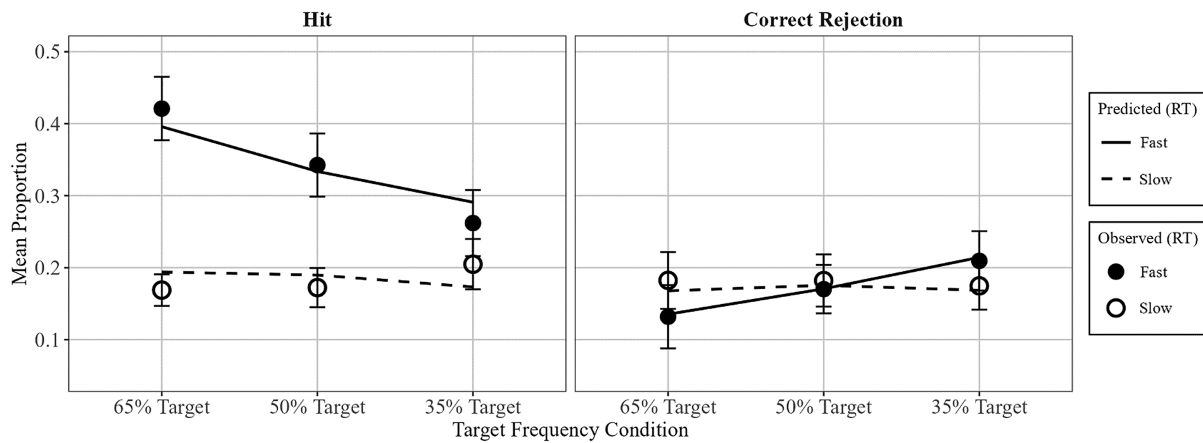
including the mean RTs for each condition, are available in [Gutkin et al. \(2024\)](#).

A striking pattern emerges in the data in [Figure 6](#) when considering correct, correct high-confidence responses. These should arise in the AJ model either when extreme familiarity values lead to fast hits or CR responses, or when intermediate familiarity values lead to an extended search producing a recollection on target trials or a failure to find the item in memory on distractor trials. If the criteria for making a decision change in response to target likelihood, then the fast, correct response proportions should change with the target frequency manipulation. The slower recollection-based responses, however, should depend mostly on the success or failure of the secondary memory search process and therefore should be more indifferent to the target likelihood, at least for high-confidence responses. Low-confidence responses should be more likely to occur as a result of guesses.

The data for correct high-confidence trials are shown in [Figures 7–9](#). These figures present a subset of the observed data shown in [Figure 6](#) (circles) and display the respective mean predictions (lines) of the AJ, 2HT, and SDT models, using parameters estimated from fitting the models to empirical data. In these figures, we can observe that the empirical pattern (the interaction) is predicted by the AJ model ([Figure 7](#)), but neither by the 2HT model ([Figure 8](#)) nor the SDT model ([Figure 9](#)). According to the AJ model, fast responses tend to be driven by familiarity values that fall outside the region of

uncertainty defined by the location of both decision criteria. If these criteria shift as a function of the expected target frequency—moving toward higher values when the proportion of targets decreases—then the proportion of extreme familiarity responses, which the AJ model claims are predominantly fast and high-confidence, should also change. Specifically, lower target frequencies should lead to fewer fast, high-confidence hits and more fast, high-confidence CRs. This pattern is not observed for slow responses, resulting in an interaction pattern that is, in fact, predicted by the AJ model, which distinguishes between fast high-confidence responses arising from extreme familiarity values and those emerging from a secondary information search process. In contrast, the 2HT and SDT models lack a theoretical mechanism to account for this interaction.

It is important to note that we cannot deterministically claim that all correct, fast, high-confidence responses reflect familiarity, nor that all slow responses reflect recollection. In fact, there is a mixture of fast and slow high-confidence responses across the familiarity-recollection continuum, and thus, these findings do not constitute definitive proof of the AJ model's predictions. However, they do reveal a pattern of results consistent with the AJ framework—specifically, that when the probability of target presentation decreases, fast, high-confidence hits decrease while fast, high-confidence CRs increase, and to a greater extent than their slow counterparts—an interaction

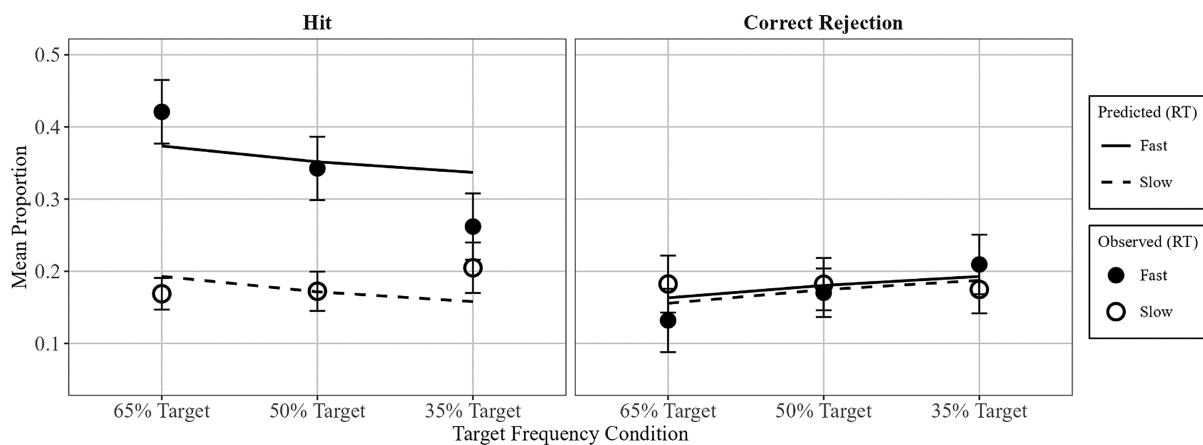
Figure 7*Observed Data and AJ Model Predictions for High-Confidence Correct Responses*

Note. The left panel displays the mean proportions of high-confidence hit responses, and the right panel shows the mean proportions of high-confidence correct rejections, both plotted as a function of the target frequency condition (65%, 50%, 35%). Observed data are shown in circles, and AJ model predictions in lines. Filled circles and solid lines represent fast responses, and open circles and dashed lines represent slow responses. Error bars represent $\pm 95\%$ confidence intervals based on variability across participants in observed data. AJ = Atkinson–Juola; RT = response time.

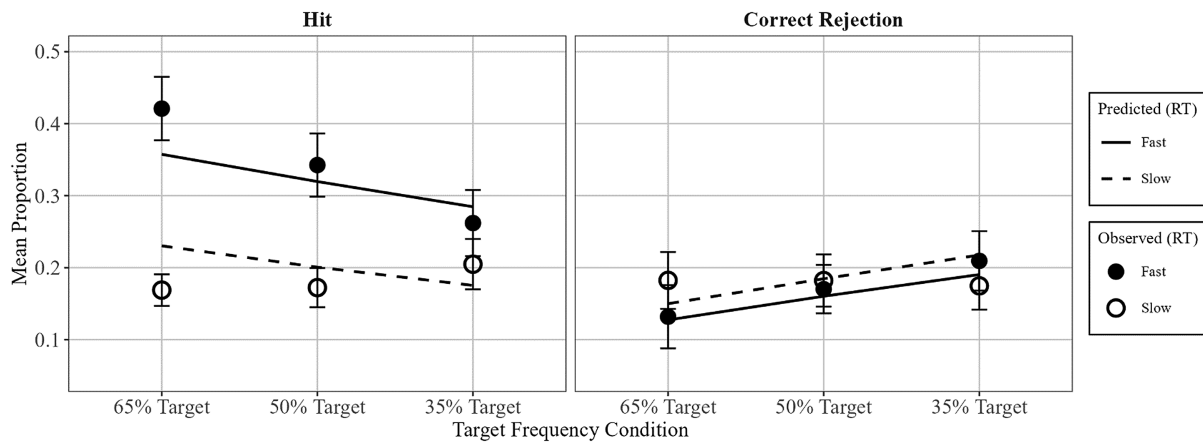
that we have not been able to account for with either the 2HT (see Figure 8) or SDT models (Figure 9), as these models lack an underlying mechanism to explain why the manipulation of target frequency would differentially affect the proportions of high-confidence hits that are fast versus slow.

In summary, Figure 7 shows the consistency between the interaction predicted from the AJ model and the empirical results, whereas Figures 8 and 9 show the lack of interaction predicted from the other two models. In the following section, we will continue comparing the 2HT, SDT, and AJ models using a model-fitting approach.

Table 1 presents the comparisons for the 2HT, SDT, and AJ models, detailing the number of parameters (left column), rejection rates of the G^2 goodness-of-fit tests (central column), and the selection frequency (and percentage; right column) as the best model according to the AIC. Initially, the top part of the table compares the SDT and 2HT models against the two AJ model versions that allow either the decision criteria or the guessing probability to vary freely across different conditions of the relative target frequency manipulation. Following these initial comparisons, the lower part of the table presents analyses between the 2HT and SDT models and two simplified versions of the AJ model. One

Figure 8*Observed Data and 2HT Model Predictions for High-Confidence Correct Responses*

Note. Mean proportions of high-confidence hits (left) and correct rejections (right) are shown as a function of target frequency (65%, 50%, 35%). Observed data are shown in circles and 2HT model predictions in lines. Filled circles and solid lines indicate fast responses; open circles and dashed lines indicate slow responses. Error bars represent $\pm 95\%$ confidence intervals across participants in observed data. 2HT = two-high-threshold; RT = response time.

Figure 9*Observed Data and SDT Model Predictions for High-Confidence Correct Responses*

Note. Mean proportions of high-confidence hits (left) and correct rejections (right) across target frequency conditions (65%, 50%, 35%). Observed data are shown as circles and SDT model predictions as lines. Filled circles and solid lines represent fast responses, while open circles and dashed lines represent slow responses. SDT = signal detection theory; RT = response time.

version assumes that the relative target frequency conditions influence the decision criteria (AJ_{cj}), while the other assumes an impact on the guessing probability (AJ_{gj}). Detailed individual analysis is provided in Table B2.

Regarding the goodness-of-fit analysis, we found that the rejection rate for the goodness-of-fit test for the SDT model was slightly higher than that of the 2HT or AJ models. The model that best fitted the data was the AJ model, with a rejection rate of 61.7%. This rejection rate decreases to 57.45% for AJ_{cj} model but increases to 76.6% for the AJ_{gj} . This suggests that there are more subjects for whom the AJ model fits acceptably.

As for the comparison between non-nested models using AIC, we observed that 27 out of 47 (57.45%) subjects' data were best fit by the AJ model. In their simplified versions, 70.21% were best fit by the AJ_{cj} and 2.13% by the AJ_{gj} . However, some subjects' data were still better fit by a nonhybrid model; six subjects (12.77%) were better fit by the 2HT model and seven (14.89%) by the SDT model.

Table 1
Summary of Model Fits

Model	K	H_0 .reject	H_0 %reject	AIC.best	AIC%best
2HT	17	31	65.96	10	21.28
SDT	11	38	80.85	10	21.28
AJ	20	29	61.70	27	57.45
2HT	17	31	65.96	6	12.77
SDT	11	38	80.85	7	14.89
AJ_{cj}	18	27	57.45	33	70.21
AJ_{gj}	18	36	76.60	1	2.13

Note. Summary of best model selection and goodness-of-fit. From left to right, the first column indicates the model to which each row refers, the second the number of parameters (K) it includes, the third (H_0 .reject) and fourth (H_0 %reject) the frequency (number of subjects, out of 47 total) and percentage of rejection of the goodness-of-fit hypothesis, and in the last two columns (AIC.best and AIC%best) the frequency and percentage with which each is selected as the best model using the AIC index. AIC = Akaike information criterion; 2HT = two-high-threshold; SDT = signal detection theory; AJ = Atkinson–Juola.

The estimators used for the goodness-of-fit test and AIC comparisons can be found in Table B2.

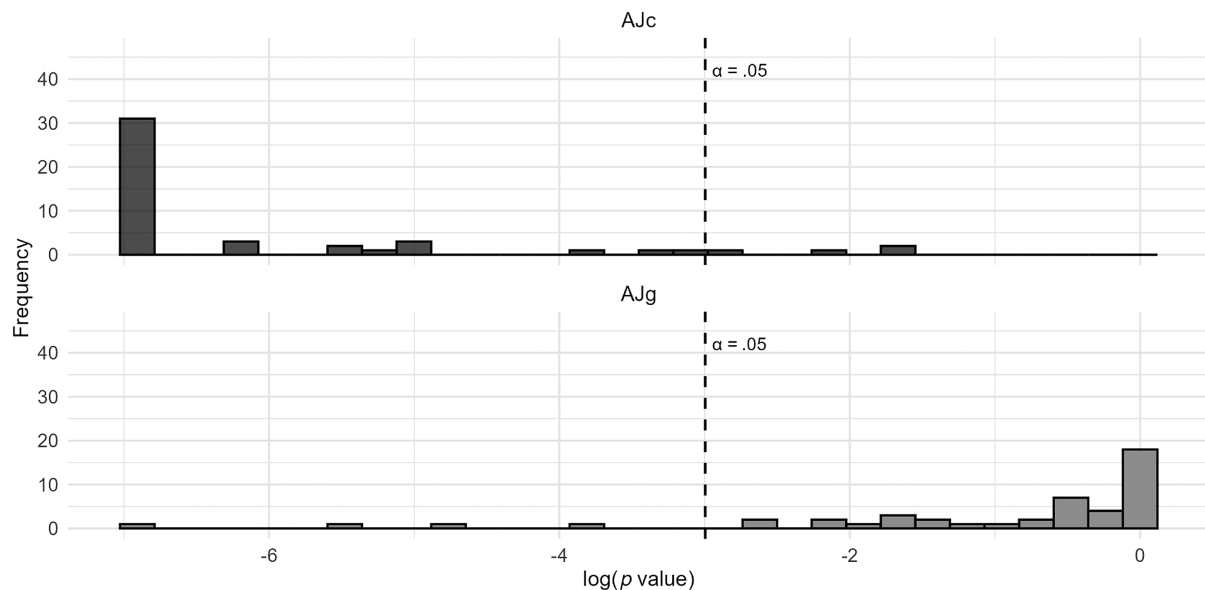
To assess the direction of the changes in c_L and g across conditions, as well as the shape of their relation (e.g., linear), we tested whether, as predicted by the AJ model, these estimated parameters have an inverse and direct linear relationship with the relative target frequencies, respectively (see Figure 10). For the AJ_{cj} version (top panel), we found a significant effect of the expected inverse monotonic relationship for 43 subjects (91.5%). However, for the AJ_{gj} version (bottom panel), we found that only four (8.5%) showed a direct relationship between g and the experimental conditions. This indicates that AJ model predictions regarding the effects on c_L and g across relative target conditions are generally supported for the AJ_{cj} but not for the AJ_{gj} model.

Next, we evaluate the RT predictions. Figure 11 presents the results of testing the relationship hypothesized by the AJ model. Specifically, in line with the AJ model, it is hypothesized that responses derived from extreme familiarity states should be faster than those associated with successful search states, and that these should be faster than guess old or guess new responses (i.e., $L_f > L_s$, $L_f > L_{go}$, and $L_f > L_{gn}$). This pattern is explained by the fact that search and guessing responses are the result of a second-stage process that is triggered only when the cognitive system determines that familiarity falls within the uncertainty region. In contrast, responses based on extreme familiarity occur before this secondary stage is completed, producing quicker decisions based only on familiarity. This relation is significant for 93.6% of subjects' data sets when fitting the AJ_{cj} model (top panel) and for 70.2% when fitting AJ_{gj} (bottom panel). This indicates that AJ model predictions regarding RTs are generally supported both for the AJ_{cj} and the AJ_{gj} models.

In Figure 12, we present the results of hypothesis testing concerning the estimated parameters modeling high-confidence responses from RT-extended branches. Specifically, these parameters represent the probability of high-confidence responses derived from processes of extreme familiarity ($H_{f, b}$), guessing responses (for old and new

Figure 10

Nested Model Comparison Log(p): Testing the AJ Model Predictions Regarding Criteria and Guessing Probability



Note. Histogram of log-transformed p values from likelihood ratio tests (LRTs) comparing nested models under different assumptions about the relationship between target frequency and the parameters c_{Lj} (for the AJ_{cj} model; top panel) or g_j (for the AJ_{gj} model; bottom panel). The full AJ_{cj} (AJ_c) model assumes an inverse linear relationship between c_{Lj} and target frequency and is tested against a nested null model assuming uniformity across conditions. The AJ_{gj} (AJ_g) model assumes a direct linear relationship between g_j and target frequency and is compared to a null model assuming no variation across conditions. The x -axis shows the log-transformed p values obtained from individual LRTs (one per participant); the y -axis indicates the number of data sets (participants) whose p values fall within each bin of the histogram. The vertical dashed line marks the significance threshold at $\alpha = .05$; values to the left indicate significant effects consistent with the model's predictions. The proportion of data sets supporting the model predictions is obtained by computing the percentage of participants whose LRT p values fall below the significance threshold ($p < .05$), that is, the proportion of histogram area located to the left of the vertical dashed line. AJ = Atkinson–Juola.

responses: $H_{go, b}$ and $H_{gn, b}$), and fast and slow successful searches ($H_{s, b}$). We evaluated whether, as predicted by the AJ model, responses derived from successful searches exhibit higher confidence than those associated with either guessing or extreme familiarity states (i.e., $H_{s, b} > H_{f, b} > H_{go, b}$ and $H_{s, b} > H_{f, b} > H_{gn, b}$). This pattern is grounded in the assumption that, within the AJ model, search responses result from a successful, information-based memory search, whereas guessing responses are triggered when this search fails. As a result, search responses are expected to be more confident than guesses. Furthermore, search-based responses may also be more confident than those based purely on familiarity, as they reflect the outcome of a more deliberate and extended retrieval process, rather than a quick, automatic reaction to familiarity. This hypothesis holds significantly for 89.4% of the subjects' data sets when fitting the AJ_{cj} model, as illustrated in Figure 12 (top panel). It also holds for 93.6% of subjects under the AJ_{gj} model (bottom panel). Suggesting that the AJ model's hypotheses about CLs are generally supported for both the AJ_{cj} model and the AJ_{gj} model versions.

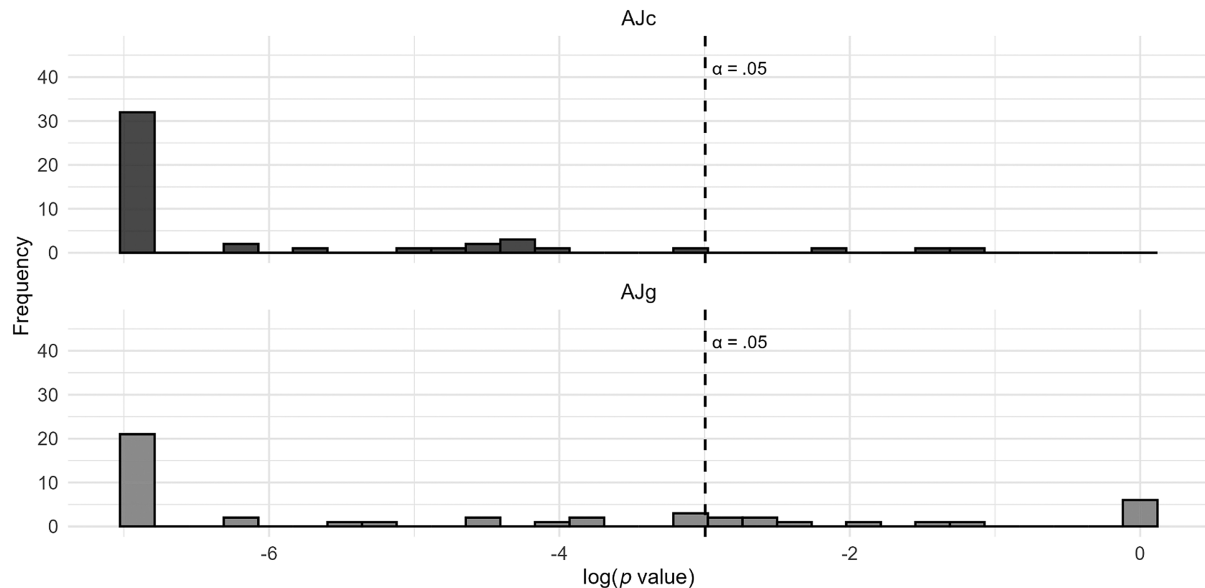
The full log-likelihood ratio test statistics used to construct Figures 10–12 are provided in Appendix D.

Discussion

The ongoing question about whether recognition memory relies on a latent continuous variable (familiarity) or on discrete states has been often studied by comparing the 2HT and SDT models (e.g.,

Atkinson & Juola, 1974; Bröder & Schütz, 2009; Green & Swets, 1966; Kellen & Klauer, 2015; Luce, 1986; Macmillan & Creelman, 2004). Research designed to discriminate between these models has provided mixed results. In the present study, we have introduced a third possibility, which we call the extended AJ model. The original AJ model, proposed by Atkinson and Juola (Atkinson & Juola, 1973, 1974; Juola et al., 1971), differs from the SDT model in that it does not assume a single criterion but rather two (low and high criteria), between which an area of uncertainty is defined. When the familiarity value is within the uncertainty area, a response is delayed until sufficient information is retrieved from a long-term memory search, or until a response criterion is exceeded and a guess is made. Consequently, search-based or guess responses should be slower than those associated with extreme levels of familiarity. We believe that our current efforts represent the first attempt to extend, through an MPT approach, the modeling of processes that occur following a state of uncertain familiarity, as described in the original AJ model proposal (Atkinson & Juola, 1973, 1974; Juola et al., 1971). Additionally, we employ variables that have proven relevant for model recovery of recognition memory data (Gutkin et al., 2024), such as RTs and CLs. The result is a hybrid model that combines elements of the SDT and 2HT models.

When comparing the extended AJ, SDT, and 2HT models, we observed that the AJ model provided the best fit for most subjects' data. Specifically, the AJ model's fit was better for 57.45% of subjects. This percentage increases to about 70.2% for the version that

Figure 11*Nested Model Comparison Log(p): Testing the AJ Model Predictions Regarding RTs*

Note. Histogram of log-transformed p values from likelihood ratio tests comparing nested models under different assumptions regarding L parameters (RT predictions). For the AJ_{cj} (top panel) and AJ_{gj} models (bottom panel), the tests assess the complex models, which include ordinal constraints on the parameters ($L_f > L_s$, $L_f > L_{go}$, and $L_f > L_{gn}$) against their null counterparts that assume all L parameters are equal (null hypotheses). The x-axis shows the log-transformed p values from these tests; the y-axis represents the number of participants exhibiting each p value. The vertical dashed line marks the significance threshold at $\alpha = .05$; values to the left indicate significant effects consistent with the model's predictions. AJ = Atkinson–Juola; RT = response time.

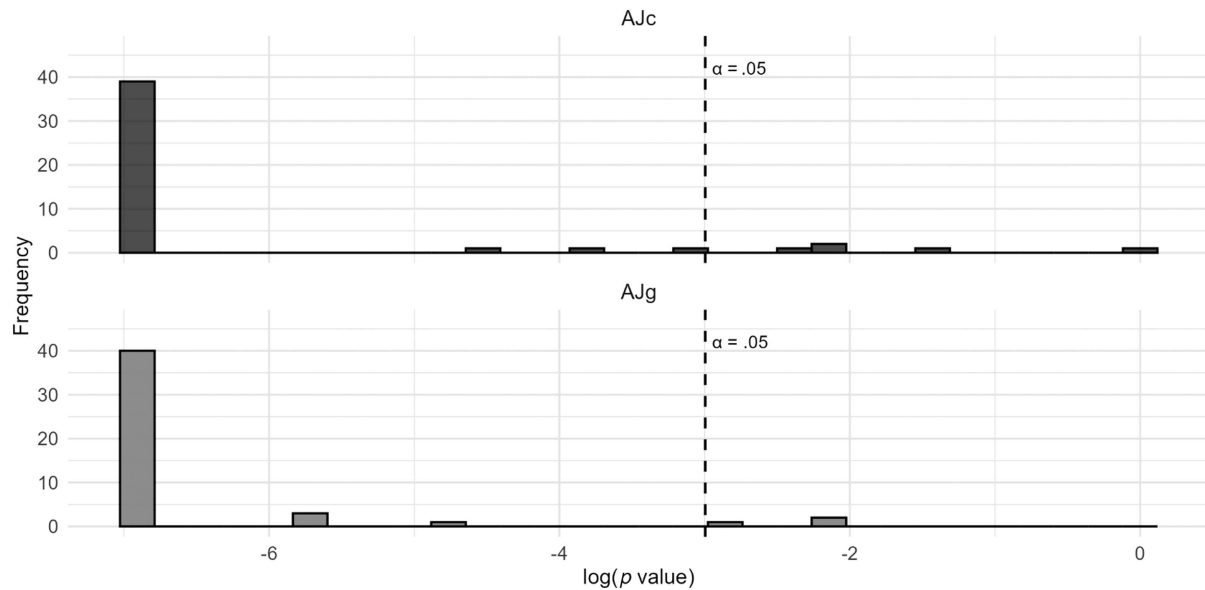
assumes that their internal criteria shift along with the manipulation of relative target frequency, while the version that allows only the guessing parameter to vary freely with target proportions gives a better fit for only 2.1% of the participants. This suggests that the superior fit of the AJ model cannot be explained solely by its larger number of parameters.

The reason why the AJ model provides a better fit to the data set could be precisely because it captures certain response patterns that the other models cannot. Specifically, the AJ model proposes a sequence of events on all trials, beginning with a familiarity assessment and possible rapid, typically confident response, or a delay in responding in which a successful or failed recollection occurs before the response. In the cases in which the participant awaits the outcome of a slower, secondary recollection process, typically a highly confident response is made if the search is successful, and a less confident response or a guess results when recollection fails. Therefore, slow, high-confidence responses show a pattern distinct from that for fast, high-confidence responses, in that slow responses are relatively immune from the effects of the target proportion manipulation (see Figures 7–9), whereas fast responses show large effects of this manipulation. This pattern is predicted by the AJ model in that fast responses should be more representative of decisions based on familiarity, while slow responses should be more reflective of the results of the extended search and recollection processes. In the nested comparison tests, we compared the full AJ model, which assumes this mechanism, with a restricted version that does not. The restricted model showed a significantly poorer fit, supporting the necessity of the extended

search mechanism to account for such patterns in the data. In contrast, the SDT model predicts that hits tend to be fast and high-confidence, as they occur when familiarity exceeds a decision criterion. Similarly, in the 2HT model, high-confidence hits are fast when they result from detection, whereas hits based on guessing are slower and lower in confidence. Thus, the pattern of slow, high-confidence hits lacks a clear explanatory mechanism in SDT or 2HT, making it a distinctive feature of the AJ model.⁵

MPT models are valuable tools for validating the structure of theories that posit latent cognitive processes behind observed responses (Cavagnaro et al., 2013; Marewski & Olsson, 2009; Myung et al., 2017). However, traditional categorical MPT models have been criticized for their limitations in effectively disentangling the nature of the processes involved in each model. While this flexibility can be seen as a weakness, it also provides a strength: MPTs allow researchers to evaluate a wide range of process architectures, such as sequential or parallel processing,

⁵ It must be acknowledged that, with the model-fitting approach we have adopted, although the tests conducted are consistent with the predictions of the AJ model, we cannot directly determine from these data whether this distinctive pattern is the reason why the AJ model provides a better fit than the SDT and 2HT models. Although it was not the original objective of the present study, as the reviewers have suggested, a critical test approach that does not require model fitting could effectively complement the current proposal and strengthen—or falsify—these distinctive predictions of the AJ model. See Appendix F for a possible proposal to evaluate such predictions without the need for model fitting; note that this would require a different type of data manipulation, which, as we comment later in the present Discussion (p. 29), would be of interest to collect in future research.

Figure 12*Nested Model Comparison Log(p): Testing the AJ Model Predictions Regarding CLs*

Note. Histogram of log-transformed p values from likelihood ratio tests testing the ordinal constraints on H parameters (CL predictions). For both the AJ_{cj} (top panel) and AJ_{gj} models (bottom panel), these comparisons contrast the complex models, which impose ordinal constraints on the parameters (alternative hypotheses: $H_{s,b} > H_{f,b} > H_{go,b}$ and $H_{s,b} > H_{f,b} > H_{gn,b}$), with their corresponding null models, which assume equality among all H parameters (null hypotheses). The vertical dashed line indicates the significance threshold ($\alpha = .05$); p values to the left reflect significant support for the predicted orderings. AJ = Atkinson–Juola; CL = confidence level.

in a mathematically elegant and intuitive framework. Also, we believe that incorporating relevant variables (e.g., CLs and RTs) can enhance the evaluation of the cognitive processes involved by testing distinctive predictions derived from such cognitive states. These predictions include the order in which different processes occur, their cognitive cost, the processing time associated with this cost, and other indirect measures of their characteristics (e.g., see Klauer & Voss, 2008; Laukenmann et al., 2023). The importance of validating predictions about variables that emerge from hypothesized latent cognitive states is illustrated here as a means to differentiate among various theories of recognition performance. If there is a contradiction between a theory and a statistical model, one or both must be incorrect. Thus, in the present case, incorporating CLs and RTs helps counterbalance the model's flexibility by imposing empirical constraints on the RT and CL predictions associated with specific cognitive processes.

There are also well-established and more constrained modeling frameworks that allow for the joint analysis of continuous and discrete data, such as diffusion models. However, only a limited number of these have been used to simultaneously account for both RT distributions and CL judgments (see Starns & Heathcote, 2024, for a detailed discussion). Among the earliest models tackling both RTs and CLs are the discrete random walk models (Audley, 1960; Volkmann, 1934), which conceptualize decisions as a series of evidence steps toward one of two boundaries. Here, the number of steps required is directly related to decision time, and therefore, CLs are entirely determined by RT. This strict coupling, however, fails to explain frequent empirical cases where high-confidence judgments follow relatively slow responses. The sliding window model (Juslin &

Olsson, 1997) proposes confidence as a process of averaging evidence over time. While this approach can account for overlap in RT distributions across CLs, it has been criticized for implausible predictions and limited empirical support (Vickers & Pietsch, 2001). A different approach is the two-stage diffusion model (Pleskac & Bussemeyer, 2010), where an initial decision is followed by post-decisional accumulation to assess confidence. Here, confidence is treated as a postdecisional evaluation, derived from additional information gathered after the decision. This framework disrupts the traditional link between fast responses and high confidence by allowing for high-confidence judgments even after prolonged deliberation. Because CLs follow their own independent stopping rules, the model can explain why confidence and RTs are not always tightly correlated. However, it does not account for how qualitatively distinct memory processes—like familiarity and recollection—may differentially influence accuracy, CLs, and RT distributions. Other models, such as RTCON and its extension RTCON2, assume that confidence is determined simultaneously with the choice (Ratcliff & Starns, 2009, 2013). These frameworks employ multiple evidence accumulators running in parallel, each linked to a specific combination of a decision outcome and a CL—for example, “sure old” or “probably old.” Ratcliff and Starns (2013) acknowledge substantial individual variability in recognition memory performance. RTCON2 captures this by allowing key parameters—such as memory strength, decision thresholds, and confidence criteria—to vary across individuals, demonstrating how different parameter sets can lead to similar or divergent behavioral patterns. However, the model does not attribute these differences to qualitatively distinct memory processes, such as a predominance of recollection-based versus familiarity-

based responses. More recent models by Osth, Bora, et al. (2017; Osth, Dennis, & Heathcote, 2017; Osth et al., 2018) offer a unified framework in which RTs and CLs emerge jointly from a single evidence accumulation process influenced by memory strength. These models incorporate global matching and retrieval dynamics to generate trial-by-trial predictions, where confidence is linked to the strength of the retrieved trace.

Ultimately, evidence accumulation models offer a principled way to account for how decision boundary settings affect the trade-off between speed and accuracy. Narrow boundaries lead to faster responses that are more prone to noise-induced errors, whereas wider boundaries result in slower but often more accurate responses and are accompanied by higher confidence due to additional accumulation of evidence. Such behavioral patterns are also captured by the AJ model, in which fast, low-confidence responses are linked to familiarity, and slower, high-confidence responses to recollection. However, evidence accumulation models typically rely on a continuous sampling process, rather than assuming two qualitatively distinct, sequential mechanisms. It is plausible that familiarity-based decisions require less evidence accumulation—reflected in lower thresholds and reduced sensitivity—while recollection entails deeper, more cautious sampling, involving higher thresholds and yielding more precise, confident judgments, at least when recollection succeeds. We find it highly promising to test the predictions of both the general and AJ models using diffusion-based approaches. Such integration could enhance our current model, which does not yet explicitly incorporate trial-by-trial variability.

We view MPTs and evidence accumulation models as complementary frameworks. Supporting this view, a recent study by Klauer et al. (2024) proposed an integration of both approaches. Traditional diffusion models conceptualize decision making as a single process of evidence accumulation regulated by decision thresholds. In contrast, Klauer et al.'s proposal decomposes decision making into multiple explicit processes, allowing different accumulation mechanisms for each branch in the model. This enables a clear separation of processes such as detection, recognition, and guessing. This integrated framework allows researchers to impose stronger theoretical constraints—borrowed from evidence accumulation models—while preserving the structural flexibility of MPTs. It presents a promising approach that should be further applied in future research and could be useful for testing the predictions of our general model in upcoming studies.

According to the AJ model, judgments based on extreme familiarities should be of relatively high confidence and result in the shortest average RTs. In fact, some authors (e.g., Weidemann & Kahana, 2016) have argued that the two should be highly correlated, enough so that RTs could be used as a simpler way to measure confidence without the necessity of adding confidence ratings after old/new responses in recognition memory tests. Our earlier results (Juola et al., 2019) demonstrated that this assumption might be unwarranted, as receiver operating characteristic curves generated from the same data set based on RT and confidence bins differed markedly. The RT receiver operating characteristic curve was convex and symmetric, whereas the confidence receiver operating characteristic was convex and asymmetric, in a way that supports the argument that the asymmetry resulted from increasingly informative target evidence beyond that derived from familiarity alone (e.g., Dobbins, 2023). It is likely that in Juola et al.'s (2019) study, in

which participants were encouraged to respond quickly and then give a confidence rating, confidence could have been influenced by additional reflection and recollection of target information after the initial response had been made. Such recollection could also have occurred during slower responses, leading to more highly confident ratings. Overall, responses based on familiarity plus the results of successful episodic search should be more confident than those based on familiarity alone, and the fastest responses should not necessarily be the most confident (see Kang et al., 2025, for similar arguments).

Incorporating RTs and CLs into the AJ, SDT, and 2HT models is feasible through the modeling approach proposed by Heck and Erdfelder (2016), which allows the integration of continuous variables into MPT models, such as the 2HT or SDT, once they have been reparametrized as multinomial trees.⁶ The inclusion of CLs and RTs has demonstrated benefits in previous studies of recognition memory (Gutkin et al., 2024), such as improving the power of selection between non-nested models (e.g., SDT vs. 2HT). Also, including CLs and RTs is particularly relevant to the AJ model, as it enables the extension of its original version and facilitates modeling processes within the “area of uncertainty,” such as a search mechanism that, if successful, leads to a correct response, or, if unsuccessful, results in guessing. Within this framework, three types of hit sources can be distinguished: (a) hits resulting from successful searches, characterized by high confidence and long RTs due to the involvement of a secondary process; (b) hits originating from guessing states, also associated with long RTs due to the secondary process but linked to low confidence because the search failed; and (c) hits generated by states of extreme positive familiarity, which should be the fastest responses with moderate or high confidence. In this study, we evaluated whether these predictions of the AJ model hold true.

Regarding our results for predictions of CLs and RTs, both versions of the AJ model (AJ_{cj} and AJ_{gj}) were validated, as post hoc tests revealed significant effects for the vast majority of participants. However, notable differences emerged between the AJ_{cj} and AJ_{gj} models in their predictions regarding the effects of the relative target frequency manipulation. Specifically, the AJ_{gj} model predicts that, for most participants, variations in the guessing probability parameter due to target frequency manipulation were minimal, enabling a simplified model in which guessing probability is assumed to remain constant. This likely accounts for the fact that 74.5% of data sets select AJ_{cj} as the best fitting model, compared to only 2.1% for AJ_{gj} . These findings highlight the importance and resulting validity of allowing criteria to vary as a function of relative target frequency manipulations within the AJ model.

Nonetheless, our results show that there are remaining participants whose data were better fit by the SDT or the 2HT model. Although the AJ model fits the data of most participants better and makes predictions consistent with theoretical expectations, variability among the selection models remains. This variability could be due to the model-fitting process itself or may genuinely reflect individual differences. Our simulation results in Appendix C

⁶ Note that we refer here only to the nonparametric approach for incorporating continuous variables into MPT models, as current parametric RT-MPT approaches and the corresponding R packages available for their estimation do not support continuous process models (such as the SDT model), which include parameters for modeling categories (node probabilities) with ranges extending beyond [0, 1].

suggest that the AJ model does not fit the data better simply because it is more flexible than the SDT or 2HT models. In fact, our findings indicate that the SDT model shows greater flexibility than both the 2HT and AJ models. While SDT has fewer parameters, the way they combine to characterize different aspects of the normal probability density function can result in greater functional complexity (Myung et al., 2017). Also, SDT handles CLs responses (from “new-high” to “old-high”) by creating three criteria that vary with target proportion manipulations, which likely allows it to fit a wider range of data sets. In contrast, the AJ and 2HT models map confidence to probability parameters (H), as in threshold models.⁷ Therefore, the reason that the AJ model fits the majority of subjects cannot be explained by model complexity alone.

The between-subjects variability regarding model selection between nonnested models has major implications. On the one hand, if different models coexist, it is incorrect to apply a hierarchical MPT model or to aggregate data from several subjects, as both approaches assume that all subjects follow the same model. On the other hand, we will have to explain why there are subjects whose data fit better with one or the other model. In our opinion, we believe that it is not enough to state that subjects have different cognitive strategies, but rather, if they do exist, these should be modeled. This is something that has not yet been done with any of the models compared in our study.

Although modeling individual strategies is beyond the original scope of this research, we suggest here that there may be a more general model, similar to the current hybrid AJ model, but one that fully encompasses both the SDT model and the 2HT model (See Appendix E). We will refer to this as the general model (see Figure E1). To understand this hypothetical model, let us consider a situation in which the criteria (low criterion and high criterion) separate symmetrically with respect to a central axis. That is, if c_H increases, then c_L decreases, causing the distance between the criteria to tend to infinity. In this case, the greater the distance, the higher the probability will be that the initial familiarity judgment will fall in the uncertainty area. Taken to the extreme, the probability of the uncertainty state would be 1 (see Figure E2), while the other branches (+ or –familiarity) would have a probability of 0. Consequently, when the distance between the criteria tends to infinity, the overall model becomes equivalent to a 2HT model. That is, no responses would be made on the basis of familiarity, and a long-term memory search for episodic information identifying whether the test item is a target would have to be made before any response. To recover the 2HT model from this scenario, it would be necessary to identify search results that either confirm or refute the test item’s inclusion with the studied material. If this process results in uncertainty, a guess will have to be made. Conversely, if the difference between criteria decreases, the probability of uncertainty tends to 0. In the latter case, the branches associated with uncertainty states would disappear, and the AJ model would become equivalent to an SDT model (see Figure E3). Therefore, by increasing or decreasing the distance between the criteria, we could obtain the SDT and 2HT models as special cases of a general AJ model.

The benefits of a general model would be, on the one hand, methodological, as comparisons would move from nonnested comparisons (2HT vs. SDT vs. AJ model) to comparisons between nested models (general vs. nested models). Thus, the question would not be whether a subject’s data are better accounted for by the discrete 2HT model or the continuous SDT model, but to what

extent a subject’s parameter vector allows for simplification of the general model without loss of fit. In this case, when the criterion discrepancy approaches zero, the log-likelihood of the general model and the SDT model would be equivalent, and we would choose the simpler model, the SDT. Conversely, when the uncertainty probability increases, we should find that the log-likelihood of the overall model and the 2HT model do not differ significantly. Furthermore, if the difference between models is indeed attributable to a specific parameter that, as proposed here, defines the distance between criteria, this would allow for the use of a partial pooling RT-MPT model (also known as hierarchical), such as the one proposed by Klauer and Kellen (2018) or Klauer et al. (2024). This approach offers several benefits: on one hand, these methods have been shown to be the most robust and consistent for estimating MPT models, accounting for subject similarities, reducing standard errors, and requiring fewer data points per subject (Chechile, 2009; Singmann et al., 2024). On the other hand, they enable an evaluation of which variables or factors influence parameter variability.

The general model would also have substantive implications. Each subject’s c_L and c_H parameters would specify the fitting distance to either the discrete or the continuous model. Moreover, we consider that the probability of each state in the general model might depend on other factors, such as the shape of the signal and distractor distributions, the family of probability density functions, etc. Essentially, modeling individual strategies through a general model that encompasses both SDT and the 2HT models revolves around the distribution of the weight assigned to each branch of the model and can be achieved by manipulating the distance between the high and low decision criteria. However, this is only a subcase of the general model, as the weight assigned to each branch may be determined by other factors (e.g., the shapes of the familiarity distributions for distractors and targets). Furthermore, it is important to note that variations in the criteria could lead to the emergence of additional model subcases beyond the 2HT and SDT models. For example, if c_H increases and c_L remains constant, the area of +familiarity could disappear while the –familiarity area remains. A model capable of accommodating all these particularities would be overly complex and probably of little use in providing a concise description of reality and deriving generalizable conclusions. Nevertheless, if we were to use a simpler model, such as the one suggested here, we would have to find out whether this is consistent with reality and what experimental manipulations are necessary to influence subjects’ strategies.

If subjects employ different strategies or have response patterns that depend on specific tendencies or traits like impulsivity and conscientiousness, these should manifest themselves through changes in the parameters that define the uncertainty area (e.g., the c_L and c_H parameters), enabling detailed studies of such variations. The adoption of one strategy or response pattern over another would not be simply binary but a matter of degree. For example, a perfectionist—whose traits can be measured—may handle uncertainty differently than

⁷ One might argue that the AJ model should handle CLs by adding more criteria along the familiarity continuum, similar to how it can be done in the SDT model. However, in the uncertainty area, where the AJ model behaves like a threshold model with discrete processes, this approach is less applicable. This area is crucial for the AJ model because it uniquely predicts slow, high-confidence search responses, which are not expected in the 2HT and SDT models. Given the significance of the processes associated with this area, which behaves like a threshold model, CLs have been mapped accordingly.

others. The same level of familiarity might lead one individual to initiate a second-stage memory search, while another person (or even the same person in a different context) may respond directly based on that initial familiarity. A perfectionist might require a more extreme familiarity signal—either high or low—before committing to a response, leading to a wider separation between c_L and c_H . A wider separation would increase the range of familiarity values that fall into the uncertainty area, leading to a greater proportion of secondary search processes made before the response. This does not mean that they have lower confidence in their responses, but rather that their internal threshold for what qualifies as “extreme familiarity” is more stringent, making them more likely to engage in a secondary memory search before responding. When a response waits for the outcome of this secondary (recollective) process, it will naturally be slower. If recollection succeeds, the response will be both more accurate and highly confident—even if the initial familiarity was not particularly extreme. In contrast, if recollection fails, the response will be slower, less confident, and will tend to be at chance level, as it defaults to guessing. Conversely, a less motivated or less perfectionist subject may be more likely to respond before engaging in a secondary search (narrow separation between c_L and c_H). Evaluating certain individual characteristics and cognitive styles in advance could provide insights into why one model sometimes fits better than another (see, e.g., Romero et al., 2022). It would therefore be interesting for future studies to evaluate, through personality assessments or test–retest methods,⁸ whether intrasubject stability can be observed in their strategies and, consequently, to identify the model that best explains their behavior.

The parametric values estimated for each participant—such as the criterion distance discussed above—may not solely reflect stable individual traits but can also be influenced by task instructions or experimental context. Importantly, there is a range of experimental manipulations, such as variations in instructions, the degree of initial learning, stimulus characteristics, or retention intervals, that can shift the balance between the relative contributions of recollection and familiarity to the recognition response (Wixted, 2009). Such manipulations could allow us to test predictions that are specific to the AJ or general model but not captured by the SDT or 2HT model. For instance, through a payoff matrix manipulation, we can set rewards for hits and CRs and adjust penalties for FAs and misses in order to affect the participants’ placements of the response criterion. In the general/AJ model, any manipulation that induces greater caution on the part of the observer could increase the uncertainty region that would result in a greater proportion of slow, high-confidence hits. Such responses are expected to occur more frequently under conditions that expand uncertainty, as opposed to those that constrain it. This represents a distinctive, model-derived behavioral prediction—one that SDT and 2HT models do not readily accommodate (see an example in Appendix F). Additional influential factors can influence the probability of a successful recollection, such as encoding depth and contextual richness. Some individuals may tend to rely more on familiarity-based “know” judgments, while others might favor recollection-based “remember” responses (see Kang et al., 2025). For example, semantic or elaborative encoding—such as forming meaningful associations—has been shown to increase the likelihood of “remember” responses, which are typically linked to conscious recollection (Gardiner, 1988; Rajaram, 1993). In contrast, shallow encoding tends to increase “know” responses and reliance on familiarity. Additional

manipulations known to differentially affect recollection and familiarity can be found in Migo et al. (2012) and Yonelinas (2002). To date, such manipulations have rarely been explored in combination with both RT and CL data in recognition memory tasks. Incorporating them would offer a more focused, theory-driven approach to testing the unique predictions of the general/AJ model, particularly those concerning the interactions among accuracy, confidence, and decision latency.

Future studies that investigate the general model will need to address complexities such as determining which types of controlled experimental conditions should be used to avoid violations of necessary constraints. These studies should incorporate accuracy measures, CLs, and RTs and explore whether there are factors that systematically affect the probability of the uncertainty zone and, if so, whether they can be experimentally manipulated or if they are stable over time or driven by the characteristics of the subjects. Another issue of interest is the utilization of a hierarchical RT-MPT model (partial pooling RT-MPT model) that allows for modeling the factors influencing variability in parameters found among participants. For this purpose, it would be necessary to create libraries capable of fitting hierarchical models that are not strictly MPTs. Since the general model contains continuous processes, the probability of some nodes in the reparametrized model is modeled through a function and not a single probability parameter, as in MPTs. To our knowledge, the only library that allows for modeling these reparametrized multinomial tree models is that of Singmann and Kellen (2013), which is not hierarchical and only allows for modeling nonparametric MPTs such as that proposed by Heck and Erdfelder (2016). Other methodological issues are due to the fact that, for some participants, the general model could be equivalent to its 2HT and SDT subcases—and therefore, some of the branch probabilities would lie on the boundaries of their parameter space. This requires verification of whether the estimates are unbiased and accurate in these situations and whether we can hold the asymptotic properties of fitting measures, such as following a χ^2 distribution for the goodness-of-fit test (see also Davis-Stober, 2009; Read & Cressie, 1988; Silvapulle & Sen, 2005). Future research on the general model should also rely on model comparison approaches that go beyond the model-fitting exercises, which is a limitation of the present study. For instance, one could adopt a confirmatory approach—such as the one we previously mentioned—by identifying “signature” predictions of each model and testing them against the data, as in a critical tests approach (e.g., Bussemeyer & Wang, 2000; Dobbins, 2023; Kellen et al., 2025; Kellen & Klauer, 2015). In Appendix F, we outline a proposal for addressing this issue by using the ratio of fast versus slow high-confidence responses under experimental manipulations that increase the separation between high and low criteria. It should be noted that this analysis was not feasible with the current data because, although we have RTs, CLs, and accuracy measures—essential for detecting the idiosyncratic

⁸ As an alternative to a test–retest approach, split-half analyses could be performed. However, no-pooling models require a large amount of data to achieve precise estimates, which makes the use of split-half analyses challenging, as they demand an unrealistically high sample size per subject. Ideally, hierarchical models should be employed, as they are more robust when dealing with a smaller number of observations per subject. The latter approach would also necessitate the development of new RT-MPT libraries to support the application of reparametrized models that are not strictly MPTs, such as SDT or AJ models (see Footnote 6).

pattern of the AJ model—the manipulation of target frequency does not predict an increase in the separation between high and low criteria, but rather a parallel shift of both. This represents a clear avenue for future research.

In short, our aim has been to derive a possible solution to the question of whether recognition memory is best represented as a discrete or continuous process. We attempted to provide a solution in the form of an extended hybrid model that, although better suited to a description of most subjects' behaviors than the SDT or 2HT models, still leaves uncertainties about the consistency between any individual's performance and the various models under consideration. Perhaps future research can lead to a general recognition model that incorporates the effects of individual differences in cognitive abilities and strategies while still being able to account for the overall trend in performance.

References

- Agresti, A. (2012). *Categorical data analysis* (Vol. 792). John Wiley & Sons. <https://doi.org/10.1002/0471249688>
- Atkinson, R. C., & Juola, J. F. (1973). Factors influencing speed and accuracy of word recognition. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 583–612). Academic Press. <https://escholarship.org/content/qt6nm5g3qj/qt6nm5g3qj.pdf>
- Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 1, pp. 243–293). W. H. Freeman. <https://escholarship.org/content/qt40x8f9s7/qt40x8f9s7.pdf>
- Audley, R. J. (1960). A stochastic model for individual choice behavior. *Psychological Review*, 67(1), 1–15. <https://doi.org/10.1037/h0046438>
- Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, 55(4), 412–428. <https://doi.org/10.3758/BF03205299>
- Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3), 929–945. <https://doi.org/10.1037/0096-1523.24.3.929>
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6(1), 57–86. <https://doi.org/10.3758/BF03210812>
- Bröder, A., & Schütz, 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. <https://doi.org/10.1037/a0015279>
- Burnham, K. P., & Anderson, D. R. (Eds.). (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer. <https://doi.org/10.1007/b97636>
- Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, 44(1), 171–189. <https://doi.org/10.1006/jmps.1999.1282>
- Cavagnaro, D. R., Myung, J. I., & Pitt, M. A. (2013). Mathematical modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 1, pp. 438–453). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199934874.013.0021>
- Chechile, R. A. (2009). Pooling data versus averaging model fits for some prototypical multinomial processing tree models. *Journal of Mathematical Psychology*, 53(6), 562–576. <https://doi.org/10.1016/j.jmp.2009.06.005>
- Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology*, 53(1), 1–13. <https://doi.org/10.1016/j.jmp.2008.08.003>
- Dobbins, I. G. (2023). Recognition receiver operating characteristic asymmetry: Increased noise or information? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(2), 216–229. <https://doi.org/10.1037/xlm0001224>
- Erdfelder, E., Auer, T.-S., Hilbig, B. E., Abfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie / Journal of Psychology*, 217(3), 108–124. <https://doi.org/10.1027/0044-3409.217.3.108>
- Erdfelder, E., Quevedo Pütter, J., & Schnuerch, M. (2024). On aggregation invariance of multinomial processing tree models. *Behavior Research Methods*, 56(8), 8677–8694. <https://doi.org/10.3758/s13428-024-02497-y>
- Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, 16(4), 309–313. <https://doi.org/10.3758/BF03197041>
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). John Wiley & Sons.
- Gutkin, A., & Heck, D. W. (2026). Extensions of multinomial processing tree models for continuous variables: A simulation study comparing parametric and non-parametric approaches. *Behavior Research Methods*, 58, 22. <https://doi.org/10.3758/s13428-025-02896-9>
- Gutkin, A., Suero, M., Botella, J., & Juola, J. F. (2024). Benefits of multinomial processing tree models with discrete and continuous variables in memory research: An alternative modeling proposal to Juola et al. (2019). *Memory & Cognition*, 52(4), 793–825. <https://doi.org/10.3758/s13421-023-01501-8>
- Hartmann, R., Johannsen, L., & Klauer, K. C. (2020). rtmpt: An R package for fitting response-time extended multinomial processing tree models. *Behavior Research Methods*, 52(3), 1313–1338. <https://doi.org/10.3758/s13428-019-01318-x>
- Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, 23(5), 1440–1465. <https://doi.org/10.3758/s13423-016-1025-6>
- Heck, D. W., & Erdfelder, E. (2020). Benefits of response time-extended multinomial processing tree models: A reply to Starns (2018). *Psychonomic Bulletin & Review*, 27(3), 571–580. <https://doi.org/10.3758/s13423-019-01663-0>
- Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2018). Generalized processing tree models: Jointly modeling discrete and continuous variables. *Psychometrika*, 83(4), 893–918. <https://doi.org/10.1007/s11336-018-9622-0>
- Juola, J. F., Caballero-Sanz, A., Muñoz-García, A. R., Botella, J., & Suero, M. (2019). Familiarity, recollection, and receiver-operating characteristic (ROC) curves in recognition memory. *Memory & Cognition*, 47(4), 855–876. <https://doi.org/10.3758/s13421-019-00922-8>
- Juola, J. F., Fischler, I., Wood, C. T., & Atkinson, R. C. (1971). Recognition time for information stored in long-term memory. *Perception & Psychophysics*, 10(1), 8–14. <https://doi.org/10.3758/BF03205757>
- Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, 104(2), 344–366. <https://doi.org/10.1037/0033-295X.104.2.344>
- Kang, M., Sinha, N., Michel, M. M., Khandelwal, R., Wang, B., Ahmed, M., & Glass, A. L. (2025). A continuous aggregated accumulation model of recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 51(4), 627–660. <https://doi.org/10.1037/xlm0001379>
- Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. <https://doi.org/10.1037/xlm0000016>
- Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal

- assumptions. *Psychological Review*, 122(3), 542–557. <https://doi.org/10.1037/a0039251>
- Kellen, D., & Klauer, K. C. (2020). Selecting amongst multinomial models: An apologia for normalized maximum likelihood. *Journal of Mathematical Psychology*, 97, Article 102367. <https://doi.org/10.1016/j.jmp.2020.102367>
- Kellen, D., Meyer-Grant, C. G., Singmann, H., & Klauer, K. C. (2025). Critical testing in recognition memory: Selective influence, single-item generalization, and the high-threshold hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 51(8), 1259–1280. <https://doi.org/10.1037/xlm0001434>
- Klauer, K. C., Hartmann, R., & Meyer-Grant, C. G. (2024). RT-MPTs: Process models for response-time distributions with diffusion-model kernels. *Journal of Mathematical Psychology*, 120–121, Article 102857. <https://doi.org/10.1016/j.jmp.2024.102857>
- Klauer, K. C., & Kellen, D. (2018). RT-MPTs: Process models for response-time distributions based on multinomial processing trees with applications to recognition memory. *Journal of Mathematical Psychology*, 82, 111–130. <https://doi.org/10.1016/j.jmp.2017.12.003>
- Klauer, K. C., & Voss, A. (2008). Effects of race on responses and response latencies in the weapon identification task: A test of six models. *Personality and Social Psychology Bulletin*, 34(8), 1124–1140. <https://doi.org/10.1177/0146167208318603>
- Langeheine, R., Pannekoek, J., & Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research*, 24(4), 492–516. <https://doi.org/10.1177/0049124196024004004>
- Laukenmann, R., Erdfelder, E., Heck, D. W., & Moshagen, M. (2023). Cognitive processes underlying the weapon identification task: A comparison of models accounting for both response frequencies and response times. *Social Cognition*, 41(2), 137–164. <https://doi.org/10.1521/soco.2023.41.2.137>
- Lin, J.-J., Chang, C.-H., & Pal, N. (2015). A revisit to contingency table and tests of independence: Bootstrap is preferred to chi-square approximations as well as Fisher's exact test. *Journal of Biopharmaceutical Statistics*, 25(3), 438–458. <https://doi.org/10.1080/10543406.2014.920851>
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.
- Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press. <https://doi.org/10.4324/9781410611147>
- Marewski, J. N., & Olsson, H. (2009). Beyond the null ritual: Formal modeling of psychological processes. *Zeitschrift für Psychologie / Journal of Psychology*, 217(1), 49–60. <https://doi.org/10.1027/0044-3409.217.1.49>
- Migo, E. M., Mayes, A. R., & Montaldi, D. (2012). Measuring recollection and familiarity: Improving the remember/know procedure. *Consciousness and Cognition*, 21(3), 1435–1455. <https://doi.org/10.1016/j.concog.2012.04.014>
- Murdock, B. B., & Dufty, P. O. (1972). Strength theory and recognition memory. *Journal of Experimental Psychology*, 94(3), 284–290. <https://doi.org/10.1037/h0032795>
- Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2017). Model evaluation and selection. In W. H. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology: Foundations and methodology* (pp. 552–598). Cambridge University Press. <https://doi.org/10.1017/9781139245913.011>
- Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, 50(2), 167–179. <https://doi.org/10.1016/j.jmp.2005.06.008>
- Osth, A. F., Bora, B., Dennis, S., & Heathcote, A. (2017). Diffusion vs. linear ballistic accumulation: Different models, different conclusions about the slope of the zROC in recognition memory. *Journal of Memory and Language*, 96, 36–61. <https://doi.org/10.1016/j.jml.2017.04.003>
- Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology*, 92, 101–126. <https://doi.org/10.1016/j.cogpsych.2016.11.007>
- Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive Psychology*, 104, 106–142. <https://doi.org/10.1016/j.cogpsych.2018.04.002>
- Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, 117(3), 864–901. <https://doi.org/10.1037/a0019737>
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition*, 21(1), 89–102. <https://doi.org/10.3758/BF03211168>
- Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, 116(1), 59–83. <https://doi.org/10.1037/a0014086>
- Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, 120(3), 697–719. <https://doi.org/10.1037/a0033152>
- Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer. <https://doi.org/10.1007/978-1-4612-4578-0>
- Rissanen, J. J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1), 40–47. <https://doi.org/10.1109/18.481776>
- Romero, M., Juola, J. F., Casadevante, C., & Hernández, J. M. (2022). Objective evaluation of goal orientation, time management and learning outcomes. *Anales de Psicología / Annals of Psychology*, 38(2), 383–394. <https://doi.org/10.6018/analesps.484851>
- Rotello, C. M. (2017). Signal detection theories of recognition memory. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (2nd ed., Vol. 2, pp. 201–225). Academic Press. <https://doi.org/10.1016/B978-0-12-809324-5.21044-4>
- Rotello, C. M., & Zeng, M. (2008). Analysis of RT distributions in the remember-know paradigm. *Psychonomic Bulletin & Review*, 15(4), 825–832. <https://doi.org/10.3758/PBR.15.4.825>
- Schmidt, O., Erdfelder, E., & Heck, D. W. (2025). How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychological Methods*, 30(4), 720–743. <https://doi.org/10.1037/met0000561>
- Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: Inequality, order and shape restrictions*. John Wiley & Sons. <https://doi.org/10.1002/9781118165614>
- Singmann, H., Heck, D. W., Barth, M., Erdfelder, E., Arnold, N. R., Aust, F., Calanchini, J., Gümüşdagli, F. E., Horn, S. S., Kellen, D., Klauer, K. C., Matzke, D., Meissner, F., Michalkiewicz, M., Schaper, M. L., Stahl, C., Kuhlmann, B. G., & Groß, J. (2024). Evaluating the robustness of parameter estimates in cognitive models: A meta-analytic review of multinomial processing tree models across the multiverse of estimation methods. *Psychological Bulletin*, 150(8), 965–1003. <https://doi.org/10.1037/bul0000434>
- Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, 45(2), 560–575. <https://doi.org/10.3758/s13428-012-0259-0>
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of*

- Experimental Psychology: General*, 117(1), 34–50. <https://doi.org/10.1037/0096-3445.117.1.34>
- Starns, J. J. (2021). High- and low-threshold models of the relationship between response time and confidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(4), 671–684. <https://doi.org/10.1037/xlm0000960>
- Starns, J. J., & Heathcote, A. (2024). Evidence accumulation and decision processes. In M. J. Kahana & A. D. Wagner (Eds.), *The Oxford handbook of human memory* (pp. 959–987). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190917982.013.33>
- Starns, J. J., & Ma, Q. (2018). Guessing versus misremembering in recognition: A comparison of continuous, two-high-threshold, and low-threshold models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(4), 527–539. <https://doi.org/10.1037/xlm0000461>
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1(1), 1–26. <https://doi.org/10.1111/1529-1006.001>
- Vickers, D., & Pietsch, A. (2001). Decision making and memory: A critique of Juslin and Olsson's (1997) sampling model of sensory discrimination. *Psychological Review*, 108(4), 789–804. <https://doi.org/10.1037/0033-295X.108.4.789>
- Volkman, J. (1934). The relation of the time of judgment to the certainty of judgment. *Psychological Bulletin*, 31(10), 672–673.
- Voormann, A., Rothe-Wulf, A., Meyer-Grant, C. G., & Klauer, K. C. (2025). Sometimes memory misleads: Variants of the error-speed effect strengthen the evidence for systematically misleading memory signals in recognition memory. *Psychonomic Bulletin & Review*, 32(1), 294–305. <https://doi.org/10.3758/s13423-024-02534-z>
- Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science*, 3(4), Article 150670. <https://doi.org/10.1098/rsos.150670>
- Wickens, T. D. (2001). *Elementary signal detection theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195092509.001.0001>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. <https://ggplot2.tidyverse.org>
- Wixted, J. T. (2009). Remember/know judgments in cognitive neuroscience: An illustration of the underrepresented point of view. *Learning & Memory*, 16(7), 406–412. <https://doi.org/10.1101/lm.1312809>
- Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, 117(4), 1025–1054. <https://doi.org/10.1037/a0020874>
- Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46(3), 441–517. <https://doi.org/10.1006/jmla.2002.2864>

Appendix A

Confidence Levels and Response Times–Extended Two-High-Threshold, Signal Detection Theory, and Atkinson–Juola Models

To ensure that the models account for their predictions, we established a set of parameter constraints that generate ordinal relationships between parameters (see Model Details section). In this work, we present simplified versions of the two-high-threshold (2HT), signal detection theory (SDT), and Atkinson–Juola (AJ) models to facilitate the interpretation of the results. However, the versions available in the repository, which are used to fit the data, incorporate ordinal relationships through shrinkage parameters (s). As a result, the repository models will differ slightly from the simplified versions presented in Figures A1–A3.

In the 2HT model, we established that $g_2 = g_3 \cdot s_{g2}$ and $g_1 = g_1 \cdot s_{g2} \cdot s_{g1}$ to ensure that $g_1 < g_2 < g_3$ (where subscripts 1, 2, and 3 refer to the target manipulation conditions of 35%, 50%, and 65%, respectively). Regarding the response time (RT) predictions, the 2HT model assumes that the probability of fast detection responses is greater than that of guessing responses. Therefore, we set $L_{gn} = L_{dn} \cdot sL_{gn}$ and $L_{go} = L_{do} \cdot sL_{go}$, as the model predicts $L_{do} > L_{go}$ and $L_{dn} > L_{gn}$. Similarly, the 2HT model assumes that the probability of high confidence levels (CLs) is greater following a successful detection than after guessing ($H_{do, b} > H_{go, b}$ and $H_{dn, b} > H_{gn, b}$), which leads

to the following constraints: $H_{gn, f} = H_{dn, f} \cdot sH_{gn, f}$; $H_{gn, s} = H_{dn, s} \cdot sH_{gn, s}$; $H_{go, f} = H_{do, f} \cdot sH_{go, f}$; $H_{go, s} = H_{do, s} \cdot sH_{go, s}$.

In the SDT model, RT parameters such as L_{miss} and L_{fa} are constrained using shrinkage parameters to ensure that hits are faster than miss responses, and correct rejections are faster than false alarms (FAs). Specifically, we establish $L_{miss} = L_{hit} \cdot sL_{miss}$ and $L_{fa} = L_{cr} \cdot sL_{fa}$. It is important to note that the decision and confidence criteria are already modeled as cumulative increments along the familiarity continuum, so the necessary constraints are inherently established within the model.

In the AJ model, the parameters for modeling RTs, such as L_s , L_{go} , and L_{gn} , are structured as follows: $L_s = L_f \cdot sL_s$, $L_{go} = L_f \cdot sL_{go}$, and $L_{gn} = L_f \cdot sL_{gn}$. This ensures that the AJ model maintains the prediction that extreme familiarity responses are faster than those based on search or guesses ($L_f > L_s$, $L_f > L_{go}$, and $L_f > L_{gn}$). Regarding CLs, search responses should exhibit the highest CLs, followed by extreme familiarity, and then guesses (i.e., $H_{s, b} > H_{f, b} > H_{g, b}$). To maintain this ordinal relationship, the following constraints were applied: $H_{f, f} = H_{s, f} \cdot sH_{f, f}$; $H_{f, s} = H_{s, s} \cdot sH_{f, s}$; $H_{g, f} = H_{s, f} \cdot sH_{g, f}$; $H_{g, s} = H_{s, s} \cdot sH_{g, s}$.

(Appendices continue)

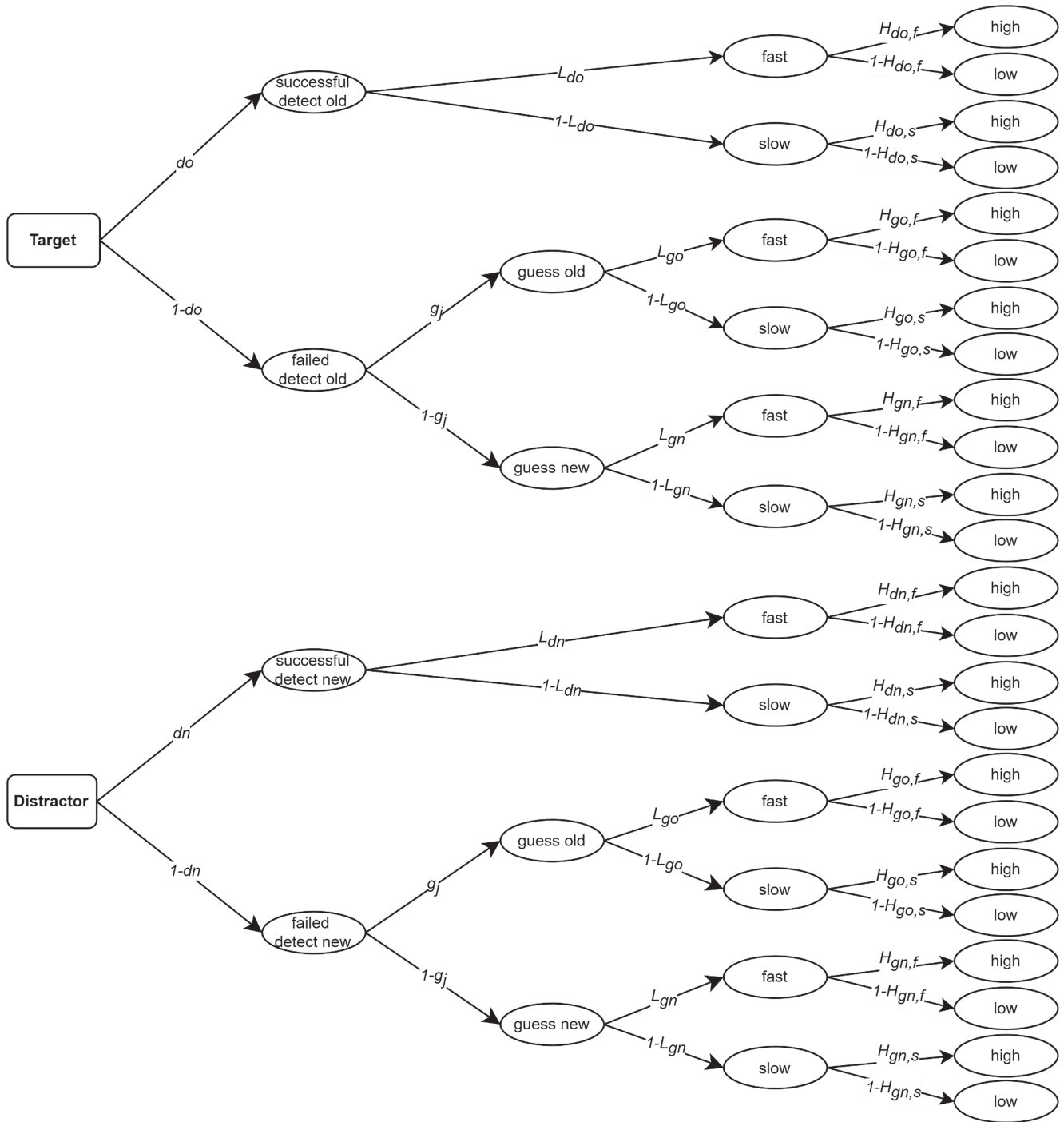
Figure A1*Confidence Level and Response Time Extended Two-High-Threshold Model**(Appendices continue)*

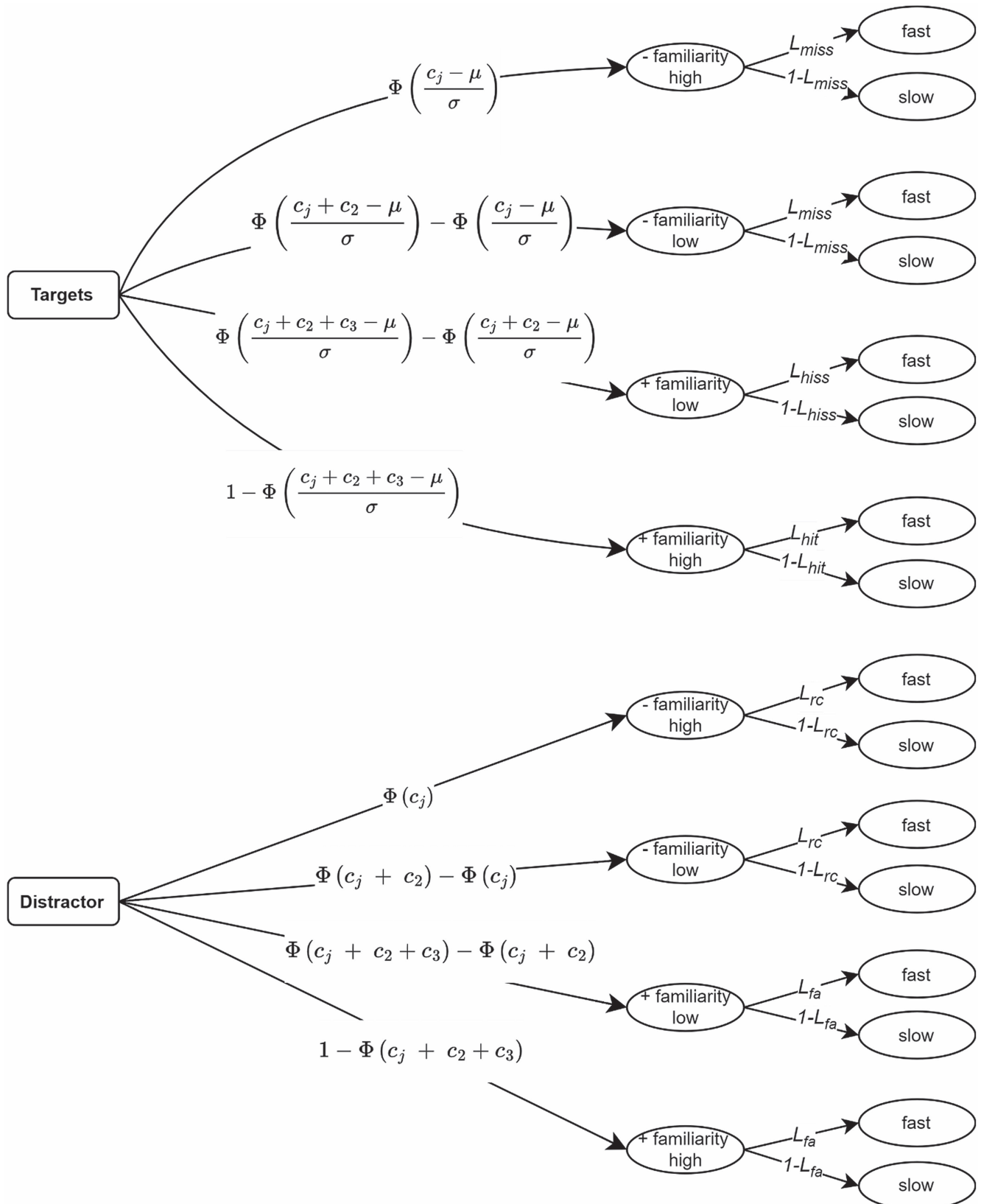
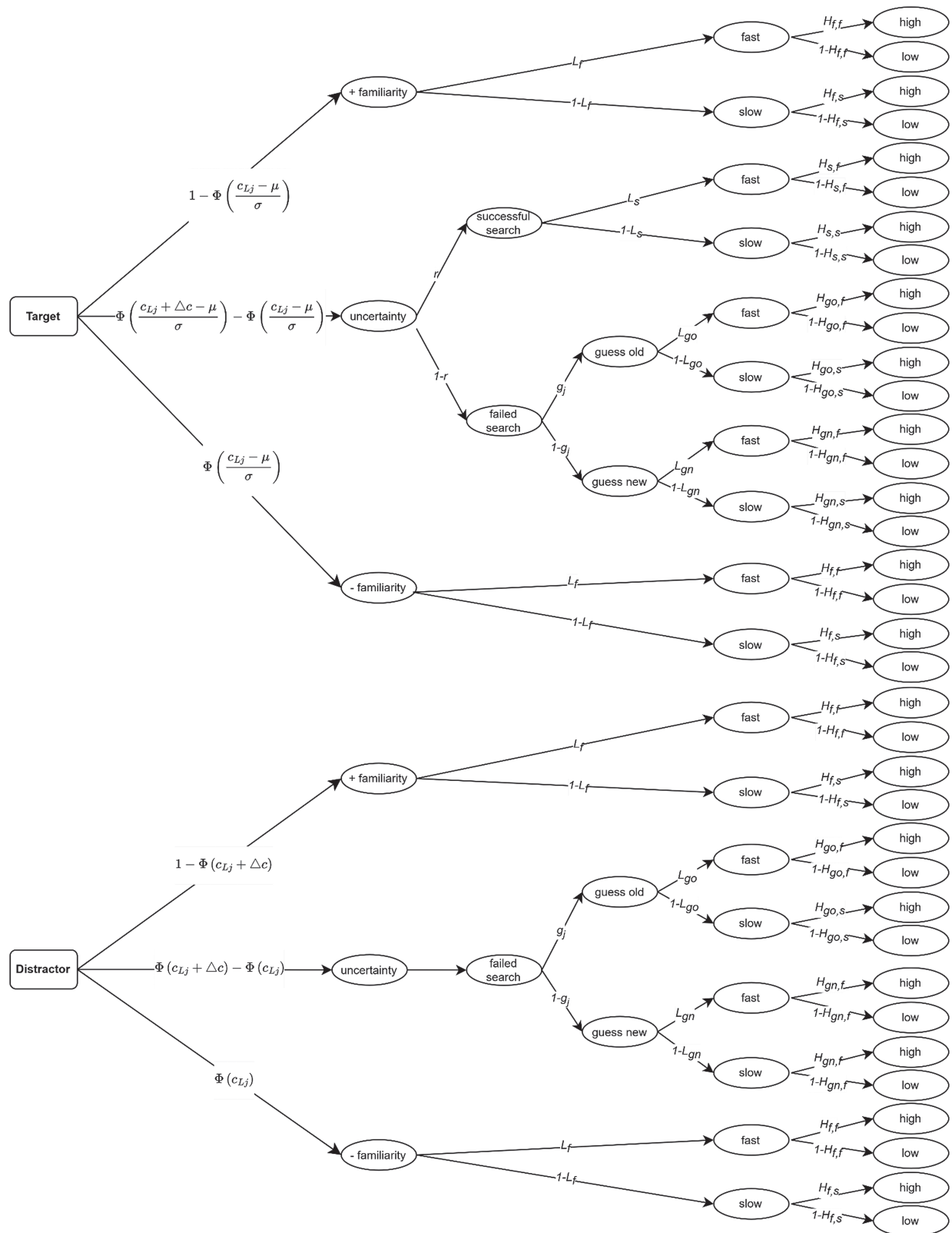
Figure A2*Confidence Level and Response Time Extended Signal Detection Theory Model**(Appendices continue)*

Figure A3*Hybrid Confidence Level and Response Time Extended Atkinson–Juola Model*

(Appendices continue)

Appendix B

Juola et al.'s (2019) Data Preprocessing

The original data from Juola et al. (2019) comprise five conditions of target manipulation, each with 100 trials, totaling 500 observations ($N = 500$). However, the most extreme conditions, those with 15% and 85% target proportions, resulted in sparse data for specific response categories. This sparsity creates significant challenges for estimating parameters associated with those under-represented response types.

Importantly, this issue is not a consequence of a small overall sample size ($N = 500$), but rather of the extremely low expected probabilities for certain categories. To illustrate this problem, consider the 2HT model. In the 15% condition, the number of old trials is only 15 (i.e., 500 trials divided by five conditions, multiplied by .15). For such a tree, the expected frequency of a “guessing old” response is computed by multiplying these 15 trials by the probability of taking the guessing-old path: $(1 - do) \times g_5 \times s_{g1} \times s_{g2} \times s_{g3} \times s_{g4}$. Based on the average parameter estimates in Table B1, this yields an expected frequency of just 1.67 observations. This low expected count highlights how extreme conditions can lead to estimation difficulties, even with a reasonable total sample size.

When the guessing branch is further divided into fast and slow responses and again into high, medium, or low confidence, the number of observations per subcategory becomes even smaller. For instance, a “guessing–high–confidence–fast” response might have an expected frequency as low as $1.67 \times L \times H$. These values are too low to reliably estimate parameters associated with guessing and CLs. Similar estimation problems occur in the SDT and AJ models, although in SDT, the effect of extreme probabilities is mediated by shifts in the decision criteria rather than the guessing process.

Incorporating other variables like RTs and CLs into the multinomial processing tree (MPT) model—as in the extended-MPT approach by Heck and Erdfelder (2016)—exacerbates the issue. Adding these variables increases the number of model branches and bins, which leads to more cells with few or zero observations. This, in turn, compromises the reliability of model estimation. As previously noted in the literature (Read & Cressie, 1988; Silvapulle & Sen, 2005), working with probabilities near the boundaries of the parameter space can introduce bias and inefficiency. Additionally, the G^2 goodness-of-fit test becomes unreliable when cell counts are low, as its asymptotic properties are no longer preserved (Langeheine et al., 1996; Lin et al., 2015). Although parametric bootstrapping techniques can mitigate this issue, it remains important to avoid

sparse data to preserve the power for testing predictions or model comparisons. Additionally, the authors of the extended-MPT approach (Heck & Erdfelder, 2016) have explicitly warned about the need to avoid binning strategies that produce RT bins with zero observations.

To address this, we considered several solutions. One possibility was to ignore RTs and CLs altogether, which would reduce the number of subcategories and improve model stability. However, we ruled this out because our main goal is to evaluate the AJ model's predictions, which requires the joint modeling of RTs, CLs, and accuracy. Another option was to collapse the extreme conditions (e.g., merging 15% and 35% trials), but this would mean combining data from distinct experimental manipulations, potentially obscuring parameter differences that should remain distinct. For example, collapsing trials where the guessing probability is 0.1 with those where it is 0.3 would undermine the validity of parameter interpretations and might lead to poor model fits. Ultimately, we chose to exclude the extreme 15% and 85% conditions. This approach, previously applied in Gutkin et al. (2024), offered the most robust and interpretable solution.

As shown in Table B2, the chi-square goodness-of-fit test was rejected in nearly all cases when using the five-condition data set. Specifically, the test was rejected for 97.9% of participants in the SDT model and for 100% of participants in the AJ and 2HT models. By contrast, Table B3 shows that the percentage of data sets for which the goodness-of-fit test was rejected decreases when only the three intermediate conditions are included: 65.96% for the 2HT model, 80.85% for SDT, 57.4% for AJ_{cj}, and 76.6% for AJ_{gj}.

It is important to emphasize that our decision to exclude extreme manipulation conditions was not based solely on the statistical significance of the goodness-of-fit tests. Nonparametric MPT-RT models require a sufficient number of observations in each response category to produce stable parameter estimates and support meaningful model comparisons (Chechile, 2009). The lack of power when comparing nested and non-nested models, as we do in the present study, is a clear limitation of this approach compared to parametric alternatives (Gutkin & Heck, 2026).

Parametric MPT-DC approaches (e.g., Hartmann et al., 2020; Heck et al., 2018) and hierarchical methods (Erdfelder et al., 2024; Singmann et al., 2024) offer promising alternatives, but

Table B1

Mean Parameter Estimates for the 2HT Model

2HT	dn	do	g_5	$H_{dn,f}$	$H_{dn,s}$	$H_{do,f}$	$H_{do,s}$	L_{dn}	L_{do}	s_{g1}	s_{g2}	s_{g3}	s_{g4}	$sH_{gn,f}$	$sH_{gn,s}$	$sH_{go,f}$	$sH_{go,s}$	sL_{gn}	sL_{go}
$K = 19$	.28	.48	.64	.53	.64	.51	.77	.75	.93	.63	.86	.79	.78	.91	.84	.91	.61	.57	.54

Note. 2HT = two-high-threshold.

(Appendices continue)

Table B2*Goodness-of-Fit Test and Akaike Information Criteria for Original Data (Five Relative Target Proportion Conditions)*

Id	2HT			SDT			AJ _{cj}			AJ _{gj}		
	$\chi^2(51)$	<i>p</i>	AIC	$\chi^2(57)$	<i>p</i>	AIC	$\chi^2(50)$	<i>p</i>	AIC	$\chi^2(50)$	<i>p</i>	AIC
1	139.09	<.001	177.09	145.67	<.001	171.67	138.83	<.001	178.83	140.93	<.001	180.93
2	189.44	<.001	227.44	188.68	<.001	214.68	184.45	<.001	224.45	167.27	<.001	207.27
3	117.97	<.001	155.97	116.45	<.001	142.45	113.07	<.001	153.07	118.81	<.001	158.81
4	107.15	<.001	145.15	113.41	<.001	139.41	106.85	<.001	146.85	112.42	<.001	152.42
5	216.07	<.001	254.07	181.83	<.001	207.83	170.28	<.001	210.28	263.78	<.001	303.78
6	205.99	<.001	243.99	157.24	<.001	183.24	108.84	<.001	148.84	267.05	<.001	307.05
7	111.82	<.001	149.82	87.21	.006	113.21	82.36	.003	122.36	118.25	<.001	158.25
8	163.86	<.001	201.86	152.32	<.001	178.32	139.52	<.001	179.52	141.14	<.001	181.14
9	189.06	<.001	227.06	164.17	<.001	190.17	155.11	<.001	195.11	215.00	<.001	255.00
10	229.90	<.001	267.90	220.33	<.001	246.33	217.68	<.001	257.68	232.53	<.001	272.53
11	182.89	<.001	220.89	139.98	<.001	165.98	142.26	<.001	182.26	257.58	<.001	297.58
12	180.55	<.001	218.55	155.17	<.001	181.17	141.65	<.001	181.65	338.22	<.001	378.22
13	146.12	<.001	184.12	175.96	<.001	201.96	121.54	<.001	161.54	218.52	<.001	258.52
14	156.46	<.001	194.46	135.57	<.001	161.57	129.09	<.001	169.09	155.51	<.001	195.51
15	152.58	<.001	190.58	133.15	<.001	159.15	137.88	<.001	177.88	169.03	<.001	209.03
16	227.52	<.001	265.52	183.46	<.001	209.46	171.11	<.001	211.11	245.19	<.001	285.19
17	192.94	<.001	230.94	165.89	<.001	191.89	170.51	<.001	210.51	228.49	<.001	268.49
18	134.78	<.001	172.78	130.52	<.001	156.52	113.51	<.001	153.51	141.31	<.001	181.31
19	154.02	<.001	192.02	106.67	<.001	132.67	88.12	<.001	128.12	173.64	<.001	213.64
20	134.43	<.001	172.43	133.09	<.001	159.09	125.86	<.001	165.86	126.46	<.001	166.46
21	155.25	<.001	193.25	124.70	<.001	150.70	112.15	<.001	152.15	169.71	<.001	209.71
22	250.81	<.001	288.81	241.93	<.001	267.93	223.40	<.001	263.40	276.51	<.001	316.51
23	260.67	<.001	298.67	257.12	<.001	283.12	249.51	<.001	289.51	272.05	<.001	312.05
24	165.74	<.001	203.74	158.06	<.001	184.06	146.15	<.001	186.15	162.98	<.001	202.98
25	127.45	<.001	165.45	120.31	<.001	146.31	127.05	<.001	167.05	174.80	<.001	214.80
26	236.79	<.001	274.79	242.04	<.001	268.04	218.04	<.001	258.04	282.52	<.001	322.52
27	155.30	<.001	193.30	158.60	<.001	184.60	144.11	<.001	184.11	180.74	<.001	220.74
28	170.98	<.001	208.98	201.73	<.001	227.73	167.36	<.001	207.36	244.07	<.001	284.07
29	134.44	<.001	172.44	114.31	<.001	140.31	110.39	<.001	150.39	174.12	<.001	214.12
30	172.71	<.001	210.71	151.82	<.001	177.82	155.16	<.001	195.16	179.82	<.001	219.82
31	123.63	<.001	161.63	148.64	<.001	174.64	125.96	<.001	165.96	330.15	<.001	370.15
32	123.84	<.001	161.84	122.70	<.001	148.70	114.04	<.001	154.04	126.49	<.001	166.49
33	191.40	<.001	229.40	168.14	<.001	194.14	178.27	<.001	218.27	211.55	<.001	251.55
34	75.86	.014	113.86	63.00	.272	89.00	72.73	.020	112.73	91.38	<.001	131.38
35	240.93	<.001	278.93	219.74	<.001	245.74	219.15	<.001	259.15	242.19	<.001	282.19
36	229.23	<.001	267.23	235.99	<.001	261.99	212.84	<.001	252.84	228.29	<.001	268.29
37	156.03	<.001	194.03	129.77	<.001	155.77	103.71	<.001	143.71	177.14	<.001	217.14
38	132.77	<.001	170.77	129.33	<.001	155.33	114.82	<.001	154.82	177.22	<.001	217.22
39	245.69	<.001	283.69	219.86	<.001	245.86	202.28	<.001	242.28	277.30	<.001	317.30
40	156.19	<.001	194.19	136.22	<.001	162.22	134.36	<.001	174.36	256.99	<.001	296.99
41	155.10	<.001	193.10	156.96	<.001	182.96	147.30	<.001	187.30	190.56	<.001	230.56
42	192.86	<.001	230.86	170.17	<.001	196.17	162.93	<.001	202.93	227.70	<.001	267.70
43	171.97	<.001	209.97	188.51	<.001	214.51	176.21	<.001	216.21	233.51	<.001	273.51
44	157.45	<.001	195.45	126.85	<.001	152.85	121.72	<.001	161.72	168.65	<.001	208.65
45	103.62	<.001	141.62	103.42	<.001	129.42	95.87	<.001	135.87	145.08	<.001	185.08
46	256.24	<.001	294.24	205.80	<.001	231.80	172.49	<.001	212.49	249.83	<.001	289.83
47	358.19	<.001	396.19	345.80	<.001	371.80	304.78	<.001	344.78	402.73	<.001	442.73

Note. The estimated goodness-of-fit G^2 values, the p value of the goodness-of-fit test, and AIC value for the 2HT, SDT, AJ_{cj}, and AJ_{gj}. 2HT = two-high-threshold; SDT = signal detection theory; AJ = Atkinson–Juola; AIC = Akaike information criterion.

currently available tools do not support fitting hybrid models such as SDT or AJ models that operate beyond the [0, 1] parameter space.

To preserve model interpretability and ensure the validity of our individual-level comparisons, we excluded the extreme conditions from the final analysis. Nonetheless, in order to increase the transparency of this decision, we conducted a sensitivity test, evaluating the results without excluding these conditions. As

shown in [Tables B2 and B3](#), model fit indices for both the five-condition and three-condition data sets are reported, and scripts for obtaining results with five conditions are available in the additional online material: <https://osf.io/rbxw7/>.

In the additional online material (<https://osf.io/rbxw7/>), one can find a script that fits both the 15% and 85% target proportion conditions. This script also shows the change in the individual model fit for each situation.

(Appendices continue)

Table B3*Goodness-of-Fit Test and Akaike Information Criteria for Final Data (Three Relative Target Proportion Conditions)*

Id	2HT			SDT			AJ _{cj}			AJ _{gj}		
	$\chi^2(25)$	p	AIC	$\chi^2(31)$	p	AIC	$\chi^2(24)$	p	AIC	$\chi^2(24)$	p	AIC
1	29.01	.263	63.01	81.21	<.001	103.21	22.50	.550	58.50	35.24	.065	71.24
2	44.92	.009	78.92	102.33	<.001	124.33	39.83	.022	75.83	28.69	.232	64.69
3	33.41	.121	67.41	56.32	.004	78.32	33.38	.096	69.38	34.96	.069	70.96
4	27.65	.324	61.65	39.90	.131	61.90	26.90	.309	62.90	29.68	.196	65.68
5	81.44	<.001	115.44	76.30	<.001	98.30	60.60	<.001	96.60	92.40	<.001	128.40
6	50.19	.002	84.19	44.39	.056	66.39	37.00	.044	73.00	74.08	<.001	110.08
7	35.45	.080	69.45	29.41	.548	51.41	17.90	.808	53.90	35.87	.056	71.87
8	33.99	.108	67.99	45.58	.044	67.58	23.92	.466	59.92	39.86	.022	75.86
9	53.49	<.001	87.49	110.10	<.001	132.10	40.51	.019	76.51	64.17	<.001	100.17
10	26.46	.383	60.46	143.73	<.001	165.73	31.75	.133	67.75	39.88	.022	75.88
11	102.59	<.001	136.59	108.11	<.001	130.11	82.81	<.001	118.81	108.87	<.001	144.87
12	51.39	.001	85.39	96.48	<.001	118.48	37.20	.042	73.20	169.29	<.001	205.29
13	63.46	<.001	97.46	72.36	<.001	94.36	37.85	.036	73.85	78.34	<.001	114.34
14	57.09	<.001	91.09	71.01	<.001	93.01	28.92	.223	64.92	58.85	<.001	94.85
15	30.44	.208	64.44	38.59	.164	60.59	33.76	.089	69.76	33.16	.101	69.16
16	80.48	<.001	114.48	98.68	<.001	120.68	55.21	<.001	91.21	77.47	<.001	113.47
17	86.63	<.001	120.63	92.35	<.001	114.35	51.13	.001	87.13	88.60	<.001	124.60
18	45.69	.007	79.69	59.97	.001	81.97	37.37	.040	73.37	49.09	.002	85.09
19	35.42	.081	69.42	38.45	.168	60.45	28.22	.251	64.22	47.93	.003	83.93
20	44.64	.009	78.64	54.15	.006	76.15	34.50	.076	70.50	42.45	.011	78.45
21	32.19	.153	66.19	54.92	.005	76.92	35.35	.063	71.35	40.58	.018	76.58
22	126.18	<.001	160.18	173.66	<.001	195.66	115.12	<.001	151.12	129.84	<.001	165.84
23	58.90	<.001	92.90	82.38	<.001	104.38	44.79	.006	80.79	55.91	<.001	91.91
24	34.08	.106	68.08	83.28	<.001	105.28	26.63	.322	62.63	34.88	.070	70.88
25	80.63	<.001	114.63	85.14	<.001	107.14	71.47	<.001	107.47	95.68	<.001	131.68
26	80.91	<.001	114.91	114.51	<.001	136.51	72.17	<.001	108.17	82.78	<.001	118.78
27	45.61	.007	79.61	102.77	<.001	124.77	44.18	.007	80.18	55.62	<.001	91.62
28	92.83	<.001	126.83	114.35	<.001	136.35	43.08	.010	79.08	116.75	<.001	152.75
29	38.47	.042	72.47	72.36	<.001	94.36	28.80	.228	64.80	34.61	.074	70.61
30	33.11	.128	67.11	54.43	.006	76.43	17.89	.808	53.89	39.17	.026	75.17
31	79.06	<.001	113.06	87.31	<.001	109.31	37.78	.037	73.78	116.13	<.001	152.13
32	43.69	.012	77.69	62.44	<.001	84.44	38.54	.031	74.54	42.59	.011	78.59
33	25.53	.433	59.53	33.26	.358	55.26	16.08	.885	52.08	30.12	.181	66.12
34	29.13	.259	63.13	36.44	.230	58.44	23.05	.517	59.05	34.21	.081	70.21
35	59.22	<.001	93.22	68.95	<.001	90.95	52.87	<.001	88.87	59.84	<.001	95.84
36	30.74	.198	64.74	96.22	<.001	118.22	27.29	.291	63.29	35.65	.059	71.65
37	58.02	<.001	92.02	51.19	.013	73.19	35.45	.062	71.45	69.01	<.001	105.01
38	44.83	.009	78.83	53.53	.007	75.53	38.57	.030	74.57	48.62	.002	84.62
39	61.13	<.001	95.13	74.04	<.001	96.04	54.92	<.001	90.92	63.46	<.001	99.46
40	38.73	.039	72.73	39.06	.152	61.06	11.13	.988	47.13	86.78	<.001	122.78
41	34.97	.089	68.97	87.75	<.001	109.75	41.51	.015	77.51	43.90	.008	79.90
42	58.36	<.001	92.36	75.21	<.001	97.21	39.67	.023	75.67	60.33	<.001	96.33
43	49.76	.002	83.76	59.06	.002	81.06	41.29	.015	77.29	56.24	<.001	92.24
44	47.80	.004	81.80	87.89	<.001	109.89	43.85	.008	79.85	50.78	.001	86.78
45	35.19	.085	69.19	44.49	.055	66.49	31.63	.136	67.63	39.09	.027	75.09
46	62.82	<.001	96.82	98.39	<.001	120.39	40.79	.018	76.79	70.72	<.001	106.72
47	130.60	<.001	164.60	166.70	<.001	188.70	121.82	<.001	157.82	128.85	<.001	164.85

Note. The estimated goodness-of-fit G^2 values, the p value of the goodness-of-fit test, and AIC value for the 2HT, SDT, AJ_{cj}, and AJ_{gj}. 2HT = two-high-threshold; SDT = signal detection theory; AJ = Atkinson–Juola; AIC = Akaike information criterion.

(Appendices continue)

Appendix C

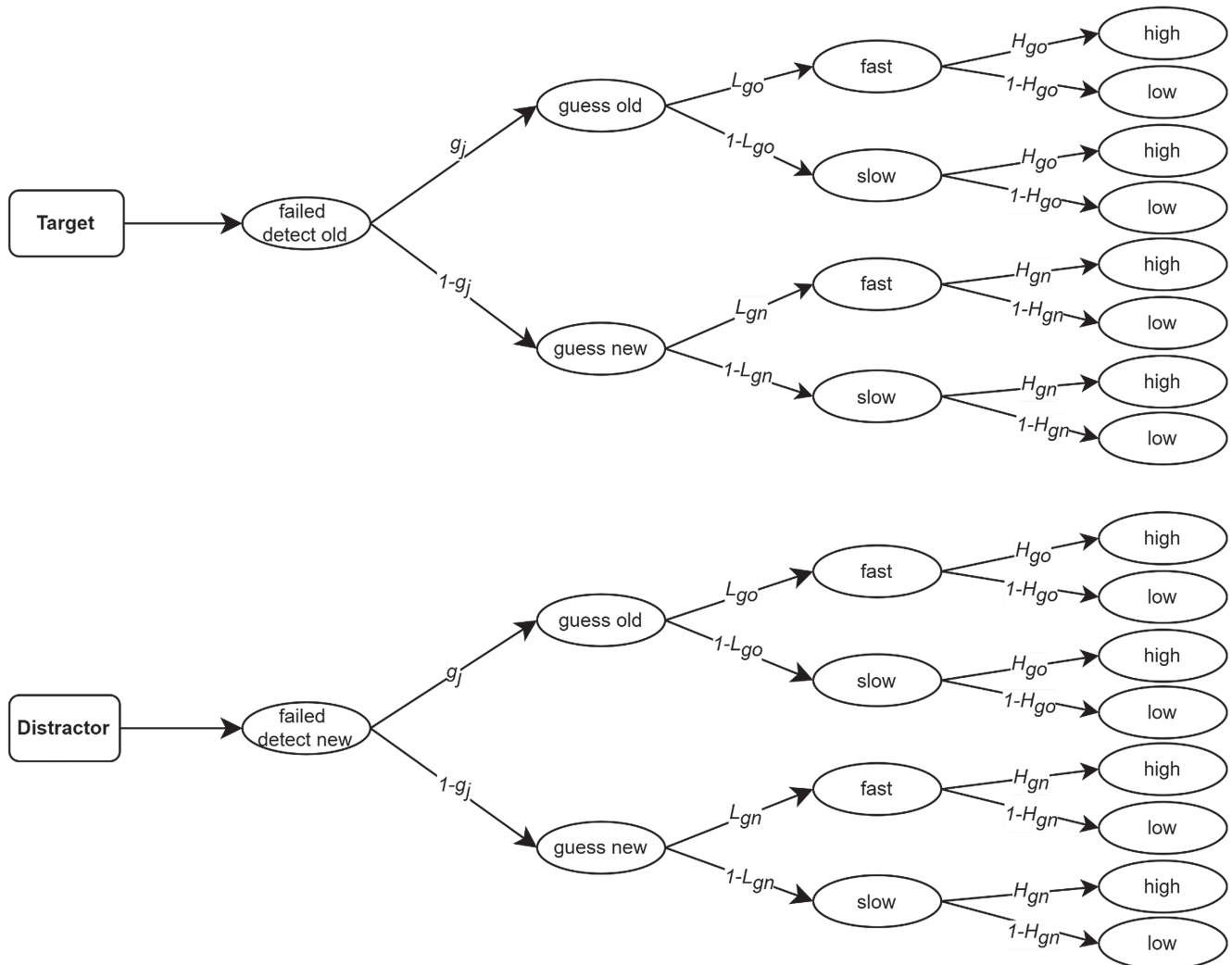
Comparison of Model Selection Criteria Through Monte Carlo Simulations

To evaluate the performance of different criteria for model selection, we conducted a Monte Carlo simulation using the 2HT, AJ, and SDT models for generating data, along with two “outside” models referred to as 2HT₀ and SDT₀ (see Figures C1 and C2).

The 2HT₀ model (see Figure C1) is a threshold model designed for no-learning situations. It is identical to the 2HT model, except that the parameters d_o and d_n are set to zero, implying that all responses are guesses. RTs are modeled using two L parameters: L_o and L_n . CLs are modeled with two H parameters: H_o and H_n . These parameters differentiate only between guess “old” and “new” responses, with subscripts o and n representing “old” and “new,” respectively.

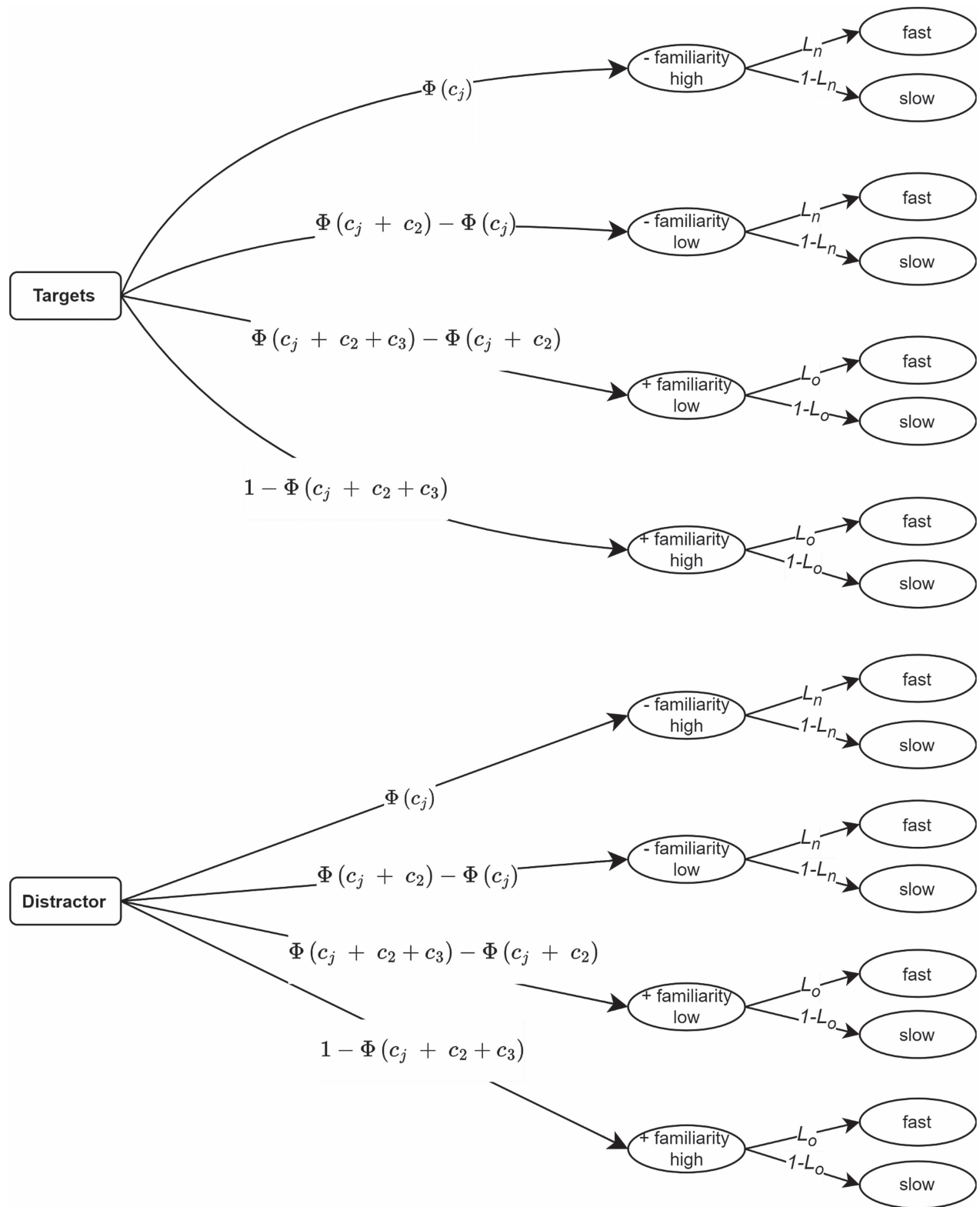
The SDT₀ model is a variant of the classic SDT for situations where no learning occurs. The familiarity for learning is assumed to increase, but if there is no learning, it should maintain its baseline familiarity. Thus, the SDT₀ model assumes that both distributions coincide and will be simplified by assuming that the mean of both familiarity distributions follows a normal probability density function where μ is set to 0, and its standard deviation, σ , is set to 1. As in the SDT model, three decision criteria along the familiarity continuum separate probability areas according to CLs and whether the stimuli are classified as “new” or “old.” In the SDT₀ model, RTs are modeled with two parameters, L : L_o for “old” responses and L_n for “new” responses (see Figure C2).

Figure C1

General Model, Subcase 2HT₀

Note. 2HT = two-high-threshold.

(Appendices continue)

Figure C2*General Model, Subcase SDT_0* 

Note. SDT = signal detection theory.

(Appendices continue)

It is important to note that we anticipated that, in the absence of learning (i.e., when data are generated by the SDT_0 or $2HT_0$), the 2HT, SDT, and AJ models should not provide good fits to the data. However, by forcing the model selection approach to choose between one of the non-data-generating models (AJ, SDT, or 2HT), we aimed to determine which model was flexible enough to adapt to this unusual scenario. Additionally, we fit the SDT_0 and $2HT_0$ models to verify whether, in cases where they generated the data, they would be selected as the best models instead of AJ, 2HT, and SDT.

Model Recovery Simulation

All simulation conditions were designed to closely resemble the experimental data analyzed in Juola et al. (2019) in terms of design, variables, number of trials, experimental manipulations, and parameter values. For each data-generating model, we generated 2,000 data sets across three target manipulation conditions, with 100 observations per relative target frequency condition ($N = 300$). The parameter values used were based on the estimated parameters from the Juola et al. (2019) experiment. Below, we show the chosen parameter for each data-generating model (see Table C1).

Note that the parameters starting with s are shrink factor parameters used to establish an ordinal relationship between parameters according to the predictions made by each model. Check Appendix B for further details.

Performance Measures

After simulating a total of 10,000 data sets, we fitted the SDT, 2HT, and AJ models to each data set and computed the rejection rate (RRate) of the goodness-of-fit test for each model. The goodness-of-fit tests were performed using the *MPTinR* package, which uses the log-likelihood ratio test statistic G^2 . The null hypothesis is rejected if the p value is smaller than $\alpha = .05$. Using Monte Carlo simulation, the RRate (Equation C1) was calculated as the relative frequency of significant test results:

$$\widehat{\text{RRate}} = \frac{\sum_{i=1}^R 1(p_i < 0.05)}{R}. \quad (\text{C1})$$

This process involves counting the number of iterations where the p value is smaller than α and dividing by the number of simulations, $R = 2,000$. Ideally, when the data-generating model matches the fitted model, the RRate should align with the Type I error level, typically set at .05.

We also calculated the frequency and proportion of selection for the SDT, 2HT, and AJ models according to different model selection criteria. The first criterion uses a measure called normalized maximum likelihood (NML). NML penalizes the functional flexibility of models based on the minimum description length framework (Myung et al., 2006; Rissanen, 1996), penalizing the model's ability to fit any data set. The penalty was computed using the *NMLmulti* package in R (Kellen & Klauer, 2020). This penalty is added to the model's estimated log-likelihood (LL_M), resulting in what we refer to as LL_{NML} —a measure of badness of fit that balances model flexibility with its fit to specific data. We also calculated the AIC and Bayesian information criterion (BIC), which were obtained directly using

the *MPTinR* package (Singmann & Kellen, 2013) and are calculated (Equations C2 and C3) as follows:

$$\text{AIC} = G^2 + 2K, \quad (\text{C2})$$

$$\text{BIC} = G^2 + \ln(N) \cdot K, \quad (\text{C3})$$

where G^2 measures the model's goodness-of-fit, K is the number of estimated parameters, and N is the sample size. As in Singmann and Kellen (2013), the AIC and BIC formulas instead of using LL_M as the first term, use G^2 , where $G^2 = 2 \cdot (LL_S - LL_M)$, and LL_S is the log-likelihood of a saturated model that fully explains the data. As a result, AIC and BIC can be understood as differences between the tested model and the saturated model, which would more precisely be noted as ΔAIC and ΔBIC , respectively (Burnham & Anderson, 2002). However, we use the terms AIC and BIC for simplicity.^{C1} Finally, we computed a measure called AIC_{NML} , calculated like AIC but using LL_{NML} , such that,

$$G^2 = 2 \cdot (LL_S - LL_{NML}). \quad (\text{C4})$$

Therefore, AIC_{NML} accounts for model fit, the number of parameters, and the functional flexibility of the model. The R and C++ scripts used for the simulations and NML calculations are available in the online repository at <https://osf.io/rbxw7/>.

Simulation Results

Referring to Table C2, in the RRate column, we observe the performance of the models in terms of their overall fit to the data. As expected, the cases where the data-generating and fitted models coincide (bold rows) show low RRates, though slightly higher than the expected .05. Notably the AJ model also shows a low RRate to data generated by the 2HT model (.08) and a relatively good fit to data generated by the SDT model (.2). In contrast, when fitting the SDT model, we observe very high RRates for data generated by the 2HT and AJ models (0.83 and 1, respectively). For the 2HT model, the goodness-of-fit test is rejected for about half of the data sets generated by the SDT and AJ models (.53 and .45, respectively). For data generated by the outside models, either SDT_0 or $2HT_0$, the RRate tends to be much higher than .05, particularly when fitting the 2HT model.

Examining the performance of the different model selection criteria, we observe that the use of LL_{NML} results in a relatively low recovery rate of the true data-generating model. Notably, the NML penalty is highest for the SDT model, followed by the AJ model, and lowest for the 2HT model. Overall, the AJ model is slightly more frequently selected as the best model across all simulation conditions. This overselection of the AJ model is evident in the underlined values, where we observe a proportion of .48 for data generated by the 2HT model and .43 for data generated by the SDT model. The same trend is seen for data generated by the outside models, with proportions of .62 and .64 for SDT_0 and $2HT_0$, respectively. In all cases, there is still no clear preference for any model, as the proportions remain far from 1, even when the AJ model both generates and fits the data (.55).

^{C1} The results using weighted AIC and BIC are very similar to those reported here; their calculation is available in the online repository.

Table C1*Parameters for Data-Generating Models*

2HT	dn	do	g_3	$H_{dn, f}$	$H_{dn, s}$	$H_{do, f}$	$H_{dn, s}$	L_{dn}	L_{do}	s_{g1}	s_{g2}	$sH_{gn, f}$	$sH_{gn, s}$	$sH_{go, f}$	$sH_{go, s}$	sL_{gn}	sL_{go}			
$K = 17$	.3	.5	.4	.6	.6	.7	.5	.7	.9	.7	.8	.8	.6	.8	.6	.6	.5			
AJ	Δc	c_{L1}	$s_{c_{L2}}$	$s_{c_{L3}}$	μ	g_3	$H_{s, f}$	$H_{s, s}$	L_f	r	s_{g1}	s_{g2}	$sH_{f, f}$	$sH_{f, s}$	$sH_{g, f}$	$sH_{g, s}$	sL_{gn}	sL_{go}	sL_s	σ
$K = 20$	1.5	-0.7	-0.3	-0.5	1.5	.4	.9	.7	.8	.6	.8	.9	.9	.8	.6	.5	.5	.6	.7	1.7
SDT	c_1	c_2	c_3	Δc_2	Δc_3	μ	L_{hit}	L_{cr}	sL_{fa}	sL_{miss}	σ									
$K = 11$	-0.7	-0.3	-0.5	1	0.5	1.5	.6	.5	.8	.7	1.5									
SDT ₀	c_1	c_2	c_3	Δc_2	Δc_3	H_n	H_o													
$K = 7$	-0.7	-0.2	-0.4	0.4	0.5	.4	.5													
2HT ₀	g_3	s_{g1}	s_{g2}	H_n	H_o	L_n	L_o													
$K = 7$	.5	.8	.9	.4	.5	.6	.7													

Note. 2HT = two-high-threshold; AJ = Atkinson–Juola; SDT = signal detection theory.

For the AIC_{NML} (which represents ΔAIC_{NML} relative to a saturated model), there is a noticeable improvement in the recovery rate of the true model, with recovery rates of .72 for data generated by 2HT, .63 for AJ, and .54 for SDT. The recovery rate shows an overall increase, but only 54% of data generated by the SDT model is correctly classified. For data generated by the outside models (i.e., SDT₀ and 2HT₀), the SDT or AJ models are the most selected models, while the 2HT model shows a lower frequency of selection.

In contrast, the traditional AIC and BIC criteria exhibit differing behaviors. AIC displays a relatively high recovery rate for the SDT

and AJ models, though not for 2HT. Even in the worst case, the recovery rate is around 60%, outperforming the other criteria. When the true model is not correctly selected, the AJ model tends to absorb these data sets. Data sets generated by the AJ model but misclassified tend to be fitted by the 2HT model. For SDT-generated data, the recovery rate is nearly 90%. For data generated by outside models, AIC tends to select SDT, even when it is not the true data-generating model.

With BIC, there is a clear overselection to the SDT model across all conditions, demonstrating a preference for the most parsimonious model.

Table C2*Model Selection for Simulated Data*

Gen model	Fit model	RRate	LL _{NML}	LL _{NML} /R	AIC _{NML}	AIC _{NML} /R	AIC	AIC/R	BIC	BIC/R
2HT	2HT	.09	918	.46	1,444	.72	1,155	.58	571	.29
	AJ	.08	965	.48	542	.27	723	.36	8	0
	SDT	.83	117	.06	14	.01	122	.06	1,421	.71
AJ	2HT	.45	859	.43	747	.37	564	.28	1,381	.69
	AJ	.09	1,101	.55	1,253	.63	1,436	.72	498	.25
	SDT	1	40	.02	0	0	0	0	121	.06
SDT	2HT	.53	648	.32	438	.22	96	.05	0	0
	AJ	.2	861	.43	482	.24	167	.08	0	0
	SDT	.06	491	.25	1,080	.54	1,737	.87	2,000	1
SDT ₀	2HT	.92	295	.15	125	.06	30	.02	1	0
	AJ	.3	1,236	.62	889	.44	383	.19	0	0
	SDT	.21	469	.23	986	.49	1,587	.79	1,999	1
2HT ₀	2HT	.88	369	.18	223	.11	60	.03	5	0
	AJ	.26	1,287	.64	1,080	.54	602	.3	0	0
	SDT	.34	344	.17	697	.35	1,338	.67	1,995	1

Note. Summary of results for simulated data: goodness-of-fit, frequency, and proportions for model selection. In the first column (gen model), the data-generating models are listed, and in the second column (fit model), the models fitted to those data are shown. The third column (RRate) provides the rejection rates from the goodness-of-fit tests. Columns four and five (LL_{NML} and LL_{NML}/R) show the frequency and proportion of selection for the best model according to the NML, where LL_{NML}/R is the proportion obtained by dividing the frequency by the number of replicates, R . Columns six and seven (AIC_{NML} and AIC_{NML}/R) provide the AIC calculated with LL_{NML}. Columns eight and nine (AIC and AIC/R) reflect the selection based on the classic AIC, and the last two columns (BIC and BIC/R) show the model selection results using BIC, frequencies, and proportions, respectively. Bold cells highlight situations where the fitted model matches the data-generating model. The italic values show cases where the model most frequently selected as the best one does not coincide with the data-generating model (overselection situations). AIC = Akaike information criterion; BIC = Bayesian information criterion; 2HT = two-high-threshold; AJ = Atkinson–Juola; SDT = signal detection theory; NML = normalized maximum likelihood; LL = log-likelihood.

(Appendices continue)

Discussion

We compared the performance of different model selection criteria in recovering data-generating models through Monte Carlo simulations. Our results show that AIC demonstrates the best overall performance. On its own, LL_{NML} shows a very low recovery rate across all simulation conditions. However, when incorporating the penalization into AIC calculations with AIC_{NML} , we detect some improvement. This improvement might be due to evaluating the model fit component LL_{NML} against a saturated model (see Equation C4) and penalizing the number of parameters.

However, while AIC_{NML} improves the recovery of the less flexible (Myung et al., 2017) 2HT model, it significantly worsens the recovery of the SDT model, which is heavily penalized despite having fewer parameters than its competitors, 2HT and AJ. It is important to note that in the SDT model, the accuracy and confidence parameters combine to characterize different normal probability functions, making the model much more complex than the 2HT model. Additionally, the SDT model is penalized more than the AJ model, likely due to how it models CLs, creating criteria along the familiarity continuum. These criteria introduce more parameters that interact with the other ones to define the normal function, further increasing the model's complexity. Focusing on the classic AIC, we observe that this criterion, while not accounting for model flexibility, does not penalize the SDT model as much, improving its recovery when it is the true model, and without overselection to the

AJ or 2HT data. Nevertheless, this results in better fits of the SDT for data generated with the SDT_0 and $2HT_0$ models. In cases where no model fits well (as indicated by high RRates), the SDT model tends to be selected by default, either because it is the most parsimonious in terms of parameter count or because its flexibility allows it to adapt to a wider range of data combinations, such as those generated by the SDT_0 and $2HT_0$ models. We also fitted the $2HT_0$ and SDT_0 models to assess whether the SDT overselection trend persisted and found that, for data generated by these models, the correct models (which assume no learning) were generally selected, rather than SDT, 2HT, or AJ. Thus, if we fit models with correct assumptions about data predictions (in this case, assuming no learning), AIC, despite not accounting for the supposed flexibility of the SDT model, does not seem to indicate overselection of these models. The AIC, for data generated by the 2HT model, shows a lower recovery rate (.58) than for the AJ (.72) or SDT (.87) data. This may be due to a mimicry between the AJ and 2HT models (a similar, though lesser, issue arises with AJ, where some data sets are misclassified as 2HT). BIC, on the other hand, clearly favors the model with fewer parameters, leading to overselection of the SDT model, demonstrating that it is not a suitable criterion in this context.

In summary, AIC proves to be the best model selection criterion overall due to its higher recovery rates. However, we must note that the recovery rate remains relatively low for the 2HT model (slightly less than 60%), with the majority of the remaining percentage absorbed by AJ. A similar phenomenon occurs with misclassified AJ data sets, which tend to be absorbed by 2HT.

(Appendices continue)

Appendix D

Testing Atkinson–Juola Model Predictions Through Nested Model Comparisons

Table D1

Nested Model Comparison: AJ Hypothesis Testing With Respect to Criteria and Guessing Probability

Id	AJ _{cj}		AJ _{gj}	
	$G^2(3)$	p	$G^2(3)$	p
1	12.74	<.001	0	.999
2	4.04	.044	15.18	<.001
3	1.58	.209	0	.999
4	3.65	.056	0.86	.354
5	33.39	<.001	1.58	.209
6	37.07	<.001	0	.999
7	17.49	<.001	0	.999
8	15.95	<.001	0	.999
9	23.67	<.001	0.02	.888
10	8.12	.004	0	.999
11	29.14	<.001	3.08	.079
12	127.54	<.001	0	.999
13	38.94	<.001	0	.999
14	30.08	<.001	0.15	.699
15	1.68	.195	2.28	.131
16	25.47	<.001	3.21	.073
17	38.4	<.001	0.93	.335
18	11.88	.001	0.15	.699
19	19.71	<.001	0	.999
20	10	.002	2.04	.153
21	5.29	.021	0.05	.823
22	22.85	<.001	8.13	.004
23	13.47	<.001	2.35	.125
24	9.85	.002	1.6	.206
25	22.76	<.001	0	.999
26	12.08	.001	1.46	.227
27	11.47	.001	0.03	.862
28	73.94	<.001	0.27	.603
29	7.31	.007	1.5	.221
30	21.28	<.001	0	.999
31	77.77	<.001	0	.999
32	4.61	.032	0.56	.454
33	14.06	<.001	0.03	.862
34	11.15	.001	0	.999
35	12.05	.001	5.08	.024
36	8.43	.004	0.07	.791
37	33.55	<.001	0	.999
38	9.18	.002	0	.999
39	10.17	.001	1.62	.203
40	73.75	<.001	0	.999
41	2.39	.122	0	.999
42	20.9	<.001	0.25	.617
43	21.71	<.001	6.77	.009
44	7.24	.007	0.32	.572
45	7.79	.005	0.33	.566
46	30.28	<.001	0.35	.554
47	7.38	.007	0.36	.549

Note. Log-likelihood ratio tests compare two nested models under different assumptions about the relationship between target frequency and the parameters c_{Lj} (for the AJ_{cj} model) or g_j (for the AJ_{gj} model). On the left, the most complex AJ_{cj} model assumes an inverse linear relationship between c_{Lj} and target frequency, and it is tested against a nested null model that assumes uniformity among the c_{Lj} parameters. On the right, the most complex AJ_{gj} model posits a direct linear relationship between g_j and target frequency, and it is compared to a null model that assumes no variation in g_j across experimental conditions. The G^2 columns display the log-likelihood ratio statistics for the nested model comparison, with the corresponding degrees of freedom ($df = 3$) indicated in parentheses, while the p columns list the associated p values. Significant results ($p < .05$) indicate rejection of the null hypothesis, supporting evidence in favor of the model's predicted relationships. AJ = Atkinson–Juola.

(Appendices continue)

Table D2*Nested Model Comparison: AJ Hypothesis Testing for Response Times*

Id	AJ _{cj}		AJ _{gj}	
	$G^2(3)$	p	$G^2(3)$	p
1	15.23	<.001	11.19	.001
2	68.6	<.001	70.62	<.001
3	19.86	<.001	19.83	<.001
4	5.51	.019	3.94	.047
5	9.03	.003	4.57	.033
6	6.13	.013	3.77	.052
7	6.18	.013	1.47	.225
8	16.8	<.001	13.78	<.001
9	21.17	<.001	19.3	<.001
10	12.02	.001	11.44	.001
11	21.69	<.001	2.54	.111
12	27.65	<.001	3.95	.047
13	6.45	.011	1.04	.308
14	59.94	<.001	34.24	<.001
15	40.44	<.001	41.75	<.001
16	37.33	<.001	18.43	<.001
17	31.34	<.001	7.24	.007
18	5.71	.017	0.34	.560
19	15.48	<.001	13.38	<.001
20	23.74	<.001	13.68	<.001
21	31.14	<.001	27.84	<.001
22	13.87	<.001	20.98	<.001
23	87.52	<.001	77.12	<.001
24	28.5	<.001	22.16	<.001
25	1.04	.308	0.17	.680
26	33.89	<.001	30.38	<.001
27	2.31	.129	5.93	.015
28	77.29	<.001	29.1	<.001
29	18.01	<.001	13.51	<.001
30	28.3	<.001	14.67	<.001
31	34.1	<.001	0.02	.888
32	18.9	<.001	15.49	<.001
33	23.14	<.001	14.31	<.001
34	6.98	.008	8.07	.005
35	4.57	.033	1.64	.200
36	68.62	<.001	58.53	<.001
37	3.94	.047	5.29	.021
38	6.21	.013	4.97	.026
39	1.49	.222	2.69	.101
40	28.07	<.001	0	.999
41	22.71	<.001	24.87	<.001
42	26.85	<.001	13.97	<.001
43	18.49	<.001	14.29	<.001
44	9.43	.002	11.75	.001
45	9.94	.002	7.86	.005
46	20.25	<.001	5.01	.025
47	7.61	.006	7.78	.005

Note. Log-likelihood ratio tests (LRTs) are presented to compare two models under different assumptions regarding the L parameters. For the AJ_{cj} and AJ_{gj} models, the tests assess the complex models, which include ordinal constraints on the parameters ($L_f > L_s$, $L_f > L_{go}$, and $L_f > L_{gn}$) against their null counterparts that assume all L parameters are equal (null hypothesis). The G^2 columns report the LRT statistics, with significant p values ($p < .05$) indicating support for the models with ordinal constraints over the null models. AJ = Atkinson–Juola.

Table D3*Nested Model Comparison: AJ Hypothesis Testing for Confidence Levels*

Id	AJ _{cj}		AJ _{gj}	
	$G^2(4)$	p	$G^2(4)$	p
1	73.17	<.001	72.42	<.001
2	50.03	<.001	43.62	<.001
3	28.66	<.001	34.43	<.001
4	20.3	<.001	19.09	<.001
5	96.11	<.001	122.39	<.001
6	6.56	.010	15.4	<.001
7	139.39	<.001	136.55	<.001
8	49.04	<.001	46.7	<.001
9	17.8	<.001	9.11	.003
10	5.09	.024	8.81	.003
11	133.09	<.001	120.36	<.001
12	15.37	<.001	16.71	<.001
13	99.5	<.001	83.69	<.001
14	60	<.001	62.52	<.001
15	151.57	<.001	153.58	<.001
16	78.32	<.001	83.81	<.001
17	30.65	<.001	25.96	<.001
18	38.95	<.001	34.53	<.001
19	18.25	<.001	22.87	<.001
20	27.84	<.001	24.52	<.001
21	67.23	<.001	78.42	<.001
22	28.73	<.001	39.82	<.001
23	2.66	.103	7.1	.008
24	24.9	<.001	17.94	<.001
25	36.05	<.001	38.7	<.001
26	0	.999	21.38	<.001
27	88.12	<.001	85.19	<.001
28	87.07	<.001	70.01	<.001
29	10.13	.001	14.74	<.001
30	45.13	<.001	46.63	<.001
31	2.27	.132	8.74	.003
32	45.41	<.001	45.63	<.001
33	36.22	<.001	34.96	<.001
34	106.28	<.001	102.03	<.001
35	134.06	<.001	128.39	<.001
36	59.95	<.001	58.68	<.001
37	73.52	<.001	57.23	<.001
38	2.29	.130	2.62	.106
39	4.01	.045	3.75	.053
40	1.45	.229	2.34	.126
41	26.54	<.001	27.59	<.001
42	111.18	<.001	102.14	<.001
43	24.28	<.001	27.57	<.001
44	23.98	<.001	19.96	<.001
45	34.97	<.001	32.87	<.001
46	30.64	<.001	27.62	<.001
47	24.88	<.001	44.92	<.001

Note. Log-likelihood ratio tests (LRTs) comparing two models under different assumptions for the H parameters are presented. For both the AJ_{cj} and AJ_{gj} models, these comparisons contrast the complex models, which impose ordinal constraints on the parameters (alternative hypothesis: $H_{s, b} > H_{f, b} > H_{go, b}$ and $H_{s, b} > H_{f, b} > H_{gn, b}$) with their corresponding null models, which assume equality among all H parameters (null hypothesis). The G^2 columns display the LRT statistics, with significant p values ($p < .05$) indicating rejection of the null hypothesis in favor of the models with ordinal constraints. AJ = Atkinson–Juola.

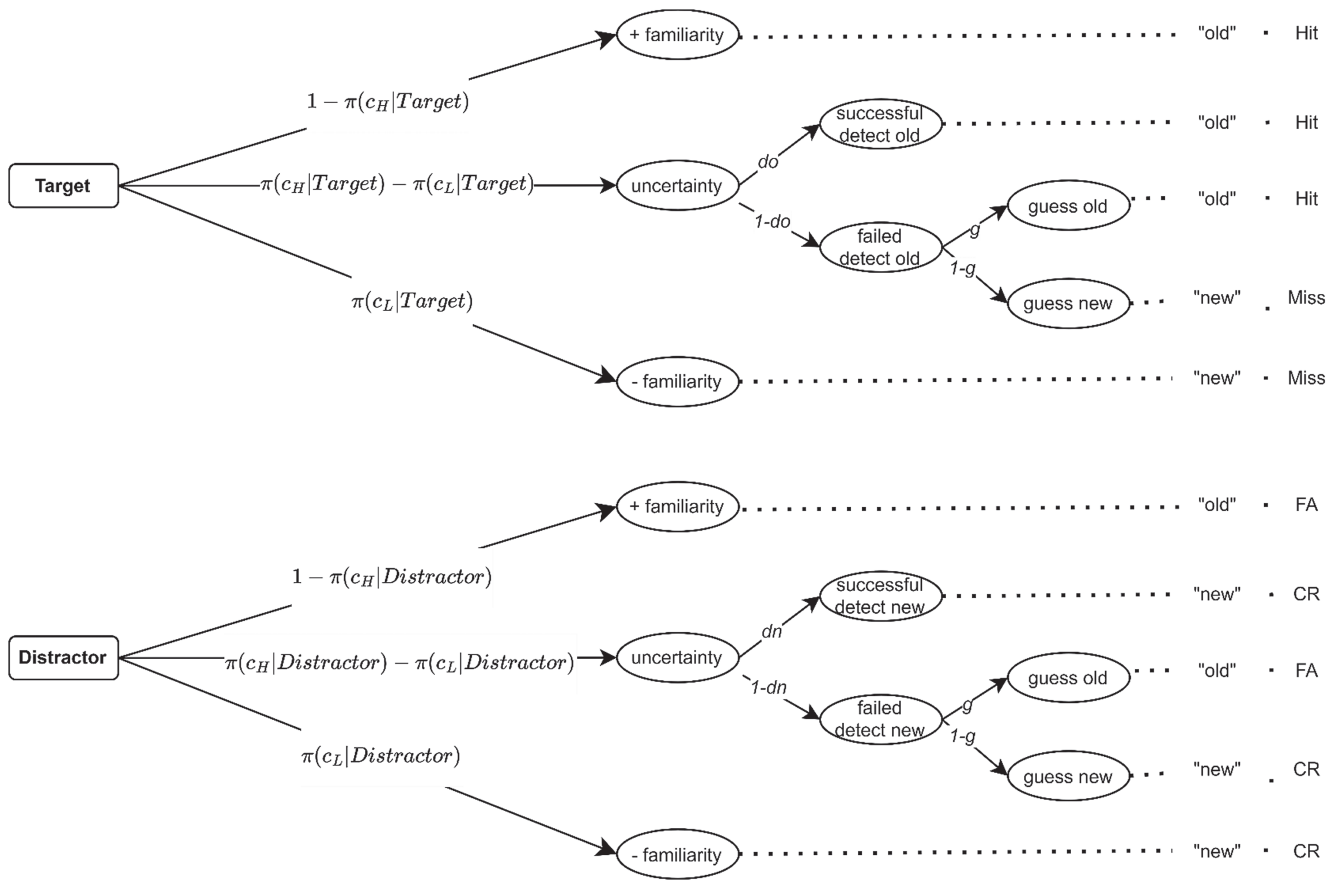
(Appendices continue)

Appendix E

Response Times and Confidence Levels—Extended General Model Corrections

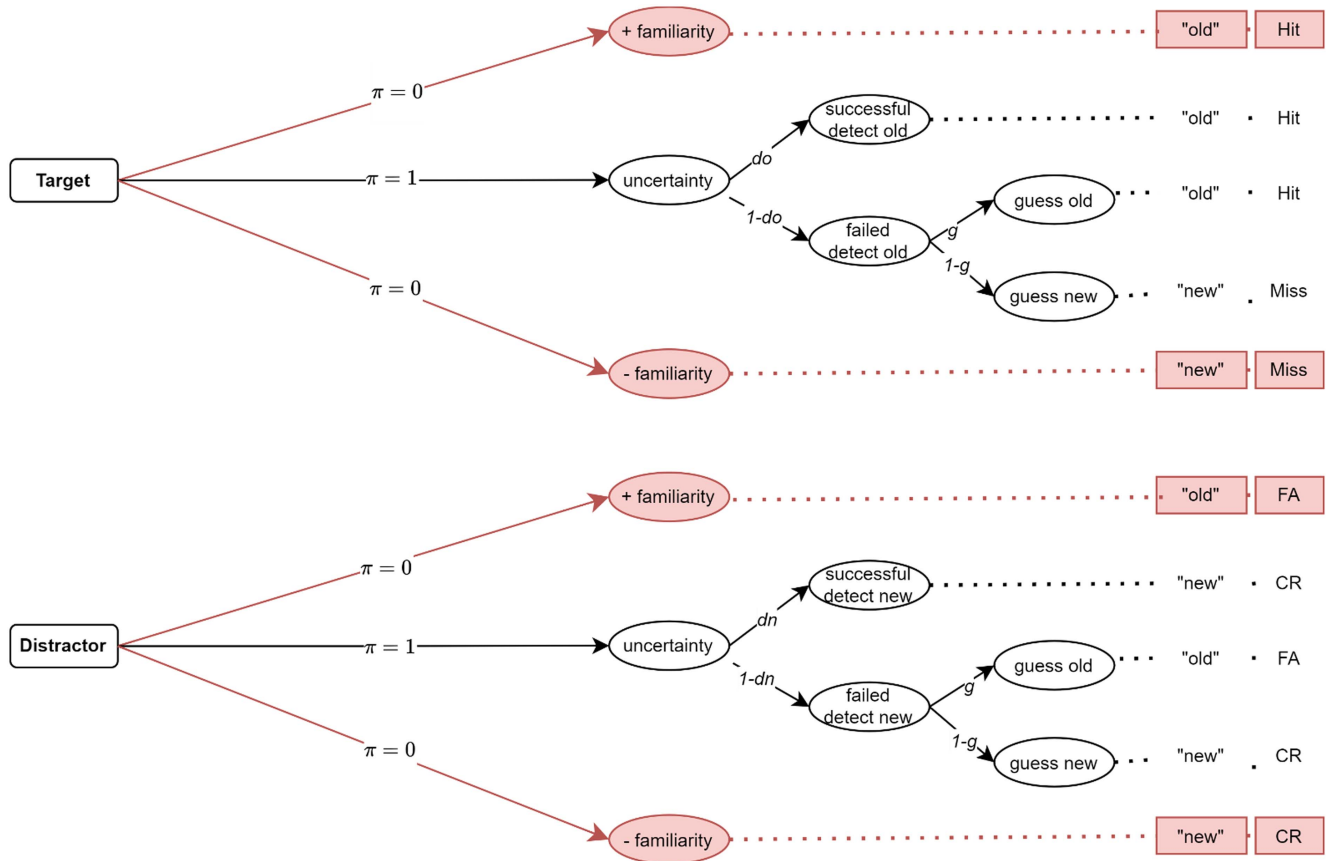
Figure E1

General Model



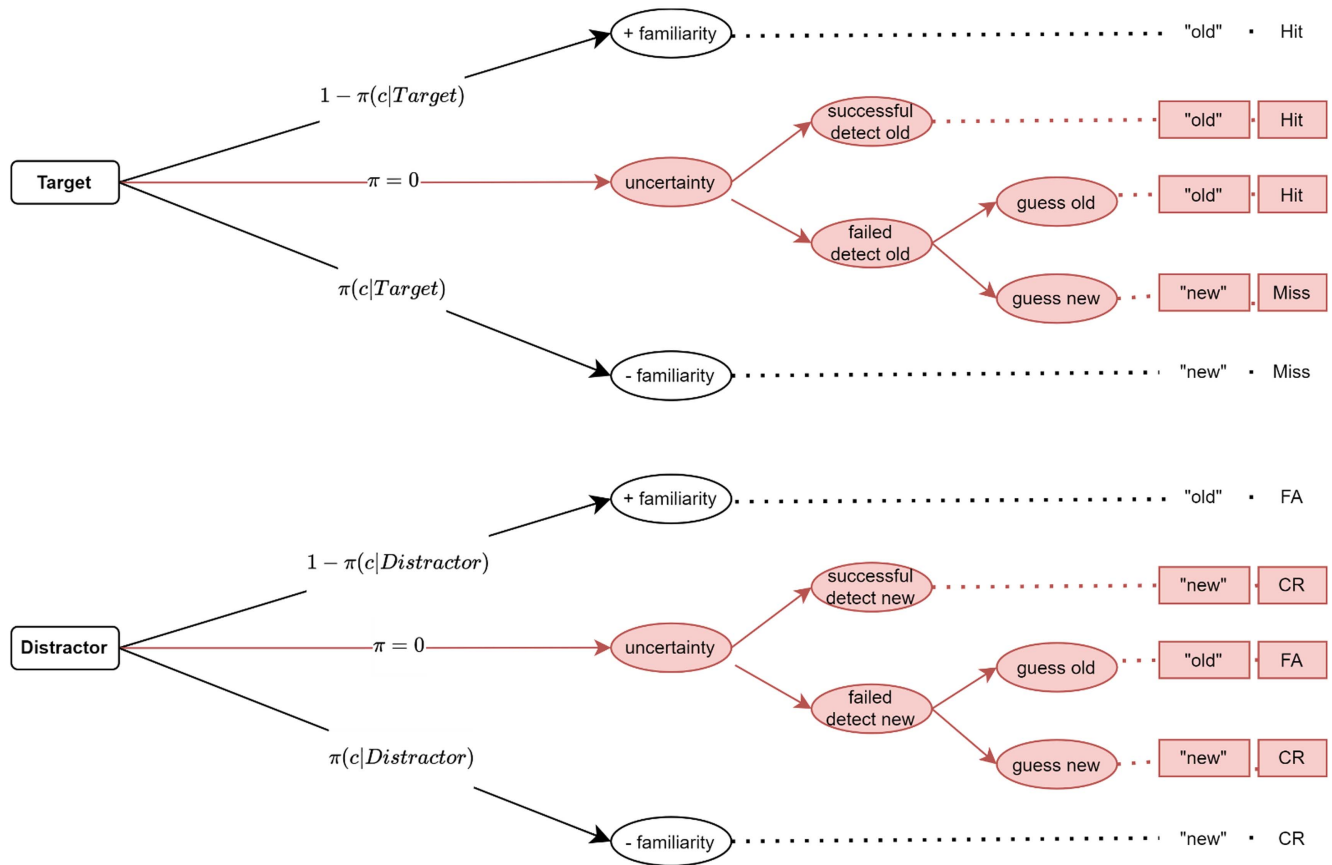
Note. General recognition memory model. A general model is presented in which the branches of +familiarity and –familiarity represent the areas above the high criterion, c_H , or below the low criterion, c_L , respectively. Meanwhile, the uncertainty region corresponds to the area between the two criteria. If the familiarity of a stimulus falls within the uncertainty region, a search for information can occur that may lead to either a successful old detection process (in the target presentation), with a probability of do , or a successful new detection (in the distractor presentation), with a probability of dn . There are complementary probabilities for these detection processes to fail. In case of failure, the guessing process would occur, with a probability of go for guessing old and gn for guessing new. FA = false alarm; CR = correct rejection.

(Appendices continue)

Figure E2*General Model, Subcase 2HT*

Note. The general model, as the 2HT model, operates with +familiarity and -familiarity representing the areas above the high criterion, c_H , or below the low criterion, c_L , respectively. In this subcase, the criteria c_H tends toward infinity or a very high value, while c_L tends toward negative infinity or a very low value. Consequently, the areas above and below the high and low criteria are practically null ($\pi = 0$). In contrast, the uncertainty region will have a probability of one, and the general model will be equivalent to a 2HT model. 2HT = two-high-threshold; FA = false alarm; CR = correct rejection. See the online article for the color version of this figure.

(Appendices continue)

Figure E3*General Model, Subcase SDT*

Note. The general model, as the SDT model, operates with +familiarity and –familiarity representing the areas above the high criterion, c_H , or below the low criterion, c_L , respectively. In this subcase, the criterion c_H and c_L coincide ($c = c_H + c_L$). Therefore, the uncertainty region will have a probability of 0, and the general model would be equivalent to an SDT model. SDT = signal detection theory; FA = false alarm; CR = correct rejection. See the online article for the color version of this figure.

Appendix F

Example for Testing Model Predictions

Certain experimental manipulations can systematically influence the degree of bias individuals exhibit when making recognition decisions, that is, deciding whether a stimulus is “old” or “new.” A well-established method for manipulating response bias is to employ a payoff matrix. Within the framework of SDT, increasing the costs of FAs and/or enhancing the benefits for correct rejections should shift the decision criterion (c_j) rightward along the internal evidence continuum, thereby promoting a more conservative response bias. Alternatively, if misses are made to be more costly and/or hits result in higher benefits, the criterion should be shifted to the left. Notably, these shifts affect not only the primary yes/no criterion but also the associated confidence criteria, making high-confidence responses more or less stringent (Swets et al., 2000). In contrast, the 2HT model posits that payoff manipulations do not influence detection probabilities (d_o and d_n), but rather modulate the decision-level guessing

parameter (g_j). When participants are cautioned against making FAs are increased via payoffs, g_j decreases, and they are less inclined to guess “old” (Kellen & Klauer, 2014; Voormann et al., 2025). The high-familiarity region in the 2HT model which should result in fast, high-confidence responses, and the analogous uncertainty region in the AJ model incorporate a recollection-based mechanism that also yields high-confidence responses, but typically with longer RTs. Therefore, as the uncertainty zone becomes broader and the fast-familiarity zone shrinks, we expect a decrease in the ratio of fast-to-slow high-confidence hits. This pattern reflects the model’s predictions that with greater caution, more high-confidence responses are derived from deliberate recollections rather than automatic familiarity judgments.

To further evaluate and clarify the model-specific predictions of the AJ, SDT, and 2HT models, we next examine the ratios of

fast versus slow high-confidence hits. From this point forward, we will refer to these ratios as R_{AJ} , R_{SDT} , and R_{2HT} , corresponding to the predictions of each respective model. This logical description is designed to assess the predictions resulting from the AJ model that differentiate it from both the SDT and 2HT models.

Let us assume a scenario with an experimental manipulation of caution with three levels $j = \{1, 2, 3\}$, where increasing values of j correspond to increasing levels of caution induced by changes in the payoff matrix. Under this framework, the 2HT, SDT, and AJ each generate distinct internal changes that influence the probability of high-confidence hit responses. In the 2HT model, increasing caution is hypothesized to selectively reduce the guessing parameter such that $g_1 > g_2 > g_3$ while detection probabilities remain constant. In the SDT model, the effect of increased caution manifests as a rightward shift in the decision criterion. Given the model's inclusion of hierarchical confidence thresholds, this shift also applies to all confidence criteria that increase progressively across the three conditions, such that $c_{1j} < c_{2j} < c_{3j}$. In contrast, in the AJ model, manipulations of caution result in a widening of the uncertainty region—defined as the probability area between the low- and high-confidence criteria—while simultaneously reducing the probability associated with responses from the high-familiarity extreme. Based on these assumptions, we now examine how each model predicts changes in the respective ratios of fast-to-slow high-confidence hit responses: R_{AJ} , R_{SDT} , and R_{2HT} . In the analysis that follows, we restrict attention to model branches that directly contribute to high-confidence hit responses.^{F1}

2HT Model Predictions

To formally characterize how experimental manipulations of payoffs affect the probabilities of fast versus slow high-confidence recognition responses (see Figure F1), we examine the behavior of the 2HT model under varying levels of decisional caution (manipulated, for instance, by a payoff matrix).

Specifically, we define the ratio R_{2HT_j} as (Equation F1):

$$R_{2HT_j} = \frac{P(\text{HIT, high, fast})}{P(\text{HIT, high, slow})} = \frac{[(do \cdot H_{do} \cdot L_{do})] + (1 - do) \cdot g_j \cdot H_{go} \cdot L_{go}}{[do \cdot H_{do} \cdot (1 - L_{do})] + (1 - do) \cdot g_j \cdot H_{go} \cdot (1 - L_{go})}, \quad (\text{F1})$$

where do represents the probability of detecting an “old” item, H_{do} the probability of a high-confidence response under detection, and L_{do} the proportion of those detection-based responses that are fast. The guessing parameter g_j , which varies as a function of experimental condition j , reflects the probability of guessing “old” when detection fails, and H_{go} the probability of high-confidence judgment rating in guessing, and L_{go} the probability that such a high-confidence judgment is fast.

To simplify the expression and analyze its behavior across the j levels of caution, we define four constants A , A' , B , and B' such that (Equation F2),

$$\begin{aligned} A &= [(do \cdot H_{do} \cdot L_{do})] \\ A' &= [do \cdot H_{do} \cdot (1 - L_{do})] \\ B &= (1 - do) \cdot H_{go} \cdot L_{go} \\ B' &= (1 - do) \cdot H_{go} \cdot (1 - L_{go}). \end{aligned} \quad (\text{F2})$$

And therefore (Equation F3),

$$R_{2HT_j} = \frac{P(\text{HIT, high, fast})}{P(\text{HIT, high, slow})} = \frac{A + g_j \cdot B}{A' + g_j \cdot B'} \quad (\text{F3})$$

According to the 2HT model, detection-based responses are typically relatively fast, whereas those based on guessing are usually slower. Therefore, we will assume that $A > A'$ and $B' > B$. Furthermore, as mentioned before, experimental manipulations are assumed to reduce the probability of guessing, such that $g_1 > g_2 > g_3$. Accordingly, R_{2HT_j} should increase across the j conditions. This occurs because both numerator and denominator decline with decreasing g_j , but the denominator declines more steeply due to the larger weight on the slower guessing term D . Consequently, as caution increases and guessing is reduced, the model predicts a relative increase in fast high-confidence hits compared to slow ones. That is (Equation F4),

$$R_{2HT_1} < R_{2HT_2} < R_{2HT_3}. \quad (\text{F4})$$

SDT Model Predictions

In the SDT model with CL, high-confidence hit responses are modeled as originating from the uppermost tail of the signal distribution, defined by the threshold c_{3j} (see Figure F2). Here, we assume that c_{3j} increases monotonically with j , as specified in the main text. Thus, $c_j + c_2 + c_3$ and the probability of a high-confidence hit, which we denote $P1_j$, are given by Equation F5:

$$P1_j = 1 - \Phi\left(\frac{c_j + c_2 + c_3 - \mu}{\sigma}\right). \quad (\text{F5})$$

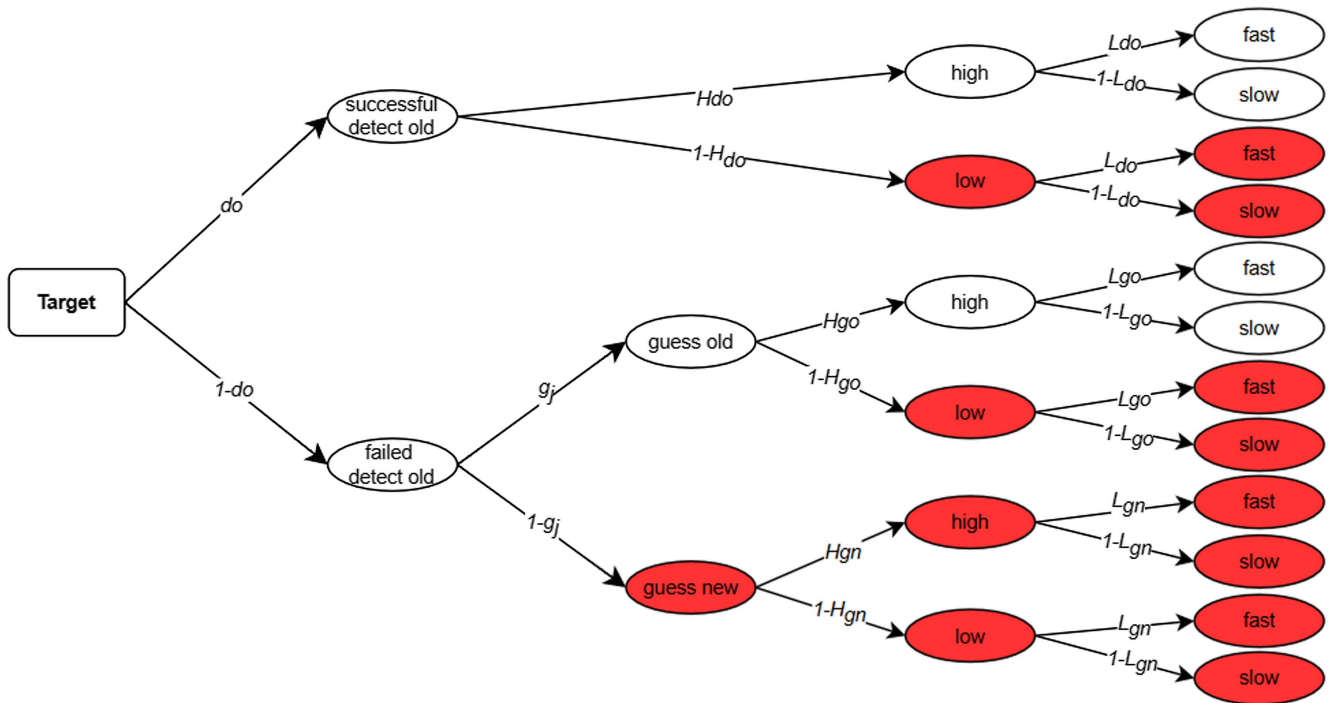
Here, μ is the mean of the target distribution, σ is its standard deviation, and c_j , c_2 , and c_3 represent decision and confidence criteria along the evidence axis. Crucially, c_j is modulated by the experimental manipulation of payoff (e.g., increased penalties for FAs), such that increasing caution results in higher values of c_j .

Therefore, as c_j increases, $P1_j$ decreases. This reflects a key prediction of SDT: Increased caution reduces the probability of issuing high-confidence hits, as a higher level of internal evidence is required to surpass the now more stringent upper threshold.

While the absolute number of high-confidence hits decreases, the nature of the evidence driving these decisions changes. Because only trials with very strong internal evidence will now surpass the more conservative high-confidence threshold, we expect the RTs to be shorter. That is, the smaller subset of high-confidence hits that still occur under high caution is likely to be those supported by especially strong, unambiguous familiarity signals. These strong signals, in turn, facilitate faster decision responses.

Consequently, although the total probability of high-confidence hits diminishes, the ratio of fast-to-slow high-confidence hits, denoted

^{F1} See Appendix A and the Model Details section for a more complete understanding of the full specifications of the SDT, 2HT, and AJ models.

Figure F1*Two-High-Threshold Model: High-Confidence Hit Responses*

Note. The multinomial processing tree diagram represents the probabilities associated with the branches that lead to fast and slow high-confidence hits. Red nodes indicate branches that do not lead to high-confidence hits and are therefore disregarded for the current purposes. See the online article for the color version of this figure.

R_{SDT_j} , is expected to increase with higher levels of caution. This is because the more conservative criterion acts as a stricter filter, allowing only the most robust, rapid responses to be classified as high confidence. Formally, this leads to the model-derived prediction:

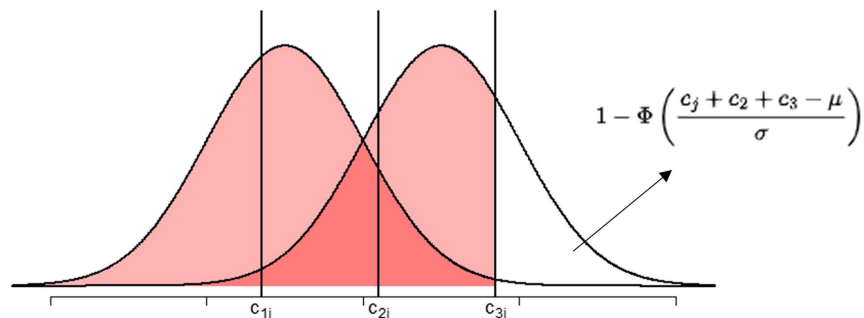
$$R_{SDT_1} < R_{SDT_2} < R_{SDT_3}, \quad (F6)$$

where conditions 1–3 represent increasing levels of payoff-induced caution. Thus, SDT uniquely predicts that increasing decision thresholds leads not only to fewer high-confidence hits

but also to an enrichment of these responses in terms of speed and decisional strength.

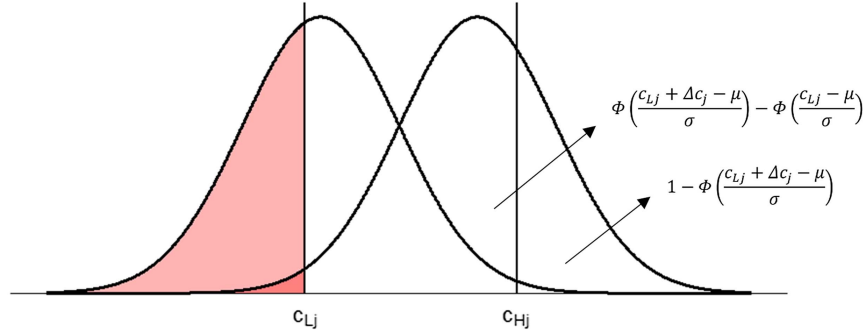
AJ or General Model Predictions

In the AJ or general model, high-confidence hit responses can originate either from the uncertainty region or from the extreme familiarity region, which are determined by the low- and high-confidence criteria (c_{Lj} and c_{Hj}) that vary along with payoff experimental manipulations j (see Figure F3). Where $c_{Hj} = c_{Lj} + \Delta c_j$.

Figure F2*SDT Model: High-Confidence Hit Responses*

Note. The SDT diagram represents the probability area associated with high-confidence hits. The red-shaded regions indicate probability areas that do not result in high-confidence hits and are therefore excluded from the current purposes. SDT = signal detection theory. See the online article for the color version of this figure.

(Appendices continue)

Figure F3*Atkinson and Juola Model: High-Confidence Hit Responses*

Note. The Atkinson–Juola diagram illustrates the probability regions associated with high-confidence hits. In this model, the area between the low- and high-confidence criteria (c_{Lj} and c_{Hj}) is referred to as the *uncertainty zone*, while the region above the high criterion (c_{Hj}) corresponds to *extreme positive familiarity*. Both regions can give rise to high-confidence hit responses. The red-shaded areas represent probability regions that do not result in high-confidence hits and are therefore excluded from the current analysis. See the online article for the color version of this figure.

To simplify the development of the model's logic predictions, we will denote the probability associated with the uncertainty region as P_u and the probability corresponding to the extreme positive familiarity region as P_f (Equation F7).

$$\begin{aligned} P_u &= \Phi\left(\frac{c_{Lj} + \Delta c_j - \mu}{\sigma}\right) - \Phi\left(\frac{c_{Lj} - \mu}{\sigma}\right) \\ P_f &= 1 - \Phi\left(\frac{c_{Lj} + \Delta c_j - \mu}{\sigma}\right). \end{aligned} \quad (\text{F7})$$

As in the AJ model, high-confidence hits may originate from two distinct cognitive sources: extreme familiarity or uncertainty-based recollection/guessing. The fast-to-slow high-confidence hit ratio is defined as Equation F8:

$$R_{AJ_j} = \frac{P(\text{HIT, high, fast, familiarity}) + P(\text{HIT, high, fast, uncertainty})}{P(\text{HIT, high, slow, familiarity}) + P(\text{HIT, high, slow, uncertainty})}. \quad (\text{F8})$$

Assuming all conditional parameters remain stable across conditions (e.g., hit rates, speed proportions, and retrieval strategies), and only the weights P_{fj} (familiarity zone) and P_{uj} (uncertainty zone) vary with caution level j , we define the following constants as Equation F9:

$$\begin{aligned} a &= H_f \cdot L_f (\text{fast familiarity}) \\ a' &= H_f \cdot (1 - L_f) (\text{slow familiarity}) \\ b &= r \cdot H_r \cdot L_r (\text{fast recollection}) \\ b' &= r \cdot H_r \cdot (1 - L_r) (\text{slow recollection}) \\ c &= (1 - r)(1 - g) \cdot H_g \cdot L_g (\text{fast guessing}) \\ c' &= (1 - r)(1 - g) \cdot H_g \cdot (1 - L_g) (\text{slow guessing}) \end{aligned} \quad (\text{F9})$$

Substituting, we get the simplified expression (Equation F10):

$$R_{AJ_j} = \frac{P_{f,j} \cdot a + P_{u,j} \cdot (b + c)}{P_{f,j} \cdot a' + P_{u,j} \cdot (b' + c')}. \quad (\text{F10})$$

Because increasing caution leads to a reduction in the extreme familiarity region: $P_{f,1} > P_{f,2} > P_{f,3}$; and a widening of the uncertainty region: $P_{u,1} < P_{u,2} < P_{u,3}$. And, responses from familiarity are predominantly fast (i.e., $a > a'$), while responses from uncertainty tend to be slower (i.e., $b + c < b' + c'$), then (Equation F10):

$$R_{AJ_1} > R_{AJ_2} > R_{AJ_3}. \quad (\text{F11})$$

That is, as caution increases across conditions j , the AJ model predicts a systematic decrease in the fast-to-slow ratio among high-confidence hit responses. This pattern reflects the shift from fast, familiarity-driven judgments to slower, recollection-based processing within the uncertainty region.

Therefore, the predictions derived from the AJ model diverge systematically from those of the SDT and 2HT models. That is, while both SDT and 2HT predict an increasing ratio of fast-to-slow high-confidence hits with greater caution, the AJ model uniquely predicts a decreasing trend under the same manipulations. This reversal reflects the model's emphasis on slower, recollection-driven contributions from the uncertainty zone as caution increases.

Received May 30, 2024

Revision received November 7, 2025

Accepted November 11, 2025 ■