Adding a speed–accuracy trade-off to discrete-state models: A comment on Heck and Erdfelder (2016)

Jeffrey J. Starns¹

Published online: 12 March 2018
© Psychonomic Society, Inc. 2018

Abstract

Heck and Erdfelder (2016) developed a model that extends discrete-state multinomial processing tree models to response time (RT) data. Their model is an important advance, but it does not have a mechanism to produce the speed–accuracy trade-off, the bedrock empirical observation that rushed decisions are less accurate. I present a similar model, the “discrete-race” model, with a simple mechanism for the speed–accuracy trade-off. In the model, information that supports detection of the stimulus type is available for some proportion of items and unavailable for others. Both the amount of time needed for detection to succeed and the amount of time that the decision maker waits before guessing are variable from trial to trial. Responses are based on detection when it is available *and* has a finishing time before the guess time for that trial. In other words, the decision maker sometimes loses opportunities to respond correctly on the basis of detection by first making a guess. These lost opportunities are more common when the guess-time distribution tends to have low wait times, which decreases accuracy. I report simulations showing that the model can accurately recover parameter values and is strongly constrained by the speed–accuracy trade-offs across conditions with different levels of response caution.

Keywords Discrete-state models · Response time (RT) models · Recognition memory

A complete theory of decision making must account for the time course of decisions in addition to their accuracy, and modeling response time (RT) distributions has important practical benefits such as improving the ability to discriminate alternative models and providing separate measures of psychologically meaningful processes (e.g., Ratcliff & McKoon, 2008). In the recognition memory literature, RT modeling has played a critical role in testing theoretical accounts (e.g., Criss, 2010; Starns, Ratcliff, & White, 2012) and measuring memory ability (e.g., Ratcliff, Thapar, & McKoon, 2004). The RT modeling literature has primarily relied on sequential sampling models such as the diffusion model (Ratcliff, 1978), the linear ballistic accumulator model (Brown & Heathcote, 2008), the Poisson race model (Smith & Van Zandt, 2000), and the RTCON model (Ratcliff & Starns, 2009). Heck and Erdfelder (2016) recently proposed a method for extending multinomial processing tree models to RT data and applied their method to data from a recognition

memory task (see Batchelder & Riefer, 1999, for a comprehensive discussion of processing tree models). Although discrete-state theorists had previously tested predictions for RT data (Kellen, Singmann, Vogt, & Klauer, 2015; Province & Rouder, 2012) and taken some steps toward developing RT models (Hu, 2001), Heck and Erdfelder’s work is an important advance that will provide the opportunity to more rigorously test discrete-state models.

One critical test for any RT model is its ability to accommodate the speed–accuracy trade-off, the ubiquitous finding that encouraging people to make quicker decisions decreases the probability that they will make correct decisions. Speed–accuracy manipulations have played a large role in testing RT models, and correctly accounting for the speed–accuracy trade-off is a critical measurement property of a model. For example, young adults and older participants differ in their pace of responding in a recognition task (Ratcliff et al., 2004), so determining the relative memory ability of the two groups requires an appropriate model to correct for this difference in response caution. Sequential-sampling models have mechanisms to produce a speed–accuracy trade-off, and as a result the models make specific predictions about how much accuracy should change for a given change in decision time. These predictions have been thoroughly tested across a wide

✉ Jeffrey J. Starns
jstarns@psych.umass.edu

¹ Department of Psychological and Brain Sciences, University of Massachusetts, 135 Hicks Way, Amherst, MA 01003-9271, USA

variety of decision tasks, and the models consistently provide a close match to data (e.g., Wagenmakers, 2009). In the context of this past work, explaining the speed–accuracy trade-off is an important goal for discrete-state RT models.

In this comment, I note that the Heck and Erdfelder (2016) model lacks a speed–accuracy mechanism, and thus cannot predict how changes in the timing of responding will affect accuracy. This property of the model has two consequences: It means (1) that the model cannot be rigorously tested with response caution manipulations that are common in the RT modeling literature, such as instructions to emphasize speed versus accuracy, and (2) that the model cannot provide a performance measure that is independent of response caution. I will propose a model similar to the Heck and Erdfelder (2016) model, but with an added mechanism for the speed–accuracy trade-off, and explain the basis of this trade-off in the model. I will also present some brief simulation results to show that the model can be tested with timing manipulations and recovers parameters accurately.

The Heck and Erdfelder (2016) model

The decision tree structure of the Heck and Erdfelder (2016) model is the same as the traditional 2HT model (as is shown in their Fig. 2). For studied or “old” items, the decision maker either experiences the “detect old” state (with probability d_O) or the “guess” state (with probability $1 - d_O$), where the former means that information from memory identified the item as studied and the latter means that memory failed (Snodgrass & Corwin, 1988). For nonstudied or “new” items, the decision maker either experiences the “detect new” state (with probability d_N) or the “guess” state (with probability $1 - d_N$), where the former means that information is available that rules out the possibility that the word could have been studied and the latter means that no such information is available. The probability of responding “old” when guessing is g , and this probability applies to both studied and nonstudied items. In other words, the model assumes that a decision maker has no way to discriminate the two stimulus types in the “guess” state.

Heck and Erdfelder (2016) extended the 2HT model to RT data by specifying an RT distribution associated with responding based on each possible information state, “detect old,” “detect new,” and “guess.” These distributions can take any functional form, or the model can have free probability parameters for the proportion of RTs in binned response categories. All error responses are based on guessing in the standard 2HT model, so the predicted RT distributions for errors are identical to the RT distribution for the guess state. Correct decisions, in contrast, are a mixture of responses based on detection and guessing, so the predicted RT distribution is a mixture of the two distributions for guessing and detection

with mixing weights determined by the detection and guessing parameters.

In the Heck and Erdfelder (2016) model, changing the RT distributions associated with each information state has no effect on accuracy. Like the traditional (accuracy only) 2HT model, their model predicts that the proportion of correct responses in a condition is $d + (1 - d)g_C$, where d is the probability of detection and g_C is the probability of guessing correctly. Thus, the parameters defining the RT distributions play no role in the accuracy predictions. This property of the model means that the model assumptions cannot be tested by manipulating the relative emphasis on accuracy versus speed, a trademark manipulation in the RT modeling literature (e.g., Ratcliff & McKoon, 2008; Starns, Ratcliff, & McKoon, 2012). The model can accommodate faster responding in speed-emphasis conditions by shifting the RT distributions, but this mechanism does not produce the drop in accuracy that is observed empirically. One could estimate different d parameters for speed emphasis and accuracy emphasis to accommodate the accuracy change, but this is just a free parameter that is in no way constrained by the change in RTs across the emphasis conditions. In contrast, traditional RT models accommodate speed–accuracy manipulations with parameters that are constrained by the change in both accuracy and RT across different levels of response caution (e.g., Ratcliff & Smith, 2004).

The lack of a speed–accuracy mechanism also means that the model cannot be used as a measurement tool to distinguish differences in memory ability from differences in response caution. For example, the model cannot answer questions like “How much would performance for young participants improve if they slowed down to the cautious pace observed for older participants?” Traditional RT models can answer such questions, at least in principle, because the speed–accuracy trade-off is produced by central model mechanisms.

The discrete-race model

A simple way to implement a speed–accuracy trade-off in a discrete-state model is to assume that early guesses sometimes preclude later detection. At the beginning of a trial, the decision maker does not know whether or not detection is available for the stimulus. As more time passes without successful detection, the probability that detection is available steadily declines. At some point in this process, the decision maker gives up on trying to detect and makes a guess response. I assume that both the time needed for detection to succeed and the time at which the participant becomes willing to guess vary from trial to trial. Participants who are very cautious in their speed–accuracy trade-off will tend to wait a long time for detection to succeed, whereas participants who are incautious will tend to quickly give up on detection. The functional

detection rate will be lower with incautious responding due to an increased number of trials in which the participant could have detected the item but lost their chance with a quick guess. I call the model the discrete-race model, because detection must race against the guess process to determine the response.

Psychologically, the discrete-race model is consistent with a decision process in which a detection attempt and the selection of a guess response occur in parallel during a trial. Successful detection terminates the race and leads to a response, whereas the outcome of the guess selection process does not terminate the race. That is, the finishing time for the guess process does not represent how long it takes to select a guess; instead, it represents how long the decision maker waits for detection to succeed before they make a guess response that they already selected. Alternative processing assumptions are possible; for example, one could assume that participants do not begin the guess selection process until they have already given up on detection, so there is a time period at the end of guess trials during which detection has no chance of succeeding. These different alternatives can be explored in future work, but the present model is a simple starting point.

The discrete-race model has detection and guessing parameters as in the accuracy-only 2HT model, but the detection parameter has a different interpretation. For the RT model, I will use D_{AV} to identify the proportion of trials for which detection is *available*, meaning that information identifying the stimulus would be retrieved if the decision maker waited long enough. Unlike the accuracy-only model and the Heck and Erdfelder (2016) model, this parameter does not uniquely determine the proportion of responses that are *actually* based on detection, which I will denote as D_{AC} . This latter proportion is also influenced by the relative finishing-time distributions of detection and guessing, because the decision maker sometimes guesses before detection has succeeded even when detection is available. In the equations below, I will use g_C to denote the probability that a guess will produce the correct response. Readers should keep in mind that the g parameter in the 2HT model is the probability that a guess will be an “old” responses, so $g_C = g$ when the equations below are applied to studied items and $g_C = (1 - g)$ when the equations are applied to nonstudied items.

I chose to model finishing times with ex-Gaussian distributions, but other forms could be easily substituted. I will denote the probability density of an ex-Gaussian distribution with $\gamma(t, \mu, \sigma, \lambda)$ and the cumulative probability density with $\Gamma(t, \mu, \sigma, \lambda)$, where t is time; μ and σ are the mean and standard deviation of the Gaussian component, respectively; and λ is the rate parameter for the exponential component. The distribution functions and parameters will be subscripted with D or G to identify the detection or guessing processes. For the equations below, I will leave out the parameter notations for brevity, but readers should keep in mind that there are three parameters for each finishing-time distribution.

The cumulative probability of making a correct response at or before a cutoff time t_{cut} is

$$\int_0^{t_{cut}} D_{AV} [1 - \Gamma_D(t)] \gamma_D(t) + D_{AV} [1 - \Gamma_D(t)] \gamma_G(t) g_C + (1 - D_{AV}) \gamma_G(t) g_C dt \quad (1)$$

where D_{AV} is the proportion of trials in which detection is available, g_C is the probability of guessing correctly given that a guess is made, Γ_D and Γ_G are the cumulative probability densities at time t on the distribution of finishing times for detection and guessing, respectively, and γ_D and γ_G are the probability densities at time t on the distribution of finishing times for detection and guessing, respectively. In this equation, each component of the sum within the integral is a different way that a correct response can be made at time t . The first component represents trials for which detection succeeds before a guess is made and produces a response at time t . This occurs when detection is available (with probability D_{AV}), a guess has not been made before time t [with probability $1 - \Gamma_D(t)$], and detection succeeds at time t [with probability density $\gamma_D(t)$]. The second component represents trials for which a correct guess is made at time t before detection succeeds. This occurs when detection is available (with probability D_{AV}), detection has not succeeded by time t [with probability $1 - \Gamma_D(t)$], a guess is made at time t [with probability density $\gamma_G(t)$], and the guess is correct (with probability g_C). The third component represents trials for which detection is not available (with probability $1 - D_{AV}$), a guess is made at time t [with probability density $\gamma_G(t)$], and the guess is correct (with probability g_C). Adding the components gives the overall probability density for a correct response at time t , and integrating from 0 to t_{cut} gives the probability of a correct response before time t_{cut} .

The cumulative probability of making an error response at or before a cutoff time t_{cut} is

$$\int_0^{t_{cut}} D_{AV} [1 - \Gamma_D(t)] \gamma_G(t) (1 - g_C) + (1 - D_{AV}) \gamma_G(t) (1 - g_C) dt \quad (2)$$

where all symbols have the same meaning as in Eq. 1. Each component of the sum within the integral is a different way that an error response can be made at time t . The first component represents trials for which an incorrect guess is made at time t before detection succeeds. This occurs when detection is available (with probability D_{AV}), detection has not succeeded by time t [with probability $1 - \Gamma_D(t)$], a guess is made at time t [with probability density $\gamma_G(t)$], and the guess is an error (with probability $1 - g_C$). The second component represents trials for which detection is unavailable (with probability $1 - D_{AV}$), a guess is made at time t [with probability density $\gamma_G(t)$], and the guess is an error (with probability $1 - g_C$). Adding the components gives the overall probability density for an error response at time t , and integrating from 0 to t_{cut} gives the probability of an error before time t_{cut} . As in the accuracy-only 2HT model, all errors are produced by guesses.

However, in the RT version this includes a mixture of trials in which detection was available (but did not succeed before the guess) and trials in which detection could not occur.

The overall proportion of correct and error responses can be found by integrating the equations above from zero to infinity. Another method to define predicted accuracy is to calculate the probability that responses are based on detection and use this probability to calculate accuracy with the standard 2HT equation. The probability of responding on the basis of detection is

$$D_{AC} = D_{AV} \int \gamma_G(t) \Gamma_D(t) dt \quad (3)$$

where D_{AC} is the proportion of trials *actually* based on detection (as opposed to the proportion in which detection is available) and all other symbols have the same meaning as in Eqs. 1 and 2. The probability of responding on the basis of detection is the probability that detection is available, D_{AV} , times the probability that detection succeeds before the time when the decision maker gives up on detection and makes a guess (the integral in Eq. 3). The terms inside the integral are the probability that a guess will be made at time t [$\gamma_G(t)$] and the probability that the detection process will succeed before time t [$\Gamma_D(t)$], and multiplying gives the probability that the participant would have made a guess at time t but didn't because the detection time was less than t . Integrating across all of the possible guess times gives the total probability that detection succeeds before a guess. Thus, the probability of detection depends on both the D_{AV} parameter and the relative finishing times for detection and guessing.

Figure 1 shows an example speed–accuracy function generated by the model across four values of the Gaussian mean of the guessing-time distribution (see the figure caption for other parameter values). Panels A and B show the .1, .5, and .9 quantiles of the RT distribution for correct and error responses, respectively, and Panel C shows the proportions of correct responses. The RT distributions are appropriately right-skewed (i.e., the .1 and .5 quantiles are closer than the .5 and .9 quantiles). This is not surprising, given the choice of ex-Gaussian finishing-time distributions. As the Gaussian mean of the guessing-time distribution increases, accuracy increases and RTs get slower for both correct and error responses. The RT distributions also increase in spread for slower guess times. Although this is easier to see for the correct RTs, it holds for the error RTs as well. Different distribution patterns are possible with different changes to the guessing-time distribution; for example, one could represent greater response caution by increasing the exponential rate instead of the Gaussian mean or by changing both the Gaussian mean and the standard deviation. However, the general pattern of an increase in accuracy for slower guess times holds regardless of how the change in guess times is achieved.

Fig. 1 Example speed–accuracy trade-off function produced by varying the mean of the Gaussian component of the finishing-time distribution for guessing (μ_G) in the discrete-race model. (a) The .1, .5, and .9 quantiles of the RT distribution for correct responses. (b) The same quantiles for errors. (c) Proportions of correct responses. Predictions are shown for target trials with the following parameters: $D_{AV-O} = .75$, $g = .5$, $\mu_{DO} = .6$, $\sigma_D = .05$, $\lambda_D = .2$, $\sigma_G = .05$, and $\lambda_G = .2$

Parameter recovery with a caution manipulation

One advantage of adding a speed–accuracy trade-off to the model is that the model can be tested by response caution manipulations, such as asking participants to emphasize speed or accuracy in their responding. To explore whether caution manipulations are indeed a strong test of the model, I simulated data from the model across conditions promoting cautious or incautious responding (the R code for all simulations is available on the Open Science Framework project “Adding a Speed–Accuracy Tradeoff to Discrete-State Models”: Starns, 2018). The caution manipulation was modeled by changing the distribution of guess times while holding the other model parameters constant. Data for both studied and nonstudied items were simulated within each caution condition, so there were four conditions in total. I set up the simulations to mimic the type of multiple-session experiment that is frequently used to test RT models, so each simulated participant had 3,000 trials evenly divided amongst the four conditions. Following the popular chi-squared fitting method for RT models (Ratcliff & Tuerlinckx, 2002), I fit the proportion of responses in RT bins defined by the .1, .3, .5, .7, and .9 quantiles of each RT distribution in the simulated data. This results in six frequencies each for both correct and error responses in a given condition, so each condition has 12 total frequencies and 11 degrees of freedom (one is lost because the 12 frequencies must sum to the total number of trials). Thus, the simulated data sets

had 44 degrees of freedom across the four conditions (targets and lures with speed and accuracy emphasis instructions).

I made the following parameter restrictions for fits to the simulated data: (1) the proportion of trials with available detection was equal for cautious and incautious responding (i.e., I assumed an experiment in which speed or accuracy emphasis is cued after identical learning phases); (2) parameters for the detection finishing-time distribution were equal for cautious and incautious responding (in other words, the time needed to retrieve information from memory is a fundamental property of the system not under strategic control); (3) the finishing-time distributions for “detect old” and “detect new” have equal σ and λ parameters, meaning that they have the same shape and spread (but they do not have to have the same position as μ could vary between them); (4) parameters for the guess finishing-time distribution are equal for studied and nonstudied items (this must be true because the decision maker cannot discriminate item types in this state); and (5) the guess finishing-time distributions have equal σ and λ parameters for cautious and incautious responding, meaning that they have the same shape and spread (but they can vary in position due to changes in μ). Parameter values used for each simulated data set were randomly drawn from uniform distributions covering a wide range of plausible values. Specifically, the D_{AV} and g parameters could range from .2 to .8, the μ parameter for target detection ranged from .4 to .8 s, the μ parameter for lure detection was the target detection μ plus a deviate ranging from -0.05 to 0.05 s, the μ parameter for guessing was the target detection μ plus a deviate ranging from 0.15 to 0.5 s, the σ parameters for detection had a lower bound of .05 and an upper bound of $1/5$ of the minimum detection μ value for that parameter set, the σ parameter for guessing was the σ parameter for target detection plus a deviate ranging from -0.02 to 0.020 s, the λ parameters for detection ranged from $1/10$ to $1/3$, and the λ parameter for guessing was the λ for target detection plus a deviate ranging from 0 to $.05$.

The model was fit to two versions of each simulated data set: an unaltered version and a “shifted” version in which the RT distributions for correct “old” and “new” responses were each shifted up by 100 ms in the cautious condition. Fits to the unaltered data will provide information about whether the model can accurately recover parameter values. The purpose of the shifted data is to artificially alter the relative position of RT distributions for correct and error responses to see if this impacts the model’s ability to match the effect of caution on accuracy. The relative finishing times for detection and guessing in the accuracy condition should determine the extent to which becoming less cautious will impact accuracy. If detection is much faster than guessing (relative to the variability in each process), then a moderate speed-up in the guessing process will have a relatively small effect on accuracy. If detection and guess times are more similar, then a moderate speed-up in guessing will have a relatively large effect on accuracy.

The simulations will show whether the relative position of the correct and error RT distribution in the accuracy condition strongly tests this property of the model. If a strong test is possible, then the fits should be substantially worse for the shifted data than for the unaltered data, and the model should consistently misestimate the size of the effect of caution on accuracy in fits to the shifted data.

Figure 2 shows the parameter recovery for the fits to the unaltered versions of 500 simulated data sets.¹ The x -axis on each plot is the true parameter value used to generate the data, and the y -axis is the best-fitting parameter values from fits to the simulated data sets. The plot points are concentrated on the positive diagonal, demonstrating accurate recovery of the parameter values.² The value displayed on the top-left corner of each plot is the proportion of variance in the fit parameter values accounted for by a model in which the predicted fit values are equal to the true parameter values (i.e., the prediction function is a line with an intercept of 0 and a slope of 1). These values are generally close to 1, indicating very accurate parameter recovery. Recovery was not quite as good for the parameters governing the shape and spread of the detection finishing-time distribution, σ_D and λ_D .

Fits to the data sets with altered RT distributions supported the claim that the model is strongly constrained by the relationship between RT and accuracy. First, altering the data substantially worsened the model fit, with an average G^2 value of 129, as compared to 33 for the unaltered data. Second, altering the RT data led to misses in the accuracy predictions, demonstrating the convergent constraints placed on the model by these two forms of data. Notably, the drop in accuracy (percent correct) from the cautious to incautious conditions was .11 for the simulated data, .12 in fits to the unaltered data, and .16 in fits to the altered data. Thus, shifting up the correct RT distributions in the cautious condition made the model overestimate the accuracy difference between the cautious and incautious conditions. This pattern makes sense when one considers the constraints on the model: The relative speeds of correct and error responses constrain the finishing times for detection and guessing, and artificially slowing the correct responses so that the correct and error RTs are more similar must be accommodated by increasing the overlap between the detection and guess finishing-time distributions. This increased overlap means that speeding up the guess times to accommodate the faster RTs in the incautious condition leads to more missed opportunities for detection—and a larger drop in accuracy—than would be the case if the

¹ I excluded the results from four data sets (less than 1%) for which the fitting program failed to converge on the best-fitting values, identified by the fact that the G^2 value for the “best-fitting” parameters was higher than the value for the parameters that had generated the data. In fits to empirical data, these failures to converge can be prevented by running several fits for each data set with different starting values.

² I am not claiming that the parameters of the full discrete-race model are identifiable for all data sets, but these results show that recovery is very good in a situation representing a typical RT-modeling experiment.

Fig. 2 Results from the parameter recovery simulation with a speed–accuracy manipulation. The *x*-axes are the true parameter values used to generate the simulated data, and the *y*-axes are the best-fitting parameter values in fits to those simulated data. The value on each plot is the r^2 for a model in which the predicted best-fitting parameter value is the true parameter value (i.e., the prediction function is a line with a slope of 1 and an intercept of 0). D_{AV-O} and D_{AV-N} are the proportions of trials with available detection for old and new items, respectively; g is the proportion of guesses allocated to the “old” response; μ_{DO} and μ_{DN} are the means for

the detection finishing-time distributions of old and new items, respectively; σ_D is the standard deviation in the Gaussian component of the detection finishing-time distributions; λ_D is the rate of the exponential component of the detection finishing-time distributions; μ_{G-ACC} and μ_{G-SPD} are the means for the guess finishing-time distributions with accuracy versus speed emphasis instructions (cautious vs. incautious responding); σ_G is the standard deviation in the Gaussian component of the guess finishing-time distributions; and λ_G is the rate of the exponential component of the guess finishing-time distributions

detection finishing times were faster (as in the unaltered data). Overall, fits to the altered data show that the model is not flexible enough to accommodate fake data and that the convergent constraints imposed by accuracy and RT data are a critical factor in limiting the model’s flexibility.

Separating accuracy and caution effects

In the recovery simulations above, detection parameters were held constant across different levels of response caution, representing an experiment in which speed–accuracy emphasis is post-cued after identical encoding conditions. I ran a follow-up simulation to address the question of whether the model can discriminate changes in detection from changes in response caution when both types of parameter are allowed to freely vary. The goal is to explore whether the model can be used as a tool to uniquely estimate memory and caution effects for manipulations that might affect either or both processes. The simulation details were exactly as above, except that different D_{AV-O} and D_{AV-N} parameters were sampled for the two speed–accuracy emphasis conditions, and the model fit to each simulated data set allowed the D_{AV-O} and D_{AV-N} parameters to vary across these conditions. If the model can

discriminate changes in detection and response caution, then the estimated detection effect should closely match the actual effect in the parameters that generated the data without being affected by the extent of the response caution effect (i.e., how much the guessing-time distribution shifts between emphasis conditions). Moreover, the estimated guess-time effect should closely match the actual effect regardless of the extent to which detection changes across conditions. If the model cannot fully discriminate the two processes, then estimates for the effect of one will be affected by changing the other; for example, the estimated increase in detection across conditions might be larger than the actual increase when there is also a large shift in the guess times contributing to increased accuracy for the more cautious condition.

I randomly sampled 1,000 parameter sets, generated data from each parameter set with the same trial numbers as the first simulation, and refit the simulated data with the model. To determine whether the model could estimate changes in both processes across conditions, I evaluated the sizes of the caution effects in D_{AV-O} , D_{AV-N} , and μ_G for each data set. Even with guess times and detection both varying across conditions, the model accurately recovered changes in both processes: the correlation between true and estimated effect sizes was .97 for D_{AV-O} , .97 for D_{AV-N} , and .98 for μ_G .

Moreover, the results showed no indication that changing one process across the caution conditions affected estimates for the effect size of the other process. The six possible pairings of the true effect size in one parameter and the estimated effect size in another parameter had correlations ranging from $-.03$ to $.04$, demonstrating that having a small versus a large effect in guess times did not impact the estimated effect on detection, and vice versa. These results demonstrate that the model can uniquely identify effects on detection and response caution, at least for data that conform to the assumptions of the model.

Single-session simulations

The simulations above have several features that facilitate accurate parameter estimation, like data sets that include many trials per participant and multiple speed–accuracy conditions. The present simulations explore parameter recovery with trial numbers typical of a single-session experiment (80 target and 80 lure trials per participant) and a minimal condition structure for recognition (only targets and lures with no extra conditions). I began by simulating 1000 data sets with the same parameter ranges as the simulations above. Data sets with fewer than ten observations defining any of the empirical RT distributions were excluded from fits, which eliminated 0.8% of the sample. The results showed decent estimation of the detection and guessing parameters, with r^2 values of .56 for D_{AV-O} , .59 for D_{AV-N} , and .67 for g .³ The parameters defining the position of the RT distributions were also estimated relatively well, with r^2 values of .71 for μ_{DO} , .69 for μ_{DN} , and .93 for μ_G . However, parameters defining the spread and shape of the RT distributions were poorly estimated, with r^2 values of .17 for σ_D , .03 for λ_D , .36 for σ_G , and .58 for λ_G . The guess RT parameters have more precise estimates because error trials have to be guesses, so the error RTs place tight constraints on the guess times. In contrast, correct RTs can reflect either guessing or detection, making it more difficult to uniquely define the detection finishing times. In light of these results, I recommend that fits to a “standard” recognition experiment should not be used for research questions that depend on accurate estimation of the RT parameters, but fits to a standard design should be effective for research questions focused on the detection and/or guessing parameters.

In fact, applying the RT model instead of an accuracy-only 2HT model can improve estimation of detection parameters in a standard recognition design even if some of the RT parameters estimate poorly. With accuracy-only data from a

minimal recognition design, the standard 2HT model cannot estimate separate detection parameters for targets and lures. A common practice in this case is to measure accuracy with hits minus false alarms ($H - F$), which measures the detection probability for both targets and lures under the assumption that the two values are equal (Snodgrass & Corwin, 1988). The 2HT version of the discrete-race model does not have to assume equal detection for targets and lures; instead, it can potentially discriminate whether unequal accuracy for the two item types is based on unequal detection probabilities or guessing biases on the basis of the different RT profiles of detection and guessing. To explore this potential advantage of fitting RT data in addition to accuracy, I conducted follow-up simulations in which I systematically manipulated either the target or lure detection rates. I kept the trial numbers at 80 targets and 80 lures and used the same parameter ranges as above for all parameters except the one that systematically varied across simulation runs. In one set of simulations, the detection parameter for targets varied across four levels with ranges of .20–.35, .35–.50, .50–.65, and .65–.80. In another, the detection parameter for lures varied with the same ranges. I simulated 250 data sets for each level and randomly sampled 25 of these data sets to define a simulated experiment. In total, I resampled the data sets 20 times to create 20 simulated experiments that each had 25 simulated participants. From each simulated experiment, I calculated the average of the detection parameters for targets and lures as well as the average $H - F$ measure. Thus, the results will be representative of conclusions drawn from average parameter values in a typical recognition experiment.

Figure 3 shows the results. Plots A–C show simulations in which the target detection parameter varied, with the true average target detection parameters (D_{AV-O}) on the x -axis and average parameter values from the fits on the y -axis. Plots A and B reveal that the RT model successfully distinguished target and lure detection, since the average target detection parameter from fits closely tracked the true values (other than a slight overestimation when the true value was low), whereas the lure detection parameter did not systematically vary across levels of target detection. Plot C shows that the $H - F$ measure does not perform as well. The poor estimation for this measure is based on two properties: (1) It is influenced by lure detection in addition to target detection, so it is relatively insensitive to changes in target detection if lure detection is held constant, and (2) it generally underestimates the proportion of trials with available detection, because the accuracy-only model has no mechanism to accommodate trials in which detection was available but was precluded by a faster guess. Plots D–F show the set of simulations with the lure detection parameter varying (D_{AV-N}). Again, the RT fits correctly identified changes in detection for the two stimulus types, with the lure detection parameter varying across levels of true lure detection but with no

³ These r^2 values are from a model in which the predicted fit parameter values are equal to the true parameter values (i.e., the prediction function is a line with a slope of 1 and an intercept of 0).

Fig. 3 Average detection parameters from simulations without a speed–accuracy manipulation. The x-axes are the mean detection parameters used to generate data, and the y-axes are the mean detection measures, where the means are taken across the 25 simulated participants in each

data set. Plots A–C (vs. D–F) show the results from a set of four simulations in which the target (vs. the lure) detection parameter varied systematically. “H – F” denotes hits minus false alarms

systematic variation in the target detection parameter. Again, $H - F$ tracked the variation in true lure detection fairly poorly.⁴ Thus, fitting RT distributions in addition to accuracy can improve estimation of the detection parameters and distinguish changes in target versus lure detection.

Differences in guessing RTs

For the simulations reported above, I assumed that the finishing-time distributions for guessing were the same for the two responses. Psychologically, this means that the participant waits the same amount of time before making a guess, regardless of whether the guess is “old” or “new.” Assuming equal guess times is reasonable for unbiased responding, but manipulations that introduce an overall bias toward one response have a large effect on RTs; that is, responses consistent with the bias are faster for both correct responses and errors (e.g., Ratcliff & McKoon, 2008).

The model can be easily extended to accommodate such guessing effects by specifying separate guess-time distributions for “old” and “new” responses such that the response consistent with bias tends to have shorter guess times. One possible

implementation would involve a three-way race between detection, guessing “new,” and guessing “old.” However, I propose a simpler model with a mixture of “guess old” and “guess new” trials that have different finishing-time distributions. Psychologically, this means that, on each trial, participants commit to guessing “old” or “new” at some point relatively early in the trial, and they allow different amounts of time for detection to succeed on the basis of which response they plan to guess. If they plan to guess the response consistent with the prevailing bias, then they wait less time for detection to succeed because their guess is likely to be correct anyway. If they plan to guess the response that contradicts the prevailing bias, then they give detection more time to succeed given that the guess will have a high risk of being incorrect.

Predictions from this model can still be defined by the equations above with the following procedure: solve Eqs. 1 and 2 separately for “old”-guess trials and “new”-guess trials using the appropriate distribution for guess finishing times, then take the average of the two trial types weighted by the proportion of “old”-guess trials (g) and the proportion of “new”-guess trials ($1 - g$). To be clear, “old”-guess and “new”-guess trials are defined by the response that the participant *plans* to guess if detection fails, but sometimes detection will succeed fast enough and the response will not actually be a guess. This property means that g cannot be interpreted as the proportion of guesses that receive “old” responses when the finishing-time distributions are different for “old”- and “new”-guess trials. Instead, g is the proportion of trials in which the participant would guess “old” *if they make a guess*.

⁴ Of course, the $H - F$ parameter would perform better if the simulations were restricted to parameter sets in which target and lure detection were equal. However, fits to recognition receiver operating characteristic data indicate that target and lure detection are generally unequal (e.g., Kellen, Klauer, & Bröder, 2013).

For example, say $g = .75$ and finishing times tend to be faster for “old” than for “new” guesses. In this case, guessing has a better chance of finishing before detection on “old”- than on “new”-guess trials. Say that guessing determines the response on .8 of the “old”-guess trials and .6 of the “new”-guess trials. Out of all trials, $.75 \times .8 = .6$ of responses will be “old” guesses, and $.25 \times .6 = .15$ of responses will be “new” guesses. Thus, out of the .75 total guess trials, $.6/.75 = .8$ are “old” responses, a value greater than g (.75).

Strengths and weaknesses of different approaches

Along with the discrete-race model and the Heck and Erdfelder (2016) model, Klauer and Kellen (2018) have developed a version of multinomial processing tree models capable of fitting full RT distributions. These approaches are suited to different research questions, and exploring all of them will be crucial to progress in discrete-state RT modeling. The primary advantage of the discrete-race model is the fact that it incorporates speed–accuracy mechanisms, which the other two models lack. However, the model’s ability to discriminate changes in evidence quality and changes in response caution rely on specific processing assumptions that must be tested, and the discrete-race model also has other limitations relative to the alternative models, as discussed below.

The Klauer and Kellen (2018) model assumes that a sequence of processes with discrete outcomes unfolds serially on each trial, and the model specifies finishing-time distributions for each process in the sequence. Thus, this model can potentially address questions about the duration of individual cognitive processes, such as the length of time needed to detect a studied item. In contrast, the Heck and Erdfelder model associates RT distributions with each possible evidence state without specifying times for component processing leading up to the states. Thus, this model cannot isolate the duration of specific cognitive processes, but it has the advantage of potentially accommodating a wide variety of different processing sequences. The discrete-race model can estimate the length of time from the beginning of a trial until detection succeeds or the participant gives up on detection and responds on the basis of guessing, but it cannot be used to estimate the precise durations of these component processes. For example, the Klauer and Kellen model has a parameter for how long it takes to select which response to guess, whereas the discrete-race model can only estimate how long participants wait before making a guess, which is plausibly a much longer duration (i.e., just because people know what they want to guess doesn’t mean they are ready to give up on detection).

A unique advantage of the Heck and Erdfelder (2016) model is the ability to use nonparametric RT distributions by establishing a set of RT bins and parameterizing the probability

that RTs will fall in each bin. This approach will not work for the discrete-race model, because the model has to have a way to determine which process had an earlier finishing time on every trial, requiring continuous distributions of finishing times for both detection and guess times. The binning approach also cannot be used for the Klauer and Kellen (2018) model, because the predicted RTs are derived by adding component RTs across a number of processing stages. Thus, the latter two approaches must make parametric assumptions about finishing-time distributions, and conclusions based on these models might not be valid if the distributional assumptions are incorrect.

The alternative models also likely differ in how easily they can be adapted to more complex paradigms (and the elaborated discrete-state models associated with these paradigms). Both the Heck and Erdfelder (2016) and Klauer and Kellen (2018) models are presented in a format that can be easily extended to different multinomial processing tree structures (see, e.g., Heck & Erdfelder, 2017, for an application of their model to the recognition heuristic). Such an extension would be more complex for the discrete-race model, given that adding new processes requires decisions about whether all processes can occur in parallel and which processes take precedence in ending the race and determining the response. For example, in a model for both recognition and source decisions (e.g., Batchelder & Riefer, 1990), the time needed for source detection to succeed should reasonably be related to the time needed for item detection to succeed across trials, perhaps even to the extent that the “clock” for source detection doesn’t begin until item detection succeeds. However, the exact method of achieving this relationship is an open question. For another example, in a model with both familiarity and recollection-based retrieval (Brainerd, Gomes, & Moran, 2014), should successful familiarity-based retrieval terminate the race and prompt a response, or should there be extra time to allow recollection-based retrieval to potentially succeed? I will not attempt to answer these questions here; instead, I leave them to be addressed on a case-by-case basis in future research.

Discrete processes and sequential sampling models

Sequential-sampling models have traditionally relied on continuous evidence representations (e.g., Ratcliff & Smith, 2004), but a recent model proposed by Donkin, Nosofsky, Gold, and Shiffrin (2013) combines discrete-state representations with a particular sequential sampling model, the linear ballistic accumulator (LBA) model (Brown & Heathcote, 2008). The model was applied to a visual working memory task in which participants decided whether or not the stimulus at a probed location changed from an earlier presentation. The model has an accumulator for each available response (“change” or “no change”), and a response is triggered when the amount of evidence in an

accumulator reaches a response criterion. The model assumes that a subset of trials are “guess” trials in which the average accumulation rate across trials is equal for the two accumulators and the remaining trials are “memory” trials in which the accumulation rates are determined by the stimulus presented (e.g., the “change” accumulator will tend to have a higher accumulation rate than the “no-change” accumulator when the probe is different than the encoding stimulus). Thus, the distribution of accumulation rates across trials is a discrete mixture of two different trial types.

Although this is similar in some ways to the present approach, there are some key theoretical differences. For example, the discrete LBA model uses continuous evidence to select responses on both memory and guess trials; that is, the information that triggers a decision is the position of the accumulators, a continuous variable. Further, decision makers are assumed to have full access to the continuous information represented in the accumulators, as the model allows the decision criteria to be placed at any value of accumulated evidence and to vary freely across conditions that should affect response caution or response biases. In contrast, decisions in the discrete-race model are based on a limited number of evidence states (three for a 2HT version). In other words, the discrete-race model selects a response based on discrete information, whereas the discrete LBA model is a continuous model with a discrete distribution of trial types.

Of course, this is not a criticism of the discrete LBA model, and it is possible that decisions in some tasks are informed by continuous evidence within trials with a discrete mixture of processes across trials. Indeed, Donkin et al. (2013) show that their model outperforms several alternatives for their visual working memory data. An interesting goal for future work is to compare the Donkin et al. model to models that assume discrete evidence states, such as the discrete-race model, the Heck and Erdfelder (2016) model, and the Klauer and Kellen (2018) model.

Nonrace time

Sequential-sampling models include a parameter for nondecision time, which represents the time needed for basic stimulus processing before the decision process can begin and the time needed to overtly respond after a response alternative has been selected (Ratcliff, 1978). The discrete-race model does not have separate nondecision time parameters, because the finishing-time distributions represent the time between the beginning of the trial and the point at which the participant responds on the basis of detection or a guess. The actual response time for a trial is the finishing time for the earliest process. As such, the finishing-time distributions already include time for initial stimulus processing and motor execution of a response.

However, researchers using the discrete-race model need to consider something I will call the “nonrace” time. In discussing the impact of nonrace time, I will delineate three types of distributions: (1) race-time distributions, which represent the period of time in which the participant is considering which response to make; (2) nonrace-time distributions, which represent any time during a trial in which the participant is *not* considering which response to make; and (3) finishing-time distributions, which are parameterized in the model and represent the sum of race and nonrace times. One possible contribution to nonrace time is any delay between the objective beginning of the trial and the point at which the participant realizes that the trial has begun. Another contribution is the delay between making an irreversible commitment to a response and actually making the overt response. Although the participant may continue to consider which response is correct up to and even past the overt response, consideration is *functionally* over when further information cannot alter the response. For example, a participant who detects that an item was studied after initiating the motor process for guessing “Not Studied” might fail to suppress the motor response and end up responding incorrectly.

Fortunately, the average duration of nonrace processes has no impact on the model’s predictions. If a constant nonrace time is assumed on every trial, then the distributions of race time are just the finishing-time distributions shifted by this constant. This shifting affects both the detection and guess finishing-time distributions equally, so it does not affect which one wins the race on each trial. *Variability* in nonrace time, however, can affect the model’s predictions. To illustrate why variability matters, consider two sets of finishing-time distributions for which the average detection time is faster than the average guess time. In Set A, all of the variability in finishing times comes from the race time. Set B has the same total variability as Set A, but a portion of this variability comes from nonrace time. Since the variance of the finishing times is the sum of the race and nonrace variance, the variability in *race* times must be lower in Set B than in Set A. This decreased variability means that there is less overlap in the detection and guess race-time distributions, and detection is more likely to succeed than for Set A. Although it can theoretically affect model predictions, nonrace variability has a negligible effect as long as the variability in nonrace time is small in relation to the total variability in response times. If the potential effects of variability in nonrace time are a source of concern for a particular research question, then the model can be extended to incorporate this process. In such an extended model, the extent of variability in nonrace times can be directly estimated if there are multiple conditions with different paces of responding. Alternatively, variability in nonrace time could be estimated by evaluating the variability of RTs in a simple reaction time task; that is, a task in which participants make a single response as quickly as possible after the onset of

a stimulus. This task closely matches the components contributing to nonrace time in a decision task; specifically, identifying that a stimulus has appeared and executing a motor response.

Conclusion

The Heck and Erdfelder (2016) model is an important advance that will hopefully encourage more widespread consideration of RT distributions in discrete-state research. Their model will likely prove useful for investigating a variety of research questions in the memory and decision making literature, but it cannot predict the magnitude of speed–accuracy trade-offs. The discrete-race model presented here is similar to the Heck and Erdfelder model, except that guessing can lead to missed opportunities for detection. This competition between detection and guessing produces a speed–accuracy trade-off that places strong constraints on the ability of the model to fit data, as demonstrated by the poor fits to simulated data sets with artificially shifted RT distributions. Fits to unaltered data show that the discrete-race model recovers all parameters accurately with trial numbers typical of multiple-session experiments. The parameters defining the variability and shape of finishing-time distributions estimate poorly with the low trial numbers that are more typical for single-session experiments, but the model accurately recovers separate estimates of target detection, lure detection, and guessing even with low trial numbers.

Author note Preparation of this article was supported by Grant No. 1454868 from the National Science Foundation.

References

- Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, 97, 548–564. <https://doi.org/10.1037/0033-295X.97.4.548>
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6, 57–86. <https://doi.org/10.3758/BF03210812>
- Brainerd, C. J., Gomes, C. F. A., & Moran, R. (2014). The two recollections. *Psychological Review*, 121, 563–599.
- Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, 57, 153–178. <https://doi.org/10.1016/j.cogpsych.2007.12.002>
- Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 484–499. <https://doi.org/10.1037/a0018435>
- Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slots models of visual working-memory response times. *Psychological Review*, 120, 873–902. <https://doi.org/10.1037/a0034247>
- Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, 23, 1440–1465.
- Heck, D. W., & Erdfelder, E. (2017). Linking process and measurement models of recognition-based decisions. *Psychological Review*, 124, 442–471. <https://doi.org/10.1037/rev0000063>
- Hu, X. (2001). Extending general processing tree models to analyze reaction time experiments. *Journal of Mathematical Psychology*, 45, 603–634.
- Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review*, 20, 693–719.
- Kellen, D., Singmann, H., Vogt, J., & Klauer, K. C. (2015). Further evidence for discrete-state mediation in recognition memory. *Experimental Psychology*, 62, 40–53.
- 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>
- Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences*, 109, 14357–14362.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108. <https://doi.org/10.1037/0033-295X.85.2.59>
- Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20, 873–922. <https://doi.org/10.1162/neco.2008.12-06-420>
- Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, 111, 333–367. <https://doi.org/10.1037/0033-295X.111.2.333>
- Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, 116, 59–83. <https://doi.org/10.1037/a0014086>
- Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, 50, 408–424. <https://doi.org/10.1016/j.jml.2003.11.002>
- Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, 9, 438–481. <https://doi.org/10.3758/BF03196302>
- Smith, P. L., & Van Zandt, T. (2000). Time-dependent Poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology*, 53, 293–315.
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117, 34–50. <https://doi.org/10.1037/0096-3445.117.1.34>
- Starns, J. J. (2018). Adding a speed–accuracy tradeoff to discrete-state models (R simulation code). Retrieved from <https://osf.io/dgx46>
- Starns, J. J., Ratcliff, R., & McKoon, G. (2012a). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, 64, 1–34. <https://doi.org/10.1016/j.cogpsych.2011.10.002>
- Starns, J. J., Ratcliff, R., & White, C. N. (2012b). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 1137–1151. <https://doi.org/10.1037/a0028151>
- Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, 21, 641–671. <https://doi.org/10.1080/09541440802205067>