The Falsifiability of Actual Decision-Making Models

Andrew Heathcote[1], E.-J. Wagenmakers[2], & Scott D. Brown[1]

[1]School of Psychology, University of Newcastle, Australia
[2]Department of Psychology, University of Amsterdam

Correspondence should be addressed to:

andrew.heathcote@newcastle.edu.au
School of Psychology
The University of Newcastle, Australia
Callaghan, 2308, NSW, Australia

Abstract

Jones and Dzhafarov (2013) provide a useful service in pointing out that some assumptions of modern decision-making models require additional scrutiny. Their main result, however, is not surprising: if an infinitely complex model were created by assigning its parameters arbitrarily flexible distributions, this new model would be able to fit any observed data perfectly. Such a hypothetical model would be unfalsifiable. This is exactly why such models have never been proposed in over half a century of model development in decision-making. Additionally, the main conclusion drawn from this result –that the success of existing decision-making models can be attributed to assumptions about parameter distributions– is wrong.

Models of Speeded Decision-Making Are Highly Constrained

Modern decision-making models have been used to uncover new insights about brain and behavior in dozens of different paradigms requiring choice among two (e.g., Ratcliff, & McKoon, 2008) or more (e.g., Busemeyer & Diederich, 2002) options. All modern models share a common and simple structure: they assume that evidence is gradually accumulated from the environment and a decision is made whenever the evidence reaches a threshold amount (e.g., the diffusion model: Ratcliff 1978, Ratcliff & Tuerlinckx, 2002; and the linear ballistic accumulator model: LBA, Brown & Heathcote, 2008). In their simplest forms, the models have three central parameters: the "drift rate" which measures how fast evidence accumulates; a "threshold" which measures how much evidence needs to accumulate before a decision is made; and "non-decision time", which measures how much time is taken up by processes other than decision-making, such as the time taken to push a response button.

Over the past fifty years (since Stone, 1960), the most basic versions of these models have been proven incomplete. For example, the earliest version of the model, described above, successfully predicted the general shape of response time distributions, and the tradeoff between urgent vs. cautious decisions, and even some fine details of response time distributions such as hazard rates. However, these early versions made such highly constrained predictions that they were unable to accommodate patterns of differing speed between incorrect and correct responses; patterns which were regularly observed in data when participants are told to respond quickly (e.g., Ratcliff & Rouder, 1998). These limitations have informed model development, and modern response time models include two key elements that address these earlier limitations: they assume that the drift rate varies randomly from decision to decision, and also that the starting point of the evidence accumulation process varies randomly from decision to decision. The distributions assumed for the trial-to-trial variability of the drift rate and start point have always been simple forms with one additional free parameter. The interested reader will find a detailed history of the development of response time models, and the implications for model constraint and falsifiability, in the supplementary material to this comment.[1]

Jones and Dzhafarov's (2013) Central Result: Infinitely Complex Models Can Be Unfalsifiable

Jones and Dzhafarov's (2013) main result extends earlier work by Townsend (1976), Marley and Colonius (1992), and Dzhafarov (1993). The key idea is that, if one allows unbounded complexity and freedom in the across-trial distribution of drift rates, the model can perfectly fit any and all data sets. This is intuitively obvious – for example, if the threshold was set at 1.0 (i.e., 1 unit of evidence required to trigger a decision) and the drift rate distribution happened to perfectly invert the observed data (i.e., each observed RT corresponded to a drift rate sample of 1/RT), then the "predicted" data from the model would perfectly match the observed data. Jones and Dzhafarov's (2013) theorems formalize this intuition.

[1] The supplement addresses in detail specific claims about (1) a lack of empirical support for the LBA and diffusion models; (2) the flexibility and testing of the LBA and diffusion models; (3) positions held by authors of evidence accumulation models about the status of different assumptions made by their models; and (4) the supposed special status of distributional assumptions over other assumptions.

It is not surprising that allowing infinite complexity in a model makes it unfalsifiable. This is not unique to decision-making or response time models, but applies to all models. For example, it is trivial to see that signal detection theory can perfectly fit any pattern of hit and false alarm rates, if one allows unbounded freedom in how the parameters (*d'* and bias) change across conditions. Similarly, a linear regression model with an unlimited number of predictors will fit any data at all.

This kind of result does not make signal detection theory or linear regression any less useful; rather it means that researchers should limit the complexity of models instantiated within these frameworks. This is exactly what has always happened in practice with decision-making models. Researchers have never proposed arbitrary and complex distributions for across trial variability, but have always restricted themselves to highly constrained and extremely simple distributions, such as the uniform distribution (for start points) or the Gaussian distribution (for drift rates). The central result of Jones and Dzhafarov (2013), while entirely correct for hypothetical, unrealistic models, applies to no actual model that has ever been proposed.

It is true that the particular forms of the across-trial variability parameters in decision-making models (Gaussian and uniform) were originally chosen arbitrarily, for practical and not theoretical reasons. However, since these forms were chosen in the original model development, they have been fixed in the dozens or hundreds of applications of the models since. This constitutes a rigorous test of the models. The simple forms chosen for across-trial variability result in falsifiable models that could easily have failed to fit new data, many times over, but this has not happened. In other words, if the precise shape of the across-trial distributions had been crucial for the model's success in fitting, one would expect these shapes to differ from experiment to experiment (or even across subjects or conditions) in order to accommodate the idiosyncrasies of different data. In reality, the models have managed to provide an excellent account of hundreds of data sets and thousands of participants using exactly the same distributional shapes.

What are the Implications for Real Decision-Making Models?

An important conclusion drawn from Jones and Dzhafarov's (2013) main result and stated prominently on the front page is that "the explanatory or predictive content of these models is determined … by distributional assumptions". This is a mistaken conclusion that does not follow from the central result. Jones and Dzhafarov showed that a new model formed by allowing infinite complexity in the drift rate distribution could be unfalsifiable. This does not imply the standard model's falsifiability was entirely due to its assumptions about drift rate.

The problem with concluding that drift rate assumptions are the key to the standard models' falsifiability is that allowing infinite flexibility in drift rate distributions is *sufficient* to create an unfalsifiable model, but it is not *necessary*. There are almost as many ways to make a model unfalsifiable as there are parameters in the model: almost any parameter, if endowed with infinitely flexible distributional assumptions, can result in a new model that is unfalsifiable. For example, if one allowed infinite complexity in the distribution of non-decision time, the model could fit any response time data at all (e.g., by assuming that the distribution of non-decision time was exactly the observed data distribution, and that the time taken for the decision process was zero). Similarly, if one allowed infinite complexity in the distribution of start points, the model could fit any data at all (e.g., by assuming a constant drift rate of 1.0, a threshold of zero, zero non-decision time and a start point distribution that was exactly the negative of the observed data). Similar arguments can be made about most parameters of a model, from the shape of the evidence accumulation curve to the location of the threshold.

These trivial examples illustrate the mistake of according special status to the drift rate assumptions (or any single assumption). Rather, a model's predictive content is determined by all of its assumptions together, and it is wrong to assign special status to particular assumptions about across-trial variability. Confusingly, Jones and Dzhafarov appear to come to exactly this same conclusion, but rather less prominently (on p.48): "one needs to consider both distributional and structural assumptions jointly". Our supplementary material further explores the tension in Jones and Dzhafarov's article between the idea that all model assumptions matter equally, vs. the idea that one particular model assumption carries all the predictive power.

Conclusions

In a provocative and mathematically sound article, Jones and Dzhafarov (2013) have proposed hypothetical response time models with infinite complexity in distributional shape, and shown that these models are unfalsifiable. This conclusion corroborates current practice that eschews such models in favor of models that are highly constrained in distributional shape. Despite their constraints, these realistic models have consistently yielded good fits to many data sets across a range of different paradigms, without changes in the distributional assumptions across hundreds of experiments and thousands of participants. The empirical success of realistic, constrained models shows that the explanatory and predictive content of realistic response time models is not determined by distributional assumptions.

In summary, Jones and Dzhafarov (2013) are right to point out that parameter distribution assumptions of decision-making models deserve scrutiny, but that scrutiny has a long history (e.g., Link & Heath, 1975) with increased recent activity (e.g., Heathcote & Love, 2012, Ratcliff, 2013, see supplementary material for more details). However, we conclude that, although Jones and Dzhafarov's main results are important for hypothetical, infinitely complex models that have never been proposed, they are much less relevant for the realistic models that are used in actual practice.

References

Brown, S.D., & Heathcote, A. (2008) The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology, 57*, 153-178

Busemeyer, J. R., & Diederich, A. (2002). Survey of decision field theory. *Mathematical Social Sciences*, *43*(3), 345–370.

Heathcote, A. & Love, J. (2012) Linear deterministic accumulator models of simple choice. *Frontiers in Cognitive Science*, 3, 292.

Jones, M. & Dzhafarov, E.N. (2013). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction Time. *Psychological Review.*

Link, S.W. & Heath, R.A. (1975). A sequential theory of psychological discrimination, *Psychometrika*, *40*, 77-105.

Marley, A. A. J., & Colonius, H. (1992). The "horse race" random utility model for choice probabilities and reaction times, and its competing risk interpretation. *Journal of Mathematical Psychology*, *36*, 1-20.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59.

Ratcliff, R. (2013). Parameter variability and distributional assumptions in the diffusion model. *Psychological Review*, *120*, 281-292.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. Neural Computation, 20(4), 873–922.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, 9, 347–356.

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*(3), 438–481.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260.

Townsend, J. T. (1976). Serial and within-stage independent parallel model equivalence on the minimum completion time. *Journal of Mathematical Psychology*, *14*, 219-238.

Supplement to "The Falsifiability of Actual Decision-Making Models"

Jones and Dzhafarov (2013, hereafter J&D) reports work expanding the scope of previous results (Dzhafarov, 1993; Marley & Colonius, 1992; Townsend, 1976) on the universality of independent race models, models which assume that choices are made by accumulating evidence about different potential responses. Universality in this context refers to the ability of a model to account for any observed pattern of behavior in discrete choice experiments as characterized by response probabilities and response time (RT) distributions. J&D draws an alarming conclusion with respect to two evidence accumulation models that have been widely applied in psychology and the neurosciences: "Although the diffusion and LBA models have been highly successful in fitting data from a variety of task domains (e.g., Brown & Heathcote, 2008; Ratcliff & Smith, 2004), this success does not imply any support for the theoretical or structural assumptions of these models." (p.47) Does this mean J&D have shown that psychologists and neuroscientists have been misled, and that there is no empirical support for the LBA and diffusion[2] models? We believe the answer is clearly no.

Although the mathematical results in J&D are clear and precise, the inferences drawn from them are often unclear and imprecise in a way that sometimes misrepresents positions held by the authors of the diffusion, LBA, and other evidence accumulation models, and which has the potential to mislead future research. These problems go

[2] A variety of different specific models have been proposed that share the assumption evidence diffuses (i.e., varies from moment-to-moment in the continuous case or sample-to-sample in the discrete case), beginning with the seminal work of Stone (1960). The diffusion model as studied in J&D is elaborated with the idea of three types of trial-to-trial variability in parameters. We follow J&D in calling a diffusion model with uniform variability in non-decision time and the starting evidence accumulation value, and Gaussian variability in the mean rate of accumulation – the most widely used diffusion model over the last decade – as "the" diffusion model.

beyond the claim just recounted from J&D about 1) a lack of empirical support for the LBA and diffusion models; they also encompass unsupported and/or incorrect inferences about 2) the flexibility and testing of the LBA and diffusion models, 3) positions held by authors of evidence accumulation models about the status of different assumptions made by their models and 4) claims about the special status of distributional assumptions over other assumptions that together constitute these models. The distributional assumptions referred to in J&D are assumptions about the mathematical form of variability in model parameters from one trial to the next.

The Special Status of Distributional Assumptions

We examine the claim about the special status of distributional assumptions first as J&D is conflicted on this point. In the abstract it says: "the explanatory or predictive content of these models is determined not by their structural assumptions, but rather by distributional assumptions" (p.1). However, in the discussion it says: "one needs to consider both distributional and structural assumptions jointly" (p.48). We agree with the second statement, that it is the joint properties of all assumptions that are critical, and worry that the contradiction will confuse careful readers and, because of the prominence of the abstract, mislead casual readers. Distributional assumptions do not have a special status because it is only when they are combined with structural assumptions – assumptions about the way that evidence is processed and a choice made – that a model could make predictions about behavior.

If anything it is a structural assumption, *evidence accumulation to a threshold*, which has a special status. This assumption states that one or more types of evidence are accumulated over time, that each type has one or more associated thresholds, and that a response associated with a threshold is selected if its associated evidence total is the first to satisfy any threshold. Participants can set the threshold strategically to modulate response caution (i.e., to trade speed for accuracy or vise versa). One might see this assumption's status as special because most of the modern literature on modeling choice RT, including J&D, does not question it. In contrast, the generality of the claim in J&D about the primacy for distributional assumptions is contradicted by Dzhafarov (1993). It shows universality for an arbitrary choice of threshold distribution by allowing complete flexibility in the form of a deterministic accumulation process. Hence, in Dzhafarov, who studied a type of evidence accumulation to threshold model called a Grice model (e.g., Grice, 1968, 1972), the form of the accumulation process (e.g., linear, exponential, sigmoid etc.) determines the model's explanatory or predictive content, not distributional assumptions.

In the context of the LBA model, which assumes linear deterministic accumulation, J&D shows that universality results by allowing complete flexibility in the distribution of rates of accumulation. In contrast to this model, which J&D names the gLBA, the LBA model is not universal because it assumes a specific form for the rate distribution (Gaussian) and a specific from for the distribution of points at which evidence accumulation starts (uniform). J&D states "universality of the gLBA … is a straightforward mathematical fact. However, its implications for the standard LBA seem to have been overlooked. Specifically, this result implies that the predictive power of the standard LBA lies in its assumptions regarding growth-rate and start-point distributions (which have heretofore been treated as implementation details)[3]" (p.18). Putting to one side that we find it unsurprising that implications for the standard LBA have been "overlooked" (after all the "mathematical fact" was first proved in J&D, and it is about a model that is not the standard LBA model but instead about a more general model that never been proposed or used by anyone but J&D), the implication drawn is clearly false; it is the joint effects of all assumptions in the standard LBA model, including linear accumulation, that determines its explanatory and predictive content.

In summary, J&D states that its "core message … is that, when a modeling framework is universal, the predictive content of any model expressed in that framework lies in whatever falsifiable assumptions that model makes". (p.33). We agree with this statement but not the implication that is immediately drawn: "For the standard LBA and diffusion models, these assumptions are the forms of the probability distributions for growth rate (Gaussian), starting points (uniform), and nondecision time (uniform, for the diffusion model)" (p.33). In contrast, we think that it is the joint effect of all of the model's assumptions that determine its explanatory and predictive content.

What does the evidence imply?

[3] J&D also examines the effect on universality of two "selective influence" assumptions about the way model parameters can vary across stimuli and experimental conditions. J&D notes that of "the two selective influence assumptions made by the standard diffusion and LBA models, the first has no impact on universality, and the second is logically suspect and perhaps even psychologically unlikely." (p.31). Recent research we have preformed applying both the diffusion and LBA (Rae, Heathcote, Donkin & Brown, submitted), as well as other research using only the diffusion, including papers cited in J&D as well as other work (Starns, Ratcliff, & White, 2012), supports the doubts raised on psychological grounds, and so we do not discuss selective influence assumptions further and omit mention of them in quotes for this reason.

What then of the claim in J&D that the success of the diffusion and LBA models in fitting data from a wide variety of task domains “does not imply any support for the theoretical or structural assumptions of these models” (p.47)? We think that this success *does* imply strong support for the *combination* of assumptions in each model (i.e., the model’s *joint* assumptions). Taken together, the assumptions make up models that could have been falsified. However this did not happen. Instead, the models provided good (but not perfect) descriptions of fine-grained details of behavior. They also provided a coherent account of that behavior in terms of latent psychological constructs with a clear correspondence to the model’s parameter estimates. That said, we do agree that the universality results in J&D have a more limited implication, that there is no necessary support for any assumption taken in isolation.

How important is it that we find support for isolated assumptions, or perhaps even subsets of assumptions? J&D states: “Whenever a cognitive model provides a good account of empirical data, it is critical to understand which of its assumptions are responsible for its predictive success. Such understanding is important for theoretical progress and for generalizing to other paradigms or domains.” (p.46). In contrast, we do not see such an understanding as critical, and think that it is rarely if ever achievable in models of sufficient complexity to provide a realistically detailed account of empirical data. J&D goes on to state: “the assumptions of most formal models can be roughly divided into ones corresponding to theoretical principles the model is meant to embody, and technical details that are necessary to generate quantitative predictions but are chosen without theoretical consideration and can be modified or dispensed with as need arises.” (p.48). We think that J&D provides a valuable service in extending the already existing demonstrations by Townsend (1976), Marley and Colonius (1992) and Dzhafarov (1993) that this rough division is indeed very rough in the context of evidence accumulation models, and so probably not a differentiation that is worth making in any strong sense. However, we see only a few examples in the past literature where this differentiation has been made, and do not believe that the field has been misled on this point.

A very different perspective is offered in J&D. One of its concluding statements is that: “Mathematical modeling has produced models that often yield impressive fits to these data with relatively few free parameters. Nevertheless, the theoretical implications of these modeling results are far less certain than they have been made out to be. As we have shown here, the models’ predictions derive not from their structural assumptions but from technical aspects that have been considered irrelevant details”. (p.50) The “technical aspects” referred to are distributional assumptions, which are claimed to have “received little attention or justification” (p.13), a claim that is later repeated and elaborated, saying they have been treated as ‘being “merely” implementation details and not part of the underlying theory.’ (p.29). In support of distributional assumptions having been treated as “irrelevant” and “mere” implementation details J&D quotes Brown and Heathcote (2008, p.160) saying that they: “chose the normal distribution for practical reasons, because it is both tractable and conventional.”. J&D reasons: “If these distributional assumptions are only a matter of convenience and tradition, then they should not be considered a critical part of the psychological theory.” (pp.15-16).

We think this reasoning in J&D confuses the origin of assumptions and their place in a model. Both the diffusion and LBA models also make a linear accumulation assumption, so that assumption might also be characterized as conventional. The linear assumption in both models is also key for mathematical tractability. Indeed, Brown and Heathcote (2008) explicitly motivated the linear assumption as about tractability with reference to their BA model (Brown & Heathcote, 2005), where the combination of the same distributional assumptions with nonlinear accumulation is not mathematically tractable. The LBA assumption of an independent race, with each racer having its own response threshold, has its basis in a long history of independent race model applications (e.g., Vickers, 1979) and in mathematical tractability (Marley & Colonius, 1992). Similarly, the assumption of deterministic accumulation facilitates mathematical analysis and also has a long history (e.g., Grice, 1968, 1972; Carpenter, 1981). In short, all of the LBA models assumptions can be motivated as "tractable and conventional"; it is the combination of these assumptions that makes the model original, and their joint effect that makes it testable. More generally, once assumptions have been chosen, for whatever reason, a model stands or falls on its ability to make specific predictions that can be subject to empirical verification: clearly both the LBA and diffusion models pass this test. We believe the reasons for which assumptions are chosen matter only to the degree that they bring with them testable predictions.

On the status of distributional assumptions

We agree with J&D that distributional assumptions are important and deserve close scrutiny. Such scrutiny was the major motivation of Heathcote and Love (2012), who defined the class of *deterministic accumulator* (DA) models with the LBA as a special case. In DA models the time to threshold for one accumulator is a ratio of random variables representing the distance from start point to threshold (i.e., response caution) in the numerator and accumulation rate in the denominator. This class is equivalent to the gLBA that J&D shows in universal. Heathcote and Love focused on the properties of a new specific (i.e., non-universal and falsifiable) model in this class where both distributions have a Lognormal form. Reflecting the fact that they did not see this change as a mere implementation detail, they gave the resulting model a different name, the Lognormal race (LNR). The LNR has even greater mathematical tractability than the LBA, particularly with respect to the case of correlated evidence and Bayesian estimation (Rouder, Province, Morey, Gómez & Heathcote, submitted), as the ratio of Lognromal variables it itself Lognormal.

In the current context, the latter property of the LNR model as an interesting consequence; the parameters of the numerator and denominator distributions combine additively. In the LNR model, therefore, and in contrast to the LBA model, response caution and accumulation rate effects are not separately identifiable without additional assumptions the about a selective influence on these parameters of experimental manipulations. Heathcote and Love (2012) concluded that distributional assumptions are important in the class of DA models because they can determine the identifiability of effects on response caution and rate parameters. Acting on the implication that it is important investigate and test distributional assumptions, they compared the fit of the LNR and LBA models to data reported by Wagenmakers, Ratcliff, Gómez and McKoon (2008). Although both performed well the LBA model did slightly better based on model selection criteria taking account of the number of estimated parameters. Clearly, at least in this case, distributional assumptions were not treated as mere implementation details.

Ratcliff (2013) investigated the effect of variations the distributional assumptions on the standard diffusion model. It concludes that the psychological implications of standard diffusion model parameter estimates were largely invariant under mild misspecifications in start point and rate distributions (i.e., when it was fit to simulated data generated with mildly different distributions). However, we believe the limitation to mild changes is important to emphasize. For example, such invariance was not found for a more marked change in the distribution of non-decision time (from uniform to exponential). It is also true that invariance would fail for more marked changes in other two trial-to-trial distributions in the standard diffusion model, and change in ways that could falsify the model. For example, as the range of the uniform start-point distribution shrinks to zero (i.e., in the limit of a change to no start-point variability) errors responses cannot be faster than correct responses. Similarly, as the standard deviation of the Gaussian distribution of accumulation rates approaches zero error responses cannot be slower than correct responses. Both faster and slower errors are observed empirically (e.g., under instructions to respond quickly vs. accurately respectively, Ratcliff & Rouder, 1998).

More broadly, we would argue that a balanced assessment of the longer-term history of evidence accumulation models reveals a healthy development and testing of distributional assumptions. For example, Link and Heath (1975) shows that, in the absence of any trial-to-trial variability, a wide range of assumptions about the form of moment-to-moment variability (including the Gaussian assumption made by the standard diffusion model) leads to equivalence in distribution correct and error RT, but for other assumptions this was not the case. Laming (1968) introduced trial-to-trial variability in diffusion start points in order to account for fast errors and Ratcliff (1978) introduced variability in the mean rate of accumulation to account for slow errors. Ratcliff and Tuerlinckx (2002) introduced trial-to-trial in the time to complete non-decision processes and Ratcliff, Gómez and McKoon (2004) presented evidence for the necessity of this addition for the success of a diffusion model of the lexical decision task when an additional selective influence assumption is imposed. Attention has also been giving to the potential for removing sources of variability in the context of different sets of joint assumptions. For example, Brown and Heathcote (2005) simplified Usher and McClelland's (2001) evidence accumulation model by removing moment-to-moment noise. They justified the simplification empirically based on its ability to fit a wide range of benchmark data.

In summary, there are many examples of investigations that have addressed distributional assumptions. Much of this debate has centered on whether a constant parameter value suffices or whether a parameter needs to be allowed to vary randomly from trial-to-trial. However, investigations have also addressed the subtler questions related to distributional form, including the sensitivity of predictions to differences in the form of trial-to-trial distributions (Ratcliff, 2013), and the appropriate form of both trial-to-trial (Heathcote & Love, 2012) and moment-to-moment (Link & Heath, 1975) distributions. Although we think that this summary makes it clear that distributional assumptions have generally been treated as more than mere implementation details, we also agree with both Heathcote and Love (2012) and J&D that distributional assumptions continue to deserve further scrutiny.

Flexibility and testing of standard models

Are the standard models overly flexible and not subject to rigorous tests that could potentially falsify them? J&D maps these models onto a universal Grice model, where all of the flexibility in the models is concentrated into the form of evidence growth functions and concludes: “the Grice representations offer a new perspective on the predictive constraints in the diffusion and LBA models arising from their parametric and selective influence assumptions. They show that the flexibility identified in the previous two subsections enables the diffusion and LBA models to match most aspects of the data in a post hoc manner.” (p. 45). We believe that our recounting of the history of the development of these models shows that their ability to fit data is anything but post hoc; instead it is based on a careful cumulative development of supporting evidence, with additional flexibility added (and sometimes removed) only when that is clearly justified.

J&D goes on to state: “That is, had these features taken on different values, the models could have matched them as well, by using different parameter values.” (p.35). For the standard diffusion model Ratcliff (2002) shows that this statement is wrong. Ratcliff notes that: “Researchers working with stochastic models for RT and accuracy have known that their models are inflexible— that is, that there are many possible patterns of data the models cannot fit.” (p.286). This claim is substantiated with simulations that show, amongst other things, that no parameter values of the standard diffusion model allow it to provide an adequate fit to symmetric (Gaussian) and highly skewed RT distributions. Figure 1 shows that the LBA also provides a poor fit to a Gaussian RT distribution and an RT distribution with unrealistically large skew and sharply increasing leading edge similar to Ratcliff's example. Perhaps R&D did not mean to include these cases in "most aspects of the data", but we do not believe this distinction is made sufficiently clear. Ratcliff's simulations and Figure 1 underline a fact that it is important to emphasize in order to avoid readers of J&D being mislead: the diffusion and LBA models are empirically falsifiable because they cannot fit any pattern of data by post hoc parameter adjustment.

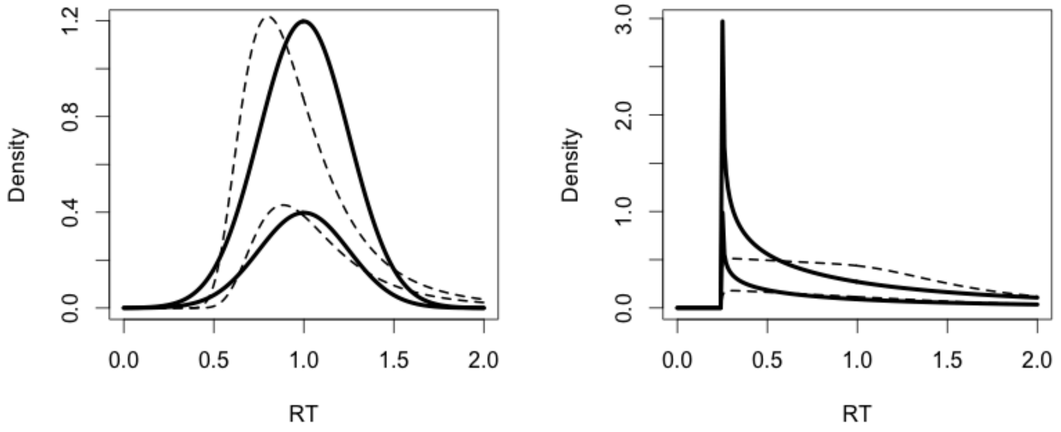

Figure 1. Thick sold lines are simulated Gaussian (left panel, mean = 1, SD = 0.25) and Weibull (right panel, shape parameter = 0.75, scale = 1, shift = 0.25) densities. Thinner dashed lines are best fitting LBA densities. The larger simulated densities have an area of 0.75 and the smaller and area of 0.25, representing 75% correct responses and 25% error responses respectively (error and correct distributions have the same parameters).

Finally, J&D claim that the standard way of testing evidence accumulation models does not provide an inherently stringent test. It says: "An advantage frequently cited for stochastic-accumulation models of speeded choice is that they jointly capture the overall response probabilities and the RT distribution associated with each response. The implied suggestion is that there is some coupling among these measures inherent in the models, so that fitting all of them simultaneously is a more stringent test. The universality results imply there is no such coupling, other than that arising from the parametric … assumptions … a universal model can fit all of these measures simultaneously and independently.” (p.28). We disagree; there is a coupling that arises between these measures in standard models and it arises from the joint effect of all assumptions, not just distributional assumptions. For the actual falsifiable models of choice RT that have been proposed the stringent test provided by determining whether they can jointly capture the response probabilities and RT distributions associated with each response is essential. As Ratcliff (2002) notes “the diffusion model can almost always, for any single experimental condition, fit the condition’s accuracy value and two mean RTs, one for correct and one for error responses” (p.286). Our experience is that the same is true for the LBA, whereas, in contrast, there are clearly patterns of RT distribution and response probability that cannot be accommodated by either model.

Supplement References

Carpenter, R. H. S. (1981). Oculomotor procrastination. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), Eye movements: cognition and visual perception (pp. 237–246). Lawrence Erlbaum.

Dzhafarov, E.N. (1993). Grice representability of response time distribution families. *Psychometrika*, *58*, 281-314.

Grice, G. R. (1968). Stimulus intensity and response evocation. *Psychological Review*, *75*, 359-373.

Grice, G. R. (1972). Application of a variable criterion model to auditory reaction time as a function of the type of catch trial. *Perception & Psychophysics*, *12*, 103-107.

Heathcote, A. & Love, J. (2012) Linear deterministic accumulator models of simple choice. *Frontiers in Cognitive Science*, 3, 292.

Jones, M. & Dzhafarov, E.N. (2013). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction Time. *Psychological Review.*

Laming, D.R.J. (1968). Information theory of choice reaction time. London: Academic Press.

Link, S.W. & Heath, R.A. (1975). A sequential theory of psychological discrimination, *Psychometrika*,*40*, 77-105.

Marley, A. A. J., & Colonius, H. (1992). The “horse race” random utility model for choice probabilities and reaction times, and its competing risk interpretation. *Journal of Mathematical Psychology*, *36*, 1-20.

Rae, B., Heathcote, A., Donkin, C. & Brown, S.D. (submitted). Emphasizing speed can change the evidence used to make decisions

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59.

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

Ratcliff, R. (2013). Parameter variability and distributional assumptions in the diffusion model. *Psychological Review*, *120*, 281-292.

Ratcliff, R., Gómez, P., & McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological Review*, *111*, 159–182.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, 9, 347–356.

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*(3), 438–481.

Rouder, J. N., Province, J. M., Morey, R.D., Gómez, P. & Heathcote, A. (submitted). The Lognormal Race: A cognitive-process model of choice and latency with desirable psychometric properties.

Starns, J. J., Ratcliff, R., & White, C. N. (2012). 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(5), 1137–1151.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260.

Townsend, J. T. (1976). Serial and within-stage independent parallel model equivalence on the minimum completion time. *Journal of Mathematical Psychology*, *14*, 219-238.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*, 550.

Vickers, D. (1979). Decision Processes in Visual Perception, Academic, New York.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*(7), 1206–1220.

Wagenmakers, E.-J., Ratcliff, R., Gómez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, 58, 140–159.