Meta-Psychology, 2020, vol 4, MP.2019.2238
https://doi.org/10.15626/MP.2019.2238
Article type: Commentary
Published under the CC-BY4.0 license

Open data: N/A
Open materials: N/A
Open and reproducible analysis: N/A
Open reviews and editorial process: Yes
Preregistration: N/A

Edited by: E. M. Buchanan
Reviewed by: D. Navarro, S. Farrell
Analysis reproduced by: N/A
All supplementary files can be accessed at OSF:
https://osf.io/2mdfj/

What factors are most important in finding the best model of a psychological process? Comment on Navarro (2019)

Nathan J. Evans
Department of Psychology, University of Amsterdam, The Netherlands
School of Psychology, University of Queensland, Australia

Abstract

Psychology research has become increasingly focused on creating formalized models of psychological processes, which can make exact quantitative predictions about observed data that are the result of some unknown psychological process, allowing a better understanding of how psychological processes may actually operate. However, using models to understand psychological processes comes with an additional challenge: how do we select the best model from a range of potential models that all aim to explain the same psychological process? A recent article by Navarro (2019; Computational Brain & Behavior) provided a detailed discussion on several broad issues within the area of model selection, with Navarro suggesting that *"one of the most important functions of a scientific theory is ... to encourage directed exploration of new territory"* (p.30), that *"understanding how the qualitative patterns in the empirical data emerge naturally from a computational model of a psychological process is often more scientifically useful than presenting a quantified measure of its performance"* (p.33), and that "quantitative measures of performance are essentially selecting models based on their ancillary assumptions" (p.33). Here, I provide a critique of several of Navarro's points on these broad issues. In contrast to Navarro, I argue that all possible data should be considered when evaluating a process model (i.e., not just data from novel contexts), that quantitative model selection methods provide a more principled and complete method of selecting between process models than visual assessments of qualitative trends, and that the idea of *ancillary* assumptions that are not part of the core explanation in the model is a slippery slope to an infinitely flexible model.

Keywords: Model selection, Science, Quantitative model comparison, Cognitive models.

Over the past several decades, psychology research has become increasingly focused on creating formalized models of psychological processes (e.g., Ratcliff, 1978; Brown & Heathcote, 2008; Usher & McClelland, 2001; Nosofsky & Palmeri, 1997; Shiffrin & Steyvers, 1997; Osth & Dennis, 2015). These *process models* are created by taking verbal explanations of a process, and formalizing them with an exact mathematical functional form. Process models make exact quantitative predictions about observed data that are the result of some unknown psychological process, and by attempting to see which models can best account for these observed data, we can better understand how this unknown process may actually operate. However, using models to understand psychological processes comes with an additional challenge: how do we select the best model from a range of potential models that all aim to explain the same psychological process? This is an area of research known as model selection (Myung & Pitt, 1997; Myung, 2000; Myung, Navarro, & Pitt, 2006; Evans, Howard, Heathcote, & Brown, 2017; Evans & Annis, 2019), and is subject to ongoing debate, both at broad levels (e.g., qualitative methods [Thura, Beauregard-Racine, Fradet, & Cisek, 2012] vs. quantitative methods [Evans, Hawkins, Boehm, Wagenmakers, & Brown, 2017]) and specific levels (e.g., Bayes factors [Gronau & Wagenmakers, 2019a] vs. out of sample prediction [Vehtari, Simpson, Yao, & Gelman, 2019]).

A recent article by Navarro (2019) provided a detailed discussion on both specific and broad issues within the area of *model selection*. Although this article was a comment on the specific critique of Bayesian leave-one-out cross-validation by Gronau and Wagenmakers (2019a), Navarro (2019) also made several broader points on the philosophy of modelling, and how we should evaluate these formalized theories. These broader points made by Navarro (2019) appear to have been the most impactful part of the entire debate so far, with Navarro's article currently (as of the 18th of January, 2019) having over 3,400 downloads and 122 shares, compared to the 678 downloads and 11 shares of the original article by Gronau and Wagenmakers. In general, Navarro (2019) suggested that 1) *"one of the most important functions of a scientific theory is ... to encourage directed exploration of new territory*" (p.30), 2) *"understanding how the qualitative patterns in the empirical data emerge naturally from a computational model of a psychological process is often more scientifically useful than presenting a quantified measure of its performance"* (p.33), and 3) *"quantitative measures of performance are essentially selecting models based on their ancillary assumptions"* (p.33). Although Gronau and Wagenmakers (2019b) provided a reply to all three commentaries made on their original article (Vehtari et al., 2019; Navarro, 2019; Chandramouli & Shiffrin, 2019), their response mostly focused on replying to Vehtari et al. (2019) with further limitations of Bayesian leave-one-out cross-validation. Their section replying to Navarro (2019) briefly mentioned that quantitative methods are useful as *"the data may not yield a clear result at first sight"* (p.42), but focused on a more specific point, which was regarding how useful simple examples (or in the more critical terms of Navarro, *"toy examples"*) are in assessing the robustness of analysis methods.

Here, I provide a critique of some of Navarro's broader perspectives, such as the function of scientific theories, the importance of qualitative patterns compared to precise quantitative performance, and the distinction between core and ancillary assumptions. Specifically, I argue that 1) all possible data should be considered when evaluating a process model (i.e., not just data from novel contexts), 2) quantitative model selection methods provide a more principled and complete method of selecting between process models than visual assessments of qualitative trends, and 3) the idea of *ancillary* assumptions that are not part of the *core* explanation in the model is a slippery slope to an infinitely flexible model. However, before providing my arguments, I would like to note that my arguments only reflect one side of the contentious debate over how models of psychological processes should be evaluated – just as Navarro's arguments only reflected another side of the debate. Therefore, I believe that researchers should read both Navarro (2019) and my comment with an appropriate level of scrutiny, in order to gain a more complete perspective on the broad issues within this debate and decide how they believe models of psychological processes should be evaluated.

What is the most important function of a process model?

Most of Navarro's (2019) perspectives regarding model selection appear to be based around one key underlying factor: what is the most important function of a scientific theory (or in these cases, a formalized process model that encapsulates a scientific theory)? From Navarro's perspective, *"one of the most important functions of a scientific theory is ... to encourage directed exploration of new territory"*. More specifically, in the section *"Escaping mice to be beset by tigers"* (p.30–31) Navarro appears to suggest that good process models – the models that provide better representations of the unknown psychological process that we wish to understand – are the ones that make accurate predictions about novel contexts, and these novel predictions are how process models – and more generally scientific theories - should be evaluated. Although Navarro's perspective may be a popular one among many researchers, I believe that this is only a single perspective on a contentious issue. Within this section I present a different perspective on what the most important function of a process model is, and how we should determine the best model(s) of a process: that 1) the most important function of a process model is to explain the unknown psychological process as well as possible, 2) process models should be evaluated based upon all known data, and 3) the most principled way of making these evaluations is using quantitative model selection techniques.

First and foremost, I agree with Navarro (2019) that encouraging directed exploration of new territory can be a useful function of a scientific theory, and making predictions about novel contexts (in Navarro's word, *"human reasoning generalization"*; p.30) can help us efficiently gain knowledge about an unknown psychological process, especially if our knowledge is quite limited. Like a state-of-the-art optimization algorithm in the context of estimating the parameter values of a model, a model that makes predictions for novel contexts provides an efficient method of searching through the space of all potential data. These novel predictions can help lead researchers to sources of data that are most informative in teasing apart different models, while avoiding less informative sources of data; a level of efficiency that a giant 'grid search' through all possible data would be unable to achieve. Providing an efficient search of the data space is where I believe predictions about novel contexts are most useful, as they provide clear directions for what experiments are most likely to discriminate between competing models most clearly.

However, I also believe that this is where the limited value of predictions about novel contexts ends. When researchers are trying to find the best explanation for a process, predictions for novel contexts do not provide any more information about which model provides the closest representation of the process than predictions for known contexts. From my perspective, data are simply observations that we make of some unknown process. Data are not inherently of *"theoretical interest"* (p.32), apart from in their ability to tell us which model provides the closest representation of this unknown process – a process that we, as scientists, wish to understand. Therefore, to be the best explanation of a process, a model should provide the best predictions across *all* possible data from *all* possible contexts that we believe are observations of this same unknown process, and not just the data that are from novel contexts. Importantly, assessing which model makes the best predictions across all available data is something that quantitative model selection methods have been specifically designed to achieve (Myung & Pitt, 1997; Evans & Brown, 2018). Quantitative model selection methods compare models in their ability to make accurate, yet tightly constrained (i.e., low flexibility), predictions about the data; factors that many have argued are important in finding a theory that accurately reflects the underlying psychological process (Roberts & Pashler, 2000; Myung, 2000; Evans, Howard, et al., 2017).

As a concrete example of why predictions about novel contexts are most important in theory evaluation, Navarro (2019) eloquently points out that the Rescorla-Wagner model (Rescorla & Wagner, 1972) served an important purpose in research on classical conditioning, with its novel predictions pushing researchers to explore new, specific directions. Exploring these novel predictions led to the discovery of many empirical phenomena – with the Rescorla-Wagner model making accurate predictions for many of these novel contexts – which helped to further shape researchers' understanding of classical conditioning. However, should we consider the Rescorla-Wagner model to be the best explanation of classical conditioning (e.g., the explanation we provide in textbooks for how the process operates) if it provides a substantially worse predictions than other models for all of the data that we already know about? For me, this is a very clear 'no'. In my opinion, Navarro (2019) has conflated two unique goals of process models in this example: the ability to provide the best explanation of what is actually happening, and the usefulness to guide us to new empirical discoveries that we may not have thought of exploring at otherwise (e.g., predictions for novel contexts that lead to new empirical phenomena). While novel predictions are useful for guiding empirical discovery, evaluating models based on their ability to successfully make these predictions ignores all other observations we have about this same unknown process from other contexts. Therefore, assessing only novel predictions provides a poor overall reflection of which model provides the best explanation of a psychological process.

Are certain data of more *"theoretical interest"* than others?

Throughout the section *"Between the devil and the deep blue sea"* (p.31–33), Navarro (2019) makes numerous suggestions that some parts of the empirical data are of more *"theoretical interest"* (p.32) than others. Specifically, Navarro states that *"To my way of thinking, understanding how the qualitative patterns in the empirical data emerge naturally from a computational model of a psychological process is often more scientifically useful than presenting a quantified measure of its performance"* (p.33), and makes numerous references throughout the concrete example of Hayes, Banner, Forrester, and Navarro (2018) to how the qualitative patterns in the data are of greater value than the quantitative fits. However, what exactly makes these qualitative patterns more *scientifically useful* than precise quantitative measurement? As I discussed previously, from my perspective data are just observations that we make of some unknown psychological process, and we use these observations to try and better understand this process. Therefore, it seems strange to me that some specific parts of the data (i.e., the data that compose the specific qualitative pattern) would provide a more theoretically interesting answer about which model best explains the psychological process of interest than the other parts of the data (i.e., the data that quantitative model selection methods would also take into account). Below I critique three general arguments for why qualitative patterns are commonly thought to be more theoretically interesting than quantified measures of performance. These arguments are each either explicitly stated, or appear to be alluded to, by Navarro (2019), and in my experience are often the beliefs of researchers who prefer qualitative assessments over quantitative model selection. The arguments that I critique are: that 1) qualitative trends are able to distinguish between models more clearly, 2) precise quantitative differences can be harder to observe and understand, and 3) qualitative trends can often avoid ancillary assumptions of the models, which model selection methods can heavily depend on. Note that I give the third argument its own section *(Where is the border between core and ancillary model assumptions?)*, as I believe that this is a more general point about core and ancillary assumptions in process models.

The 'qualitative trends often distinguish between the models more clearly' argument

One argument for qualitative trends being more theoretically interesting than precise quantified measures of performance is that qualitative trends are able to distinguish between the models clearly. I think many would argue that the 'proof of the pudding is in the eating' here, as qualitative trends have been one of the main methods in psychology for deciding between competing models, and many of these robust qualitative trends end up serving as benchmarks for new potential models to meet before being taken seriously. However, this general argument seems to imply that quantitative model selection methods cannot distinguish between models clearly, and that qualitative trends are able to magically capture something that quantified measures of performance cannot.

First, it seems important to define what exactly is meant by 'distinguishing' between models. I think a reasonable definition is something along the lines of 'situations where evidence can be shown for one model over another, to reduce ambiguity in which model provides a better explanation of psychological process of interest'. If this is an accurate definition of what it means to distinguish between models, then I believe that it is categorically false to suggest that quantitative model selection methods cannot clearly distinguish between models, or that the distinction obtained through quantitative model selection methods is in any way inferior to the distinction obtained from qualitative trends. For example, in the case of the Bayes factor (Kass & Raftery, 1995), a value of 1 indicates no distinction between the models, whereas larger (or smaller) Bayes factors reflect greater distinction between the models, until the evidence becomes overwhelmingly in favour of one model over the other. Therefore, quantitative model selection appears to both have the ability to reduce the ambiguity in which model is better, and to know the strength of evidence for one model over the other (i.e., the amount that the ambiguity was reduced by), meaning that quantified measures of performance can just as clearly distinguish between models as qualitative trends.

The 'qualitative trends are easier to observe and understand than quantitative differences' argument

Another argument for qualitative trends being more theoretically interesting than precise quantified measures of performance is that qualitative trends can be visually observed in a clear manner, whereas the more precise quantitative differences can be harder to see, and it can be harder to understand why one model beats another. Navarro (2019) states in the example of Hayes et al. (2018) that *"It is clear from inspection that the data are highly structured, and that there are systematic patterns to how peoples judgements change across conditions. The scientific question of most interest to me is asking what theoretical principles are required to produce these shifts. Providing a good fit to the data seems of secondary importance."* (p.32). Here, Navarro seems to suggest that the difference between the models can be clearly seen in the qualitative trends, making these trends of theoretical interest, and that accounting for the rest of the trends in the data, which the quantitative fit detects, is less important as these trends are less clear. I agree with Navarro – and others who make this general argument – to some extent here. Understanding *why* one model is better than others is an important scientific question that increases our understanding of a process, and provides us with future directions for model development (e.g., **'model X** misfits **data pattern Y**, so therefore, we should look into **mechanism Z** that may be able to deal with **data pattern Y'**). Gaining insights into this 'what went wrong?' question is most easily achieved through visual assessments of qualitative trends, as we can clearly see that the certain models miss certain trends, and that certain models capture certain trends. However, 'selecting the model that provides the best explanation of the unknown process' and 'understanding what specific trends in the data certain models cannot explain' are two completely different goals, and the ability of qualitative trends to achieve the latter does not make them better than quantitative model selection at performing the former, in contrast to what Navarro appears to suggest.

More generally, I do not believe that being able to visually observe a trend – based on the way that the data have been plotted – means that the observed trend should have priority over all other possible qualitative and quantitative trends in the data. Realistically, there are always likely to be several trends that can potentially be visually observed in the data, which may be shown or obscured by different ways of visualizing the data. A clear example of this can be seen in the comparisons of the diffusion model and the urgency gating model in Evans, Hawkins, et al. (2017), who show that only looking at certain trends in the data (such as interactions in summary statistics over conditions) can be misleading, and plotting the entire distributions show other, clearer trends that distinguish between the models (see Figure 1 for a more detailed walk-through of this example). However, even in cases where we manage to plot the data in every way possible, and find every qualitative trend present in the data, how do we weight these different trends? As the number of trends increase, it seems unlikely that every trend will be best accounted for by a single model, making selecting a model based on qualitative trends difficult. In contrast, quantitative model selection methods are able to simultaneously account for all of the trends in the data that they are applied to, and provide a principled approach for weighting for all of the trends together. Essentially, quantitative model selection methods are able to take into account everything that visually assessing a finite number of qualitative trends can, and more. The only reason that assessing qualitative trends can give different results to quantitative model selection is that assessing only a subset of the data – as is the case when assessing qualitative trends – *ignores* all other aspects of the data. If researchers are only interested in explaining the single qualitative trend in the data, then I can see why only assessing the single qualitative trend makes sense. However, in cases where researchers want to explain the entire psychological process – which I think is most situations – then only assessing these visually observed qualitative trends is limiting practice, rather than a theoretically interesting one.

Where is the border between core and ancillary model assumptions?

One last argument for qualitative trends being more 'theoretically interesting' than precise quantified measures of performance is that assessing qualitative trends focuses on the core assumptions of the models, whereas model selection methods can heavily depend on the ancillary assumptions of the models. Navarro (2019) states in the concluding paragraph that *"it seems to me that in real life, many exercises in which model choice relies too heavily on quantitative measures of performance are essentially selecting models based on their ancillary assumptions"* (p.33). Here, Navarro seems to suggest that we should only be interested in specific assumptions of models – deemed to be core to the explanation – and attempt to ignore other assumptions – deemed to be ancillary to the explanation. I agree with Navarro that all assumptions in the model can have a large influence on quantitative model selection methods, and researchers may consider some of these assumptions to be ancillary. However, what are the implications of starting to classify certain assumptions as ones that the model is committed to, and others as ones that are flexible and interchangeable?

The idea of core and ancillary assumptions appears to come up quite regularly in theory and model development. Models can have core assumptions, which are fundamental parts of the model's explanation of the process that cannot be changed, and ancillary assumptions, which are only made because they are required for the formalization of the model (e.g., for simulation, or fitting). However, what exactly makes one assumption *core*, and another *ancillary*? The distinction may seem like common sense while speaking in abstract terms, but I believe that these different types of assumptions become much harder to distinguish between in practice. In practice, the lines between core and ancillary assumptions can often be blurred, and declaring certain assumptions in a model as being ancillary allows a researcher to still find evidence in favour of their preferred model – success that they attribute to the core assumptions – and dismiss evidence against their preferred model – failure that they attribute to incorrect ancillary assumptions. Importantly, being allowed to adjust these ancillary assumptions can make a model infinitely flexible, even if any instantiation of the model with a specific set of ancillary assumptions is not infinitely flexible. Jones and Dzhafarov (2014) provide a clear example of this issue with the diffusion model, where if the distribution of trial-to-trial variability in drift rate is considered an ancillary assumption of the model – a common thought among researchers in the field – then the diffusion model has infinite flexibility in explaining choice response time distributions. However, the diffusion model has been shown to be quite constrained in its predictions when assuming a specific distribution of trial-to-trial variability in drift rate (Smith, Ratcliff, & McKoon, 2014; Heathcote, Wagenmakers, & Brown, 2014), such as the normal distribution (Ratcliff, 2002), suggesting that the change in flexibility is created by the choice of whether the assumption is labelled as *core* or *ancillary*. This suggests that breaking models into core and ancillary assumptions can be a slippery slope, and the flexibility of a model can rapidly increase by labelling certain assumptions as being ancillary.

In contrast to Navarro (2019), who wished to avoid making interpretations based on ancillary assumptions, I believe that when a formalized model of a process is defined, then this model represents the complete explanation of the process. There are no *core* and *ancillary* assumptions of the model: just *assumptions*. This is similar to the point made by Heathcote et al. (2014) in their reply to Jones and Dzhafarov (2014), where they suggest that Jones and Dzhafarov targeted a straw-man definition of the diffusion model, and that the distributional assumptions should not be considered ancillary. Therefore, I believe that the ability to remove the influence ancillary assumptions of the models does not make qualitative trends more theoretically interesting to assess, and instead, creates a slippery slope towards infinite flexibility.

Having said this, I can also understand why researchers may be reluctant to commit to all assumptions of a model as being core to the explanation. As Navarro (2019) points out in the Hayes et al. (2018) example, there are often difficult decisions that need to be made to create a formalized model, and some of these choices can end up being somewhat arbitrary. However, making each model a complete explanation with only core assumptions does not mean that assumptions that would normally be considered ancillary cannot be tested. Specifically, multiple models can be proposed as explanations of the unknown psychological process, with each model containing a different instantiation of these ancillary assumptions, and these models being compared using quantitative model selection methods. However, each model with different assumptions is now a different, separate explanation, and researchers cannot switch between these different models for different paradigms while still claiming that this represents a success of a single explanation. I believe this presents a principled way to address the issue of ancillary assumptions in models, providing the robustness against potentially arbitrary modelling choices desired by Navarro, while preventing the ancillary assumptions from making models infinitely flexible.

A brief digression: Is automation actually a bad thing?

One final point that appears to be implied by Navarro (2019) is that model selection methods being automated is a negative. Although this isn't a central point of Navarro, the statements *"To illustrate how poorly even the best of statistical procedures can behave when used to automatically quantify the strength of evidence for a model"* (p.31) and *"I find myself at a loss as to how cross-validation, Bayes factors, or any other automated method can answer it"* (p.32) both appear to show some level of negative connotation around the methods being automated. I agree that there is a general issue with 'black box' approaches, which can be applied and interpreted incorrectly when users do not understand them properly. However, I believe that automated quantitative model selection methods, which are applied in a consistent manner from situation-to-situation, do not belong in this category. More generally, why would a method being consistent in how it is applied be considered a bad thing?

Generally, I would consider automation to often be a good thing, and that the case of model selection is no exception. Instead of calling these methods *automated*, I would refer to them as *principled*. Quantitative model selection methods are based on statistical theory, are clearly defined, and follow a systematic procedure that always compares the models in the same way. From my perspective, this seems like a good thing; being methodical is often what makes science robust. In the context of experimentation, a lot of the designs that researchers commonly use are essentially automated: these designs are based on methodological theory (i.e., minimizing measurement error and potential confounds), are clearly defined, and researchers implement them in an almost automatic, systematic fashion from experiment to experiment. However, I do not think that many researchers would argue that systematic experimentation using robust, well-developed experimental designs is a negative. So why does this principled, systematic nature of science suddenly become a negative when it comes to our method of inference? I would instead suggest that automation is the natural next step after an approach becomes rigourously defined, and when a method cannot be automated, we should question how rigourous the method actually is. I believe that it is actually problematic that there is no automated process for visually assessing qualitative trends, where the same result would always be reached by entering the same information at the beginning, and that the lack of automation suggests that this approach lacks clear principles in at least some regard.

Author Contact

Correspondence concerning this article may be addressed to: Nathan Evans: nathan.j.evans@uon.edu.au

Conflict of Interest and Funding

There are no potential conflicts of interest regarding the current article. NJE was supported by an Australian Research Council Discovery Early Career Researcher Award (DE200101130) and European Research Council Advanced Grant (UNIFY-743086).

Author Contributions

NJE conceptualized the project, reviewed the relevant literature, and wrote the manuscript.

Open Science Practices

This article is a commentary without data, material or analysis of the type that could have been pre-registered and reproduced. The entire editorial process, including the open reviews, are published in the online supplement.

References

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive psychology*, 57(3), 153–178.

Carland, M. A., Marcos, E., Thura, D., & Cisek, P. (2015). Evidence against perfect integration of sensory information during perceptual decision making. *Journal of neurophysiology*, 115(2), 915–930.

Carland, M. A., Thura, D., & Cisek, P. (2015). The urgency-gating model can explain the effects of early evidence. *Psychonomic bulletin & review*, 22(6), 1830–1838.

Chandramouli, S. H., & Shiffrin, R. M. (2019). Commentary on gronau and wagenmakers. *Computational Brain & Behavior*, 2, 12–21.

Cisek, P., Puskas, G. A., & El-Murr, S. (2009). Decisions in changing conditions: the urgency-gating model. *Journal of Neuroscience*, 29(37), 11560–11571.

Evans, N. J., & Annis, J. (2019). Thermodynamic integration via differential evolution: A method for estimating marginal likelihoods. *Behavior research methods*, 1–18.

Evans, N. J., & Brown, S. D. (2017). People adopt optimal policies in simple decision-making, after practice and guidance. *Psychonomic Bulletin & Review*, 24(2), 597–606.

Evans, N. J., & Brown, S. D. (2018). Bayes factors for the linear ballistic accumulator model of decision-making. *Behavior research methods*, 50(2), 589–603.

Evans, N. J., Hawkins, G. E., Boehm, U., Wagenmakers, E.-J., & Brown, S. D. (2017). The computations that support simple decision-making: A comparison between the diffusion and urgency-gating models. *Scientific reports*, 7(1), 16433.

Evans, N. J., Howard, Z. L., Heathcote, A., & Brown, S. D. (2017). Model flexibility analysis does not measure the persuasiveness of a fit. *Psychological review*, 124(3), 339.

Gronau, Q. F., & Wagenmakers, E.-J. (2019a). Limitations of Bayesian leave-one-out cross validation for model selection. *Computational Brain & Behavior*, 2, 1–11.

Gronau, Q. F., & Wagenmakers, E.-J. (2019b). Rejoinder: More limitations of Bayesian leave-one out cross-validation. *Computational Brain & Behavior*, 2, 35–47.

Hayes, B., Banner, S., Forrester, S., & Navarro, D. (2018). Sampling frames and inductive inference with censored evidence.

Heathcote, A., Wagenmakers, E.-J., & Brown, S. D. (2014). The falsifiability of actual decision making models. *Psychological Review*, 121(4).

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological review*, 121(1), 1.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association, 90*(430), 773–795.

Kiani, R., Hanks, T. D., & Shadlen, M. N. (2008). Bounded integration in parietal cortex un derlies decisions even when viewing duration is dictated by the environment. *Journal of Neuroscience*, 28(12), 3017–3029.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of mathematical psychology*, 44(1), 190–204.

Myung, I. J., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, 50(2), 167–179.

Myung, I. J., & Pitt, M. A. (1997). Applying occams razor in modeling cognition: A bayesian approach. *Psychonomic Bulletin & Review*, 4(1), 79–95.

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*, 2, 28–34.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological review*, 104(2), 266.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological review*, 122(2), 260.

Pilly, P. K., & Seitz, A. R. (2009). What a difference a parameter makes: A psychophysical comparison of random dot motion algorithms. *Vision Research*, 49(13), 1599–1612.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, 85(2), 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(2), 278–291.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 2, 64–99.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological review*, 107(2), 358.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: Remretrieving effectively from memory. *Psychonomic bulletin & review*, 4(2), 145–166.

Smith, P. L., Ratcliff, R., & McKoon, G. (2014). The diffusion model is not a deterministic growth model: Comment on jones and dzhafarov (2014). *Psychological Review*, 121(4).

Thura, D., Beauregard-Racine, J., Fradet, C.-W., & Cisek, P. (2012). Decision making by urgency gating: theory and experimental support. *Journal of neurophysiology*, 108(11), 2912–2930.

Tsetsos, K., Gao, J., McClelland, J. L., & Usher, M. (2012). Using time-varying evidence to test models of decision dynamics: bounded diffusion vs. the leaky competing accumulator model. *Frontiers in neuroscience*, 6, 79.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, 108(3), 550.

Vehtari, A., Simpson, D. P., Yao, Y., & Gelman, A. (2019). Limitations of "limitations of Bayesian leave-one-out cross-validation for model selection". *Computational Brain & Behavior*, 2, 22–27.

Winkel, J., Keuken, M. C., van Maanen, L., Wagenmakers, E.-J., & Forstmann, B. U. (2014). Early evidence affects later decisions: Why evidence accumulation is required to explain response time data. *Psychonomic bulletin & review*, 21(3), 777–784.

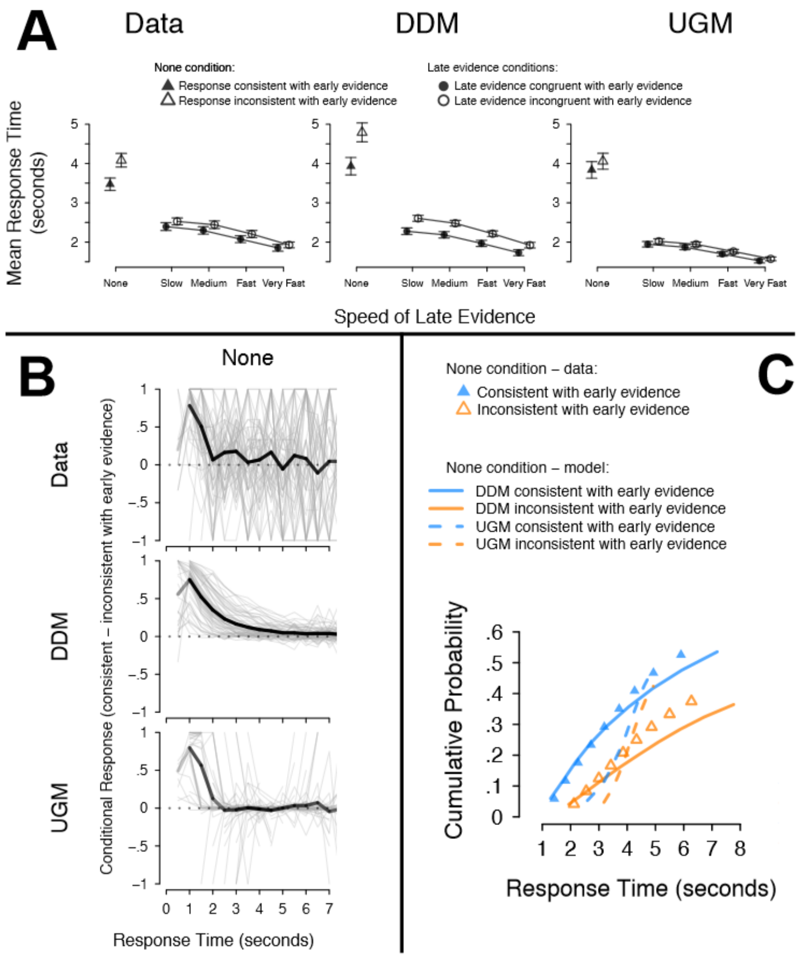

Figure 1. An example of three different ways (**A, B, C**) that the data could be, and were, visualized in Evans, Hawkins, et al. (2017). Evans, Hawkins, et al. (2017) attempted to compare the diffusion model (DDM; Ratcliff, 1978) and urgency-gating model (UGM; Cisek et al., 2009) by using a random dot motion task (e.g., Pilly & Seitz, 2009; Evans & Brown, 2017), where the evidence for each alternative changed over the course of each trial. Each trial began with a brief burst of ‘early evidence’, which was either the same as (congruent) or different to (incongruent) the ‘late evidence’, creating a variable of ‘congruency’. The late evidence either increased over time at one of four rates (slow, medium, fast, very fast), or was not present (none), creating a variable of ‘ramp rate’. **Panel A** plots a single qualitative trend, being the interaction between congruency and ramp rate on mean response time. Both models appear to capture the overall pattern of the interaction, though the UGM is able to account for the negligible effect of congruency on mean response time, whereas the DDM overpredicts the effect. **Panel B** plots a single, but different, qualitative trend, being the change in the difference between congruent and incongruent trial accuracy over time (i.e., a conditional accuracy function; CAF) for the ‘none’ condition. Again, both models appear to capture the overall pattern of change over time, though the UGM is able to account for the quick decrease of the CAF to zero, whereas the DDM underpredicts the rate of the decrease. **Panel C** plots the entire choice response time distributions for the ‘none’ condition; data which includes both of the qualitative trends shown in A and B. However, in this case the DDM clearly provides a better account of the entire distributions than the UGM, with the UGM display sizable misfit in several aspects of the data. These sources of misfit for the UGM were obscured by the methods of visualizing the data in **A** and **B**, where **A** and **B** appeared to suggest that both models explained the data well, but the UGM did so somewhat better. Importantly, most previous studies comparing these models (Cisek et al., 2009; Thura et al., 2012; Winkel et al., 2014; Carland, Thura, & Cisek, 2015; Carland, Marcos, et al., 2015) or similar models (Kiani et al., 2008; Tsetsos et al., 2012) had only focused on the qualitative trends seen in **A** and **B**, drawing into question the validity of the findings of previous studies.