*Notes:*

*New pitch: General problem of testing dynamic theory. We need to have tools that are flexible enough to accurate map onto theory. Different tools have been proposed, used in bespoke ways. Latent change models etc, are used to make inferences statistically about dynamic processes, but have lots of problems eg, inflexibility, failure to examine process (bottom=up phenomena). On the other hand, computational models are becoming more widely used for theory building. These have several advantages such as being able to simulate highly dynamic systems and examine how lower-level processes give rise to higher level phenomena. But challenge is validation. Vary difficulty to fit models to data using current approaches. Thus, hard to validate theory and make inferences about parameters. We propose a framework that unites these…*

*Paper would have same structure, existing models would be first section. Emergence models would be second section. Strip out stan code.*

*First step would be to get Jeff’s model working for Phillip’s data.*

*Bliese & Lang (2016) – Understanding relative and absolute change in discontinuous growth models.*

*Need macro examples, but don’t need to be analyzed.*

Action Editor’s Comments

*1. A central concern relates to the contribution, “uniqueness” and value-add of the manuscript.  In particular, it is unclear how researchers studying dynamic processes would benefit from using Bayesian techniques.  Yes, there has been a great deal of discussion regarding the potential benefits of Bayesian approaches, but as Reviewer 1 notes, these benefits may be overstated in many instances.  Moreover, there are potential questions about whether the benefits of these suggested Bayesian techniques are worth the added complexity of these approaches.  For example, the manuscript currently uses real data from a recent empirical study involving a dynamic goal revision process (Gee, Neal & Vancouver, 2018), yet it is unclear how the Bayesian approach used in the manuscript differs from those used in the original study or allows researchers to extend and/or look at different aspects of this study/phenomenon.  It is therefore critical that more be done to clarify and/or develop a stronger case for the importance of the approach suggested in the paper.  Along these lines, I wonder whether it makes sense to retain the current focus on the Gee et al. (2018) study as your modeling example. There are techniques available to explore dynamic processes, in general, and the reader is left wondering why and how the Bayesian approach is needed (and/or provides unique benefits and new insights) if there was already an approach available for Gee et al. (2018) to use in their study.  Alternatively, it might make more sense to give greater attention to the applicability of the Bayesian approach suggested in the manuscript to examine bottom-up processes.  As is noted in the manuscript, there are very few techniques available to model the multilevel bottom-up oriented dynamic processes involved with emergence and there is the potential to make a significant methodological contribution if the technique suggested in the current manuscript allows a way to model and examine these sorts of theories.  Thus, rather than briefly mentioning this as an application of the technique in the discussion section, I would like to ask that you consider giving this more explicit and detailed attention in the manuscript.  Of course, there could be a compromise where you start with a more “straightforward” example using the Gee et al. (2018) data/model and then expand this to apply the approach to look at dynamic bottom-up processes involved with an emergent phenomenon. If you decide to retain the sole focus using the Gee et al. (2018) data, it is essential that you demonstrate how your approach provides specific and important benefits compared to traditional analytical techniques to assess/examine the associated dynamic processes.*

We see now how the contribution of our paper could have been clearer in the originally submitted manuscript. The contribution that we intended to make was broader than simply highlighting benefits of the Bayesian approach, it was about demonstrating a framework that has the flexibility required to more directly test dynamic theories. The point we were trying to make was that traditional statistical approaches (e.g., latent growth, latent change score modeling) do not have the flexibility to implement many inherent features of dynamic theories such as dynamic (i.e., stock or level) variables and feedback loops, which makes these theories difficult to test using conventional approaches. Our primary contribution was to demonstrate an approach to testing dynamic theory that is flexible enough to be able to accurately represent the theory being tested, which therefore provides a more direct test of the theory.

Notably, while the approach we demonstrate takes advantage of some recent developments in Bayesian parameter estimation, for our purposes the Bayesian approach is a means to an end. Programs like Stan that are used for Bayesian parameter estimation allow us are ideally suited to this problem because they are extremely flexible with regard to the models that can be specified and can be used to analyze highly complex and dynamic models. That being said, we do recognize that the broader contribution of improving the way dynamic theory is tested may have been overshadowed by our lengthy introduction to the benefits of the Bayesian approach. In our revised, we have removed much of this discussion so that we could focus the introduction more squarely on the issue of dynamic theory testing.

As requested, we have strengthened the case for the importance of the approach we are suggesting. In the revised manuscript, we make the point that although theory development and testing is important for the cumulative progress of science, testing dynamic theory is difficult. Testing dynamic theory requires a) a model that operationalizes the key processes described by the theory, and b) a way of quantifying the extent to which the model accounts for a set of empirical observations. Different tools have been proposed for testing dynamic theory, but each has its own limitations. One common approach is to use statistical modeling (e.g., latent growth curve modeling, latent change score modeling). This approach is well suited to quantifying the extent to which a model accounts for empirical data. However, using this approach, it is difficult to directly represent the types of phenomena inherent in dynamic theory. For example, it is difficult to simulate the effects of dynamic variables or feedback loops using this approach. As a result, there is often a misalignment between the theory and the model, which makes the theory difficult to corroborate.

Another approach that is often advocated is computational modeling, which is the practice of articulating theory in the form of mathematical equations and/or computer code and evaluating the dynamic properties of theory by simulating the model. Computational modeling allows for a more direct mapping between theory and model, because the model is custom built to represent the processes described by the theory. However, in organizational sciences, computational models are often primarily used for theory development, rather than theory testing. The correspondence between the model output and the empirical data are often not directly examined. As a result, the potential for computational models to be used for theory testing as yet to be realized.

(finish description of pitch, possibly make it briefer).

Finally, to respond to the query regarding how the data in the original paper was analyzed, these data were analyzed by fitting a computational model that was programmed directly in R, using maximum likelihood parameter estimation. However, the approach that was used in the original paper is not well suited to analyzing more complex models like the ones we introduce in this paper for two reasons. First, the user has to know a great deal about R programming in order to create the infrastructure required to analyze the model. For example, the user has to manually code the likelihood function, which becomes more complicated as model complexity increases. By contrast, the Stan package does this automatically. Second, the parameter estimation algorithm implemented by Stan (Hamiltonian Monte Carlo) is more robust, especially when dealing with more complex and/or hierarchical models. We believe that the approach presented in our paper is a more generally useful way to implement these sorts of models.

*2.  The accessibility of the manuscript to the typical ORM reader could be improved.  As Reviewer 1 notes, most readers will not be familiar with the Bayesian approaches discussed.  Thus, it is important to, wherever possible, use words and terminology that are more accessible and familiar to researchers with a basic understanding of traditional multilevel techniques. Many of the technical details (including the Stan code and explanations) can be moved to an appendix. Reviewer 1 has a great suggestion about finding your favorite ORM articles and emulating their structure and Reviewer 2 also has a helpful suggestion about starting from a traditional multilevel modeling approach and then using this platform to highlight where and how the Bayesian technique differs.  Another helpful practice is to give your manuscript to a recent graduate of a management PhD program (ideally one that provides no or very surface level background on Bayesian techniques) and see if they can follow your points and suggestions – places where they get lost indicate that these need more exposition and discussion. It can also be helpful to describe the key modeling points and considerations in multiple ways: equations, word explanations, and figures.*

We thank the Action Editor and Reviewers for the advice regarding how to ensure our manuscript is accessible to the ORM audience. As suggested, we have moved all the Stan code (and accompanying text) to an appendix.

ETC ETC  
  
*3. There is an overemphasis on micro examples in the manuscript.  As we want the articles in the FT to speak to both micro- and macro-oriented researchers, it would be good if you added more discussion on the applicability of the approach to potential macro-oriented research questions.  In doing so, it would be good to think about potential research applications in traditional strategic management (strategy process and content areas) and organizational theory topics, as well as research on microfoundations (the notion of emergence is often central in this area -- e.g., see Felin et al., 2015 Acad Mgmt Annals and Barney & Felin, 2013 Acad Mgmt Perspectives for overviews). Additionally, please replace the term “organizational psychology and organizational behavior” with a more inclusive and all-encompassing term like “organizational studies” or “organizational sciences”.*  
*In addition to the above key issues, the guest editor and me also have the following more particular concerns/comments  
  
4. The section “A flexible Bayesian approach” illustrates the ability of Bayesian methods to make probability statements without any reference to the necessity of the prior for doing so…since the next chunk discusses the utility of the prior for hierarchically-structured models, this limitation seems glaring. Often priors are just default options that contain little realistic information – and this is true for the “highest” level prior for any hierarchical model, too. I agree that all of this is a potential strength for Bayesian models, but it is also a potential weakness, and the importance of prior choice for making concrete probability statements about posterior distributions is important, especially when making claims about null versus model hypotheses, as this section does.  
  
5. In discussing the weaknesses of RMSEA-type measures of complexity, and then saying that Bayes methods “naturally” account for this sort of complexity, it would be good to go into more detail. This gives the impression that Bayes is a “magic bullet” that solves inference problems, as opposed to requiring very serious consideration about prior and likelihood functions for these sorts of inferences to have any meaning whatsoever.  
  
6. While the Rhat rule of thumb of 1.1 is a generally accepted convergence criterion, there are some criticisms associated with its use.  It would be good to acknowledge and note these criticisms (even if in a footnote).  
  
Lastly, it is important that you fully address the reviewers’ comments and concerns about your manuscript. As I have noted above, several of the reviewers’ comments relate to the key issues identified by me and the other guest editor. For example, Reviewer 1 has several comments about the unique value of Bayesian approaches and the need to improve accessibility of the manuscript and Reviewer 2 also raises questions about your arguments regarding the benefits of Bayesian approaches and also accessibility. However, the reviewers also raise a number of other important points that need to be addressed fully.  For example, Reviewer 1 has specific comments about the manuscript details and Reviewer 2 raises a number of specific comments about model comparison and other important details.  If you read the reviewer comments carefully, you will notice that Reviewer 1 has disclosed their identity.  While reviewers always have the option to do this, please keep in mind that it is important that your identity as authors remains blinded to both reviewers (i.e., please do not contact or otherwise disclose your identity to Reviewer 1 and remain anonymous in your revised manuscript and response to reviewer comments)*

Reviewer 1’s Comments

*1) This paper has been written in a suboptimal way given that the stated goal of the authors (and a probable expectation at ORM) would involve a kind of tutorial on the topic of dynamic models with a Bayes estimator using the Stan package. The authors note that many ORM readers will not be familiar with these topics, but then proceed to describe them in a way that presumes a substantial familiarity with them vis-a-vis the language/terms used by the authors, including the proposed benefits of the authors' recommended Bayes approach. If everyone is a frequentist who doesn't know much about Bayesian methods or how to use Stan -- as the authors note -- then I recommend starting with familiar concepts and terms, and then grafting these onto the alternative logic proposed by the authors, even if by way of explicit contrast (e.g., most likely parameter estimates vs. posterior distribution means, SEs vs. SDs, etc.). Unfortunately, in my view, if this were done in earnest, what the reader would probably realize is that frequentist and Bayesian approaches are not actually all that different unless informative priors are used. This deserves its own paragraph:  
  
The authors talk about a Bayes approach as if it's inherently different or better than a frequentist approach, but this isn't very accurate. Consider that with uninformative priors, the results of analyses will be the same in the long run under both approaches. Consider also that the proposed benefits of posterior distributions can be gotten through various types of frequentist bootstrapping procedures -- Brad Efron has a paper on the 'Bayesian bootstrap' that makes this and related points. Furthermore, it's not true that frequentist inference is limited to the Yes/No case of hypothesis testing (as the authors state). Maybe this is the way frequentist information is used, but this is actually NHST rather than frequentism, and Bayesian approaches can be subjected to this same Yes/No logic -- the authors note on p.16 that "The CI or HDI is often used to decide whether particular parameter values should be accepted or rejected"! Indeed, the whole point of the likelihood is that it's applied to models rather than data -- a point Fisher made long ago, including with his 'fiducial inference' approach -- and inference using the likelihood can be applied to models even if it violates probability theory (why is everyone so addicted to probability theory, as if it should be uncritically applied to everything using formal logic?). Also, the lack of a relevant sampling theory might be considered a Bayes disadvantage, which is perhaps by de Finetti and others worked so very hard to address it in the 20th Century -- eventually creating the 'exchangeability' criterion to help facilitate formal inference to future observations. The Bayes community typically over-hypes their favorite differences to their own advantage. I don't blame 'em, for who doesn't do this when attempting to get ahead in science (or politics)?  
  
To address my points here, the authors need to substantially revise their paper with the following in mind: Bayes and Frequentist approaches are actually not that different, and both can be used for very similar purposes -- Bayes' theorem describes how to reverse the inferential process from data to parameters/hypotheses. As such, the authors should tone down their support for Bayes approaches, which do offer pragmatic benefits in some cases but little more than that, I think -- and, to be clear, I'm familiar with the debates here (I'm the first author of the Zyphur & Oswald, 2015 article in JOM that advocates for Bayes methods... I've since tempered my passions on the topic). By looking for similarities between frequentist and Bayes approaches, the authors can rewrite their paper so that instead of jumping into unfamiliar Bayes concepts early (in an unprompted defence of Bayes), the authors could introduce relevant Bayes concepts through the application of their proposed approach, noting how what they're doing would be conceptualized using more typical frequentist terms/logic. If the authors actually want to help the average ORM reader, something like this will be needed -- a tutorial for the uninitiated. Currently, the paper seems like it's as an introduction to something that only people already introduced to the topic will care about and understand.  
  
To help here, perhaps the authors could use their favorite ORM article (or set of articles) as examples. Typically, a substantive problem will be described and then tackled using a method that uniquely addresses the problem, or pre-existing methods will be described as being problematic when trying to answer specific questions. Along the way, familiar concepts and terms can be used, whether it's multilevel modeling (as the authors could draw on) or multi-group modeling in SEM, etc., the point is that the average ORM reader is going to need more help than the authors are offering. Of course, if the authors aren't actually interested in talking to the average ORM reader and instead want to talk to a subgroup, then this should probably be stated up front and the paper written for this group.*

We thank the Reviewer for the advice regarding how to improve the framing of the manuscript. We see now how the lengthy discussion of the benefits of the Bayesian approach in the introduction may have detracted from the general point we were trying to make in the paper. The contribution that we are hoping to make is broader than simply demonstrating a Bayesian alternative to existing statistical approaches, it is about testing dynamic theory. We have attempted to communicate the contribution more clearly in the revised manuscript.

Our argument is that testing dynamic theory requires a) a model that operationalizes the key processes described by the theory, and b) a way of quantifying the extent to which the model accounts for a set of empirical observations. The problem is that there are not many tools available that simultaneously address both a) and b). For example, statistical approaches such as latent growth curve modelling and latent change score modelling are useful for quantifying model-data correspondence and estimating relationships between variables. However, using this approach, it is difficult to directly represent certain types of phenomena that are common in dynamic theory. For example, it is difficult to examine the effects of dynamic variables or feedback loops using this approach, because they typically need to be modelled via simulation. As a result, there is a misalignment between theory and model that can make the theory difficult to corroborate.

Another approach that is often advocated is computational modeling, which is the practice of articulating theory in the form of mathematical equations and/or computer code and evaluating the dynamic properties of theory by simulating the model. Computational modeling allows for a more direct mapping between theory and model, because the model is custom built to represent the processes described by the theory. However, in organizational sciences, computational models are often primarily used for theory development, rather than theory testing. The correspondence between the model output and the empirical data are often not directly examined. As a result, the potential for computational models to be used for theory testing is yet to be realized.

The contribution this paper makes is to demonstrate an approach to theory testing that enables a) a close coupling between the theory of interest and the model used to test it, and b) a direct examination of the extent to which the model can account for empirical observations. We believe this approach can provide stronger, more direct tests of dynamic theory than many of the approaches typically used in organization sciences.

Importantly, while the approach we demonstrate takes advantage of some recent developments in Bayesian parameter estimation, for our purposes the Bayesian approach is a means to an end, and we no longer argue that our use of this approach is a primary contribution of the manuscript. We use the Bayesian approach because programs like Stan are ideally suited to the approach we demonstrate in the paper. One reason for this is that Stan is flexible enough to allow for the implementation of dynamic, computational models that need to be simulated in order to make predictions. This enables us to implement models in which feedback loops operate over time and at different levels of analysis. Another reason we use Stan is that the Hamiltonian Monte Carlo (HMC) algorithm that Stan uses is ideally suited to estimating the parameters of the complex models we illustrate in the paper. However, we recognize that these are primarily implementation-level issues that do not warrant a great deal of attention in the introduction. In our revised manuscript, we have removed much of the discussion of Bayesian statistics so that we could focus the introduction more squarely on the issue of dynamic theory testing.

*2) The authors' model is poorly motivated. I haven't been told enough about the variables and the process being studied to know what Gt is, not to mention the other model terms. This make it very difficult for me to appreciate what the authors are trying to do with the model and why the results are being understood in the way(s) they are.*

This is a good point. We now devote an entire subsection to introducing the problem domain, and describing how the model operationalizes the various components of the theory.

*3) The introduction to Stan is not very helpful. At a minimum, the input code requires annotation. Either way, I think it should be moved to an appendix so that the authors can focus on the substantive example of interest in a way that the reader will better appreciate. In my view, the pages of an ORM article should not be treated as a user's guide for specific programs. Appendices are better for this.*

As requested, we have moved all the code to the appendix. In the original manuscript, we opted not to annotate the code because we were concerned that the additional text would be confusing for readers (we annotated the code in the supplementary materials, just not the code in the manuscript itself). However, we now agree that annotation is helpful, and have therefore annotated the code in the appendix. *4) The authors call a prior ~N(0,1) uninformative (p.13). This is very strange. Perhaps ~N(0,10^10) could be called uninformative, but not a variance of 1 by any means.*

This is a good point. We have removed the sentence in which the ~N(0,1) prior was described as uninformative. *5) The HDI is sometimes called the highest posterior density (HPD). Perhaps note this, as Mplus uses the latter terminology.*

We have now noted this on page X of the revised manuscript.  
  
*I would offer more comments/critiques, but I am unsure they will be very helpful. The paper requires such substantial revision that areas such as the discussion section can't possibly resemble their current state if my previous comments are acted on.*

Reviewer 2’s Comments

*This paper describes a method for estimating parameters for dynamic models and for evaluating those models using the Bayesian approach. I like much of what was in the paper, but am wondering if there is more to be had. In particular, the method seems very appropriate for parameter estimation, but the material on model evaluation was disappointingly thin. I am wondering if this can be improved. Indeed, most of my comments are meant to improve the manuscript; at least that was the spirit in which they were intended.*  
  
*1.      P. 4 (based on upper left corner indexing), line 49-50. In this section you are explaining the complexity of statistically modeling data to assess a dynamic theory. Your first two points (limited observations and, essentially, possible unmeasured moderators) are generally a limitation for statistical tests. What seems more relevant here is that the parameters are not merely descriptive (e.g., of relationships; variance), but often are substantive. The problem though is thus more related to measurement. That is, the parameters often reflect latent constructs, measurement errors exist, and the functional forms might be incorrect to different degrees. This makes estimating them from the data difficult. As far as the dynamic variable issue, this can be done with standard models, but they are complex because they must be modeled correctly (e.g., with autocorrelation). But it seems like the discussion misses the point that the parameters are often not of the statistical kind (e.g., they represent a rate or a bias rather than an association or an intercept).*

*2.      P. 6, line 32. I find the argument that a more complex model is less generalizability more compelling that less parsimonious. That is, because the parameters need to be estimated they can take on more values and thus any one model (with its set of parameters) is less predictive. An interesting aspect of this, given a Bayesian approach, is that the narrowness of the priors will increase generalizability though does not affect parsimony as typically conceived.*

*3.      P. 7. Lines 6-12. I have not heard this argument that deviating from the sampling plan invalidates statistical tests. Sure, there are assumptions of randomness and p-hacking can affect probabilities of errors, but “invalidate” is a strong sentiment that does not seem warranted. Moreover, lines 23-28 seem overly optimistic regarding Bayesian. Often data will fit many models because the data was not designed to diagnose them (i.e., is insufficient in measurement and design to handle overlapping parameters). I think you can tone this paragraph down a bit and still have a point to make.*

In the revised manuscript, we have toned down our discussion of the benefits of the Bayesian approach, and have therefore removed paragraph in question. However, we respond to the Reviewer’s queries below.

Our use of the word “invalidate” might have been too strong in this instance. The point we were alluding to is that many have argued that failing to adhere to a pre-specified sampling plan (e.g., by peeking at the data and then deciding whether to continue sampling) can increase the type I error rate of null hypothesis significance tests. References below:

Rouder, J. N. (2014) Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review,* 21, 301-308.

﻿Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences, *Psychological Methods, 22,* 322-339.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p-values. *Psychonomic Bulletin & Review,* 14, 779-804.

We also agree with the point about data fitting many models. We could have been more specific here. The point we were trying to make is that the Bayesian paradigm is less dependent on the sampling plan than NHST (e.g., optional stopping is less problematic within the Bayesian framework).

*4.      P. 10, lines 7-10. In the simple example presented there are measured variables and estimated parameters (much like in statistical models). However, computational models often have many endogenous variables that are not measured. Because they are endogenous, they are not parameters. How are these represented in these figures? If the answer is they are not, so be it. My point is that one might get hung up one how to handle a model that is a bit more complex because of these endogenous variables.*

This is a good point. The model presented in the original manuscript did not have any endogenous variables. The two variables that were modelled—goal and performance—were observed variables. We have however, revised the model in the new version of the manuscript, and the revised model does include endogenous variables (e.g., skill). We now explicitly distinguish between observed variables and endogenous variables (which we refer to as “unobserved” variables in the manuscript). In the model diagram (Figure 1) Observed variables are depicted as……..whereas unobserved variables are depicted as…..

*5.      Perhaps my problem in the above situation is the word “model.” For your next line talks about the sources of variability in the model, but this is not the computational model, because different Bayesian (or graphic plate) models are tested that reference the same computational model. So you are doing what is commonly done with traditional statistics, which is to conflate (or confuse) the statistical model with the theoretical model being tested. I guess I want you to be explicit about the distinction.*

*6.      P. 10, line 25-30. Describing G as influencing itself is one of the reasons I do not like the path analytic approach to specifying the statistical model. The dynamic variable does not necessarily influence itself. It can, but that is part of the reason one needs to be clear about what is being represented. That is, in the present case, past goal level is not influencing current goal level; it is just the place from which the goal level is adjusted. In contrast, the hole in a bucket will cause the water level in the bucket to drop (water level is dynamic) as a rate determined somewhat by the water level (more pressure the higher the level). Thus, the dynamic variable is also a variable that influences its own change to some extent. These are two different conceptual ideas. Are they represented the same way in a Bayesian model? No, as the functional equation would be different.*

If we are understanding this comment correctly, the query addresses the distinction between dynamic variables in which the rate of change independent of the variable’s level versus variables in which the rate of change is dependent on the variable’s level. We agree that this is an important distinction, and that the path diagram does not provide all the details required to understand the nuances of how the dynamic variables change over time.

To answers the reviewer’s question, the Bayesian framework is agnostic regarding how the dynamic variable operates. Either type of process can be represented, just as would be the case in a computational model that was analyzed using a maximum likelihood approach.

In our model, the change in the dynamic goal variable is not completely independent of the goal itself. It is true that the rate of change is not directly proportional to the value of the dynamic variable, as would be the case in the water bucket example. However, the change in goal level at time *t* is determined in part by the difference between the goal level at time *t* – 1 and the level of performance achieved at time *t* – 1. As such, the goal level has an indirect influence on the rate at which the goal changes.

*7.       P. 12, line 10. Given you are being explicit, seems you should add that the distribution is normal because of the “normal” command.*

We have removed the section in question at the request of the editor.

*8.      P. 14, line 20. What size samples should we use? When should we worry not enough?*

*9.      P. 15, lines 3-6. What does it take for the researcher to feel confident in the obtained posteriors or that it has “converged”? For example, should we not be worried that the standard errors are zero and rhats are 1? Does not fill me with confidence.*

The main rule of thumb used to determine that the model has converged is to check whether the potential scale reduction factor (Rhat) is less than 1.1. This statistic has a lower bound of 1, with a value of 1 indicating complete convergence.

The se\_mean column in the Stan output is the Monte Carlo standard error, which is the standard deviation of the posterior (the “sd” column) divided by the square root of the number of effective samples (the “n\_eff” column). Lower standard errors are better, so values of 0 in this column indicate that there is likely to be very little error in the estimated posterior distributions.

Monte carlo standard errors of 0 and rhats of 1 should make the researcher very confident in the obtained posterior distributions.

*10.     P. 15, top. It does not appear that information for HDI is provided. And is there a test of symmetry? Is something in the model specification likely effect symmetry (e.g., if sigma was near zero).*

*11.     P. 15. Middle. For a parameter that is a weight (or multiplier), seems one would be the value it should differ from to matter. That is, a parameter value for alpha was basically 1, then it might be unnecessary for the model. Seems worth mentioning as most would think zero would mean parameter is not important (like it would be if beta was basically zero), though of course the effect of variance around these means at the person level may still matter a lot.*

This is a good point. We’ve incorporated this into the revision…

*12.     For multilevel models the formatting of the data (input) was important. Seems a similar thing would be so here. For example, would the model (e.g., person-level versus single model) matter? Indeed, evoking HLM-type models when discussing the person level might help connect readers to the model.*

All the models we include in the paper are run based on long-formatted data, where each row represents an observation (as opposed to wide-formatted data, where each row might correspond to a subject and different observations within the time series are represented as different variables). We’ve now made this point explicit (page #).

*13.     P. 21. Would it be useful to contrast the known group membership analysis described here against one that simply used t-tests or ANOVA on the person-level derived parameters? My sense is that if the priors were normally distributed, there would be no different. If that is true, might help people connect the old with the new. If different, might reinforce how different.*

This could certainly be done. However, the problem with estimating model parameters and then conducting further statistical analyses such as t-tests or ANOVAs on those parameters is that doing so can bias the results of the analyses (see reference below). Because of this, we would prefer not to include this type of analysis in our manuscript.

Boehm, U., Marsman, M., Matzke, D., & Wagenmakers, E-J.(2018). On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. *Behavior Research Methods, 50*, 1614-1631.

*14.     P. 23, lines 8-10. The result reference is not the separation, but the nature of it (that those in the approach condition tended to have higher alphas). Disambiguate “results” and what the “this” is on line 16.*

*15.     P. 23, lines 36-46. Again, for alpha the issue is both that it is above zero, suggesting the discrepancy mechanism is involved, and less than one, suggesting a parameter is needed to weight that involvement.*

*16.     P. 25. The notion that the parameter values might clump by some known or unknown group property does not necessarily mean that the groups are using different processes. Different processes would suggest that the model in eq. 1 was only valid for some, not others. In the approach/avoid finding, it seems the findings imply that a more elaborated model that includes type of goal, or better a general model where type of goal would affect variables in the model, should be pursued. But to say it is a different process is a strong sentiment.*

In the original manuscript, we used the term “process” in the way that it has been used in the context of mixture modelling. In this context, the clustering of parameter values has often been described as being attributable to different “processes”. We see now how the use of this terminology was confusing, particularly because throughout the paper we use the term “process” to mean something much more precise.

*17.     P. 25. It seems the mix\_weight variable works if one or two mixtures, but not three or more. Am I missing something?*

This is correct. If there are three or more mixtures, the model requires a slightly different specification. We included a model that can accommodate three or more mixtures in the supplementary materials that were associated with the original manuscript. We did not include this section in the main text because we wanted to keep the number of models in the manuscript itself to a minimum. In the revised manuscript,….

*18.     I am a fan of qualitative model evaluation (what you call visual), but my biggest issue with this paper is the terseness of the quantitative model comparison section. The questions that researchers might be asking are often not at the parameter level, but at the model level, and while probabilistic, it seems Bayesian and especially the Bayes Factor, can help. For example, you tested 4 versions of the Gee et al model (single, person, known groups, unknown 2-groups). I want to know whether person level adding much information compared to single (would be surprised if it did not); if known groups added over person (trickier here because you did not “know” the groups by their priors); whether the known groups captured essentially the same information as the mixed 2-group implying that is was the goal-type context that was the primary sources of the unknown group difference. Bottom line though, one comes away from this paper thinking that the Bayesian approach will be provide an index of model fit. Yet, you mention the Bayes Factor briefly in one paragraph and then describe its challenges in the next couple of paragraphs. You never fit any of the models using the Gee et al. sample. It is like you ran out of steam. To be sure, I fear over-reliance on such summary statistics, but without some discussion, researchers will use ML or other models to derive such statistics. Is that what they should do? I am okay with pointing out the dangers of the indexes, but you do not really even give them a shot.*

*19.     On a related note, you mention that visual (and I would say other methods) might not properly respect the model fitting advantage complexity adds (p. 29, lines 12-17). You then suggest that parsimony should be a key determinant. I am sympathetic with that argument, but for me, scientific plausibility is the main factor to consider. Many have complained that connectionist models are not parsimonious (and they are not), but if weights are developed with experience and cannot be measured or modeled directly (e.g., include the “experiences” as inputs to the model), then they would need to be estimated from the data. This is not a flaw of the theory, but a reality of the measurement environment. One can make many a mathematical theory, but if the theory is not reflecting the entity represented, it is not scientifically plausible. Alternatively, if a set of functions does not work to produce a stable system that is a problem for the set of functions (unless we know the system being modeled is unstable). These are the most important ways to evaluate models. The more quantitative methods become more useful as knowledge in measurement properties and priors are increased. These are often very weak in organizational science and thus such fitting needs to be taken with a large grain of salt, but for a paper like this it seems you should at least explain what model fitting looks like and how to qualify conclusions based on model fit comparisons. Indeed, on page 30, line 57, you say that you demonstrate how to test a dynamic theory, but you never do really. You demonstrate how to estimate parameters for the theory. How do we test the theory?*

*20.     P. 30, lines 41-43. I cannot agree that “Testing a dynamic theory requires a statistical model that accurately reflects the processes described by the theory.” I would say that the statistical model should accurately quantify or qualify the diagnostic predictions stemming from the model. The problem with the Collins view is that it makes it seem that all the variables involved in the process need to be either measured or estimated as parameters (exogenous). Thus, any model where some latent processes are occurring could not be tested by that requirement. In the Gee et al. example, you did not have this problem. But what about the Ballard et al. (2016; 2018) models? There is more going on in those dynamic models that was measured, manipulated, or estimated. Indeed, when you say “As a result, researchers are often forced to distill rich, dynamic theories into simple predictions regarding the direction of relationships between variables” I have no issue with simple predictions. It is the “regarding the direction of relationships between variables” that is limiting. So in the Vancouver, Weinhardt, & Schmitt (2010) model, they needed a model that produced the reversal effect (allocate to goal with greater discrepancy then with smaller discrepancy if approaching deadline and not going to get both goals) and the incentive effect. These were the key postdictions that needed to be accounted for. The notion that one can boil the question of evaluating a theory/model to some set of predictions that diagnose the model, particularly in comparison to an alternative model that can predict only some of the set of predictions, seems how we evaluate theories in the past and moving forward (the Ballard et al. 2018 is a great example of this [can be found in online first at JAP]). Meanwhile, it does not matter if they are static or dynamic. The problem with “the direction of the relationship between two variables” prediction is that there are too many models that would create that prediction. What we need is unique predictions. But those can certainly be simple.*

We suspect there might be some confusion here resulting from our use of the term ‘statistical model’. In the original manuscript, we used the term ‘statistical model’ to refer to any model that can be applied to the analysis of data. This of course includes standard models such as latent growth curve or latent change score models. However, in our view, this category also includes computational models that are instantiated in such a way that they can be fit to data (which was the case in Ballard et al., 2018). These models can be considered ‘statistical’ insofar they can be used to estimate parameters from the data that describe components of the dynamic process. Under this view, the use of a ‘statistical model’ does not preclude the ability to operationalize latent or unobservable components of the process.

This comment helped us recognize however that the term ‘statistical model’ is indeed ambiguous, because computational models are not often thought of as statistical models. In the revised manuscript, we'

The Reviewer also makes a good point about the extent to which the predictions differentiate between alternative models being the important factor to consider. We agree that even simple predictions are capable of doing so. In the revised manuscript, we’ve edited the relevant paragraph so that it emphasizes this point, and have removed the suggestion that simple predictions are problematic.

*21.     P. 31, lines 12. Careful here. There are ways to use standard statistical procedures for testing dynamic models. They are not pretty and not used all that often because of the data requirements (though same requirements as your approach). But really unless you were to directly compare, seems a dangerous statement to make.*

The point we were trying to make in that sentence was specifically about dynamic (i.e., stock or level) variables. Our argument was that this type of variable cannot be implemented in conventional models (e.g., latent growth curve or latent change score models), because they require a model that needs to be simulated in order to generate predictions (e.g., a systems dynamics model). We weren’t arguing that dynamic models could not be tested using standard statistical procedures, but just that that particular type of variable could not be modelled.

In the revised manuscript,

*22.     P. 31, lines 31-34. So here you say that the multiple-groups model quantifies differences between known groups. Does it allow any conclusions regarding what is happening? Can we make causal conclusions or is cause a non-entity with dynamic models? My point is that quantifying parameter differences is not very scientifically interesting if we cannot even conclude that we can have some confidence in the model’s validity.*

We now see that we could have made the sentence in question clearer if we had added “…in the model parameters” to the end of it. The multiple-groups model can be used to quantify differences between groups in the model parameters. Of course, the extent to which doing so enables conclusions regarding what is happening depends on the extent to which the model provides an accurate representation of the psychological process being investigated. If the model is valid in the sense that its parameters accurately reflect some psychological quantity, then differences in the model parameters are likely to be of interest. If the model is not valid, then as the Reviewer suggests, differences in the parameters are less meaningful. Model validity can be assessed by considering the psychological plausibility of the model, examining the extent to which the predictions account for the data, and comparing the model to plausible alternatives.

Regarding the question of causality, as with statistical models, this depends on how the groups were generated. For example, if the groups represent conditions in an experiment to which participants were randomly allocated, there will be reasonably strong grounds for making causal claims. On the other hand, if groups were formed, for example, by measuring some variable and performing a median split, the grounds for making causal claims will be less strong. The point is that causal inferences are handled the same way in this framework than in most statistical approaches.

*23.     You have some editing issues, like the last line of the paper is grammatically incorrect.*

Thank you for bringing these issues to our attention. We have fixed the last line in the paper, and have carefully proofread the entire manuscript prior to resubmitting.   
  
*Overall I liked this paper and I think it is important and timely. But there needs to be more on model comparison/testing than some use the Bayes Factor.*