An integrated approach to testing dynamic, multilevel theory: Using computational models to connect theory, model, and data

Timothy Ballard[a], Hector Palada[a], Mark Griffin[b], & Andrew Neal[a]
[a]The University of Queensland
[b]Curtin University

Abstract

Some of the most influential theories in organizational sciences explicitly describe a dynamic, multilevel process. Yet the inherent complexity of such theories makes them difficult to test. These theories often describe multiple sub-processes that interact reciprocally over time, at different levels of analysis, and over different time scales. Computational (i.e., mathematical) modeling is increasingly advocated as a method for developing and testing theories of this type. In organizational sciences however, efforts that have been made to test models empirically are often indirect. We argue that the full potential of computational modeling as a tool for testing dynamic, multilevel theory is yet to be realized. In this paper, we demonstrate an approach to testing dynamic, multilevel theory using computational modeling. The approach uses simulations to generate model predictions, and Bayesian parameter estimation to fit models to empirical data and to facilitate model comparisons. This approach enables a direct integration between theory, model, and data, that we believe enables a more rigorous test of theory.

Keywords: Dynamic theory | Computational modeling | Multilevel research | Bayesian parameter estimation | Self-regulation

An integrated approach to testing dynamic, multilevel theory: Using computational models to connect theory, model, and data

Many organizational phenomena are dynamic in nature, meaning that they evolve over time (Lord, Diefendorff, Schmidt, & Hall, 2010; Neal, Ballard, & Vancouver, 2017; M. Wang, Zhou, & Zhang, 2016). Examples include individual motivation, performance and well-being; team cohesion

T. B. was supported by an ARC Discovery Early Career Research Award (DE180101340). H. P. was supported by an Australian Government Research Training Program Scholarship. A. N. was supported by an ARC Discovery Project (DP150102658). We thank Michael D. Wilson for providing feedback on the manuscript. The data and code used to conduct all the analyses presented in this paper are publicly available, and can be found at https://osf.io/4euhr/.

and effectiveness; and organizational capability and culture. There are many theories that seek to describe how these phenomena emerge and change from one time point to the next. Examples include theories of self regulation (Kanfer & Ackerman, 1989; Neal et al., 2017), the reoccurring phase model of team processes (Marks & Mathieu, 2001), the cultural dynamics model (Hatch, 1993), and McKinley, Latham, and Braun's (2014) model of organizational decline and innovation. Although dynamic theories such as these have been influential, they are difficult to test. These theories propose a set of component processes that interact reciprocally over time and that may not be directly observable. The challenges of testing dynamic theory are magnified when the theorized processes operate at multiple levels of analysis. The behavior of the system as a whole is often an emergent property of the interactions amongst the lower-level processes. Such phenomena are difficult to assess using traditional multilevel methods (Kozlowski, Chao, Grand, Braun, & Kuljanin, 2013).

Computational modeling has been increasingly advocated as a method for developing and testing complex theories of this type (e.g., Ballard, Yeo, Loft, Vancouver, & Neal, 2016; Harrison, Lin, Carroll, & Carley, 2007; Rudolph, Morrison, & Carroll, 2009; Salas, Kozlowski, & Chen, 2017; Tarakci, Greer, & Groenen, 2016; Vancouver, Wang, & Li, 2018; M. Wang et al., 2016). Computational modeling is the practice of articulating theory in the form of mathematical equations and/or computer code and evaluating the dynamic behavior of the theory by simulating the model. This approach offers unique advantages to researchers attempting to understand multilevel phenomena. Simulation of the theory enables the researcher to examine the dynamic behavior that emerges at different levels of analysis as the system is played out over time. For example, Grand (2017) used a computational model to demonstrate that stereotype threat, which has a relatively small effect on performance at the individual level (in statistical terms), has a substantial negative effect on performance at the organizational level. This practice can be used to assess both top-down and bottom-up effects, as well as processes that unfold over different time scales.

In this paper, we present an integrated approach to testing dynamic theories of multilevel organizational phenomena using computational modeling. Figure 1 shows the three key elements required for testing dynamic theories. The first element is a theory of the change process. Such a theory describes how the system evolves from one time point to the next. The theory identifies the dynamic variables in the system and describes the underlying mechanisms that explain how these variables change over time. The second element is a mathematical or statistical model that operationalizes the key components of the dynamic theory. The third element is empirical data that is collected by observing the system over time. Rigorous theory testing requires the seamless integration of all three components (Collins, 2006).

The upper half of Figure 1 depicts the most common way of linking theory, model, and data. Under the conventional approach, a conceptual model (A) translates the theory into a set of verbal hypotheses that summarizes the relationships between variables that are theorized to be involved in the change process. The conceptual model is tested against empirical data using a statistical model (B) that maps as closely as possible to the conceptual model. For example, Li, Fay, Frese, Harms, and Gao (2014) used a conceptual model to generate a series of hypotheses regarding the reciprocal effects between personality and work characteristics and tested these hypotheses using a latent change statistical model. The challenge with this approach, however, is that it often difficult to directly represent the complex processes described by dynamic theories in a conventional statistical model. One reason for this is that statistical models require at least one observed variable to map

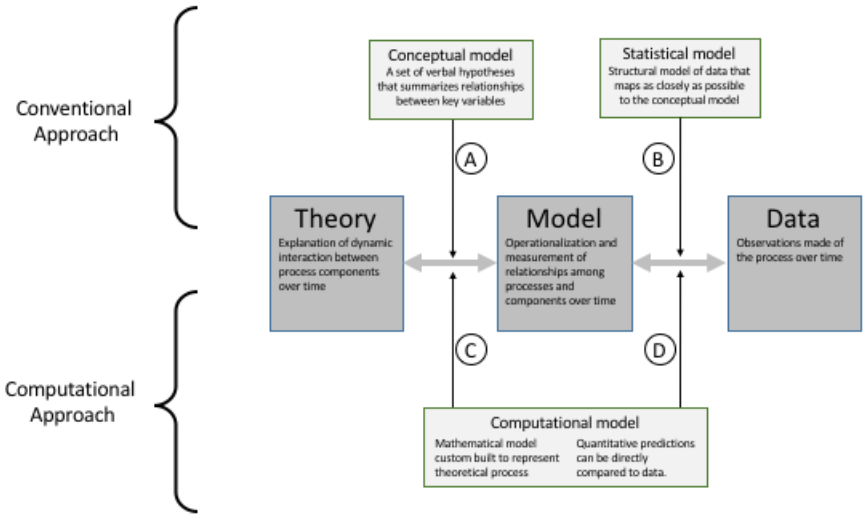

Figure 1. Integration of theory, model, and data under conventional and computational modeling approaches to dynamic theory testing.

to each latent construct. However, the change process often involves constructs that may be hard to measure (e.g., skill acquisition; Kanfer & Ackerman, 1989). This can make it hard to achieve a close mapping between theory and model, which ultimately makes the theory hard to test using standard methods.

We propose that an approach based on computational modeling, depicted in the lower half of Figure 1, integrates theory, model, and data more directly. The value of this approach comes from the fact that computational modeling allows for direct mapping between theory and model (C), because the model is custom built to represent the process described by the theory. At the same time, computational models produce quantitative predictions that can be directly compared to empirical data (D). However, the efforts that have been made to test computational models in organizational sciences are often relatively indirect. One common approach is to examine whether the model can reproduce key results from previous empirical research (e.g., Kennedy & McComb, 2014; Vancouver & Purl, 2017). For example, Vancouver et al. (2018) found that their integrated computational model of work motivation reproduced Latham and Baldes's (1975) finding that employees respond to increases in goal difficulty by applying more effort. Another common approach is to simulate data from a computational model, and then compare differences in the results of statistical tests run on the simulated data with the results of tests run on data collected from actual participants (e.g., Grand, Braun, Kuljanin, Kozlowski, & Chao, 2016). For example, Lin, Yang, and Demirkan (2007) implemented a computational model to explore whether firm ambidexterity was related to firm performance. They compared regression analyses of the simulated data with regression analyses of eight years of empirical data to compare the effects of contingency factors such as firm size. The challenge with these approaches is that they can only be used to determine whether the model produces the same qualitative effects that are present in the empirical data. They cannot be used

to assess how precisely the predictions of the model correspond to the data, which is important for comparing alternative models. The latter approach also limits the set of possible predictions that can be tested to those which can be assessed using standard statistical tests (e.g., linear relationships between variables).

Although there is much to be gained from the approaches described above, a more comprehensive test of the theory can be achieved by going a step further and fitting the computational model directly to the data (Farrell & Lewandowsky, 2018). Fitting the model to the data enables the model parameters to be informed by the empirical observations, just as would be the case with coefficients in a regression model. An advantage of this approach is that it enables the correspondence between the data and the model predictions to be quantified, making it easier to conduct model comparisons that rule out alternative explanations. In this paper, we propose and demonstrate a four-step approach to testing a computational model that involves: 1) simulation, 2) model fitting, 3) model comparison, and 4) interpretation of the parameter estimates.

In the next section, we elaborate on the challenges associated with testing dynamic theories of multilevel phenomena and the associated advantages of computational modeling. Following that, we introduce an example research question that we use to demonstrate our approach. The research question is derived from the resource allocation theory of self-regulation (Kanfer & Ackerman, 1989), and concerns the way that people adjust goals and effort over time. This phenomenon involves a set of interacting processes that unfold at different levels of analyses, a mix of top-down and bottom-up multilevel effects, and at least one internal process that is not directly observable. We begin by describing the theory, and translating it into a computational model. We then demonstrate how to test this model using the four-step approach described above.

A four-step approach to testing dynamic theory

A dynamic system is one that changes over time (Galor, 2007; Luenberger, 1979). Testing dynamic theories in the organizational sciences poses some specific challenges that can be addressed by computational modeling. One challenge is the existence of feedback loops in which the outputs of the system influence the inputs to the system at some future point in time. Theories involving dynamic systems seek to understand the how the processes that form the feedback loop unfold over time. For example, McKinley et al. (2014) proposed a theory of organizational decline in which a spiral of declining performance could be counterbalanced by innovation activities. This theory describes a feedback loop in which a decline in performance leads to innovations that in turn either accelerate or counteract the downward performance trajectory.

Dynamic theories are also often multilevel in nature. They often incorporate loops that operate at multiple levels of analysis. For example, the self-regulatory system is often conceptualized as a set of nested loops operating at different levels (Carver & Scheier, 1998; Powers, 1973). The lower-level loop governs the allocation of effort as the person pursues the goal. During the goal striving episode, effort is applied in response to a discrepancy between the goal and one's current level of performance, which enables the person to make progress towards their goal. This process operates within goal striving episodes. The higher-level loop adjusts the difficulty of the goal that the person pursues across successive goal striving episodes. This process operates between goal striving episodes. The behavior of the system as a whole reflects a mix of processes that cut across these different levels of analysis. Effort exerts an effect on performance, which in turn exerts an

effect on goals (a bottom-up process), while goals exert an effect on effort and performance (a top-down process). These processes may, in turn, shape and be shaped by processes at higher levels (e.g., the individual, team or organizational level).

As a more macro level example, the literature on organizational routines attempts to understand the dynamic, multilevel process by which patterns of action and interaction between individuals within the organization influences the practices that form at the firm level (Abell, Felin, & Foss, 2008; Pentland, Feldman, Becker, & Liu, 2012). In this context, the lower level of analysis represents the actions taken by individual actors and the higher level represents the outcomes for the organization as a whole. The emergence of organizational routines at the higher level of analysis is thought to be largely influenced by processes of learning and selective retention that operate at the lower level.

Within the organizational sciences, dynamic theories also incorporate internal mechanisms that cannot be observed or measured directly, but which are used to explain observed phenomena. For example, the resource allocation theory of self-regulation uses the skill acquisition process to explain why the relationship between effort and performance changes over time (Kanfer & Ackerman, 1989). Specifically, it is assumed that the reason why less effort is required to achieve the same level of performance after practicing a task is because the person has become more skilled. Skill acquisition is an internal mechanism that cannot be directly observed independently of performance and effort. Similarly, in the organizational routines literature, the process of action retention is an internal mechanism that must be inferred based on the patterns of actions observed (Feldman & Pentland, 2003).

It is because of the complexities described above that many researchers have advocated the use of computational modeling for developing and testing dynamic, multilevel theory (Kozlowski et al., 2013; Weinhardt & Vancouver, 2012). Computational modeling allows researchers to formally (i.e., mathematically or via propositional logic) represent the rules that govern how the system changes, and examine the dynamic behavior of the system at each level of analysis as it is played out over time. As such, this approach is well-suited to theories that contain the dynamics described above. In the sections below, we demonstrate a four-step approach to testing theory using computational models. The first is to conduct a simulation study (sometimes referred to as a virtual experiment) in order to demonstrate the model's *sufficiency*. Sufficiency refers to the ability of the assumptions that form that model to reproduce the pattern of behavior that that theory purports to explain (Epstein, 1999; Farrell & Lewandowsky, 2018; Fum, Del Missier, & Stocco, 2007).

The second step is to conduct a quantitative test of the model by fitting the model to the empirical data. Fitting a model involves using the empirical data to estimate the values of the model's parameters and generating predictions from the model based on the estimated parameter values. This gives the model the best possible chance of accounting for the data. We use Bayesian parameter estimation for model fitting, because it provides a robust method for estimating the parameters of complex models. It also provides information about the uncertainty inherent in the parameter estimates, making it easy to compare parameter values across experimental conditions or participant groups and to examine the range of observations that are most probable according to the model. The Bayesian approach also offers useful ways of quantifying model-data correspondence that facilitate model comparison.

The third step is to compare the model to alternatives in order to rule out competing expla-

nations. This step allows the researcher to establish the necessity of the model's assumptions, for example, by demonstrating that the model can no longer account for the empirical trends when these assumptions are changed or removed (Farrell & Lewandowsky, 2018). The fourth step is to interpret the estimated values of the model parameters. This allows the researcher to quantify latent components of the dynamic process based on the empirical data. These quantities can then be compared across experimental conditions or participant groups.

In the next section, we begin our demonstration by describing a theory of dynamic self-regulation and representing it as a computational model. We then describe the longitudinal data that we use to test the model. In the following sections, we evaluate our theory using the approach described above. It is important to note that the approach we demonstrate below is a general one. It is not limited to the theory or model we consider in this paper, or the levels of analysis at which the processes described by the theory operate. We return to the issue of generality in the discussion.

A theory and model of dynamic self-regulation

To illustrate the approach, we show how it can be used to model a set of reciprocal processes involving both top-down and bottom-up multilevel effects unfolding over different time scales. The problem that we focus on is dynamic self-regulation. We start with a brief description of the theory that we are drawing on. We then translate the theory into a computational model and evaluate the extent to which that model can account for data collected in an experiment. Note that the primary aim of the analysis presented is to demonstrate the approach to testing computational models. We acknowledge that there are likely many plausible ways in which the processes described below can be represented, and that more work will be needed to refine the more nuanced assumptions of the model. The goal here is not to propose a definitive model of self-regulation, it is to demonstrate how this approach can be applied to test theories of complex, multilevel phenomena.

Theory

Self-regulation theories describe the process by which people interact with their environment to achieve their goals (Neal et al., 2017). In this analysis, we focus on two interrelated processes: a) the process by which people adjust the level of effort that they apply whilst striving for a goal; and b) the process by which they adjust the difficulty level of the goals that they pursue. The type of effort that we focus on is perceived mental effort, which is the subjective feeling of trying hard (Humphreys & Revelle, 1984). Theories of self-regulation typically assume that effort is a limited capacity resource (Kanfer & Ackerman, 1989), and that it is regulated by a negative feedback loop (Vancouver, 2005). Specifically, it is assumed that the person has a goal that represents the level of performance that they want to achieve, and they monitor the discrepancy between the desired level of performance and the current level of performance, and adjust their effort accordingly. This discrepancy is commonly referred to as the *goal-performance discrepancy*. Larger goal-performance discrepancies indicate that more effort is needed to reach the goal, and therefore result in more effort being applied (Brehm & Self, 1989; Kruglanski et al., 2012; Wright, 2008). As a result, people are expected to reduce their effort when the goal-performance discrepancy decreases [1].

[1] An anonymous reviewer commented that the subjective experience of effort may not always correspond to objective indicators of effort, such as the allocation of attentional resources. This issue relates to a long-standing debate over the

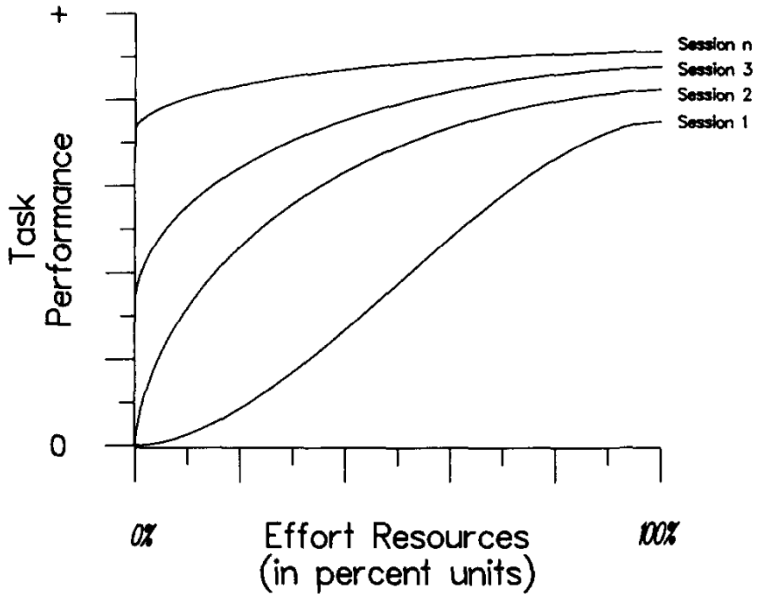

Figure 2. Effects of effort and on task performance throughout skill acquisition. From "Motivation and Cognitive Abilities: An Integrative/Aptitude-Treatment Interaction Approach to Skill Acquisition," by R. Kanfer and P. L. Ackerman, 1989, *Journal of Applied Psychology, 74*, p. 660. Copyright 1989 by the American Psychological Association.

Theories of skill acquisition assume that the amount of effort required to achieve a given level of performance changes as the person becomes more skilled at the task (Kanfer & Ackerman, 1989). During the early phases of skill acquisition, most people will require a relatively high level of effort to achieve a modest level of performance (see Figure 2). For this reason, performance during the early phases of skill acquisition is said to be "resource limited", because it is sensitive to the amount of resources that are applied (i.e., effort; Norman & Bobrow, 1975). As the person becomes more skilled at the task, they require less effort to achieve the same level of performance. Performance during the later phases of skill acquisition is said to be "data-limited", because it is constrained by the nature of the task (Norman & Bobrow, 1975).

Theories of self-regulation also assume that people adjust their goals based on past performance. If a person fails to achieve their goal, they may lower their goal for the next attempt, whereas if they achieve their goal, they may raise it (Donovan & Williams, 2003). Adjustments to the goal level have an impact on subsequent effort, because they affect the goal-performance discrepancy. So, whilst the amount of effort that the person applies may decrease over time as the person becomes more skilled at the task, this may be counteracted by an increase in effort caused by the individual setting more difficult goals. The amount of effort that the person applies at any point in time reflects a mix of these processes, which unfold over different time scales.

---

measurement of effort and the relative merits of subjective measures (of the type used here) by comparison with objective measures (either behavioral or physiological; Humphreys & Revelle, 1984; Tsang & Wilson, 1997). Although these issues are important to consider, they are beyond the scope of the current paper.

Model

In this section, we translate the theory presented above into a dynamic, computational model. In developing the model, we assume that the person performs a task that is broken into a series of discrete goal striving episodes, and that prior to each episode, the person sets a goal for their output in the next goal striving episode. This goal is then revised prior to the next episode. A summary of the model is presented in Figure 3. In this diagram, observed variables are represented as shaded squares, unobserved variables are represented as white squares, and estimated parameters are represented as white circles. The figure uses plates to differentiate levels of analysis. The *observation* plate represents the lowest level of analysis. Variables positioned within this plate (i.e., practice, skill, effort, and performance) vary across observations within each goal striving episode. The *episode* plate represents a higher level of analysis. Variables positioned within this plate (but not within the observation plate; i.e., goal) vary across goal striving episodes but do not vary within each episode.

As can be seen, the model assumes that effort is determined by the goal level and the performance variable. Effort in turn combines with skill, which is an unobserved variable that increases with practice, to influence performance. The dynamics of effort and performance over time within each goal striving episode feed through to influence the higher-level goal revision process that operates across goal striving episodes. The model therefore captures the top-down effects of goal on effort and performance as well as the bottom-up effects of these variables on goal level. The model also assumes that there are several free parameters that affect these processes, which are explained in more detail below. As can be seen, the effort, performance, and goal level nodes all include arrows that emerge from and return to that same variable. These *self-feedback* loops indicate that the levels of these variables depend in part on their previous values. In the sections below, we elaborate on the details of the model. We begin by describing the model of the effort regulation process. We then describe how the model accounts for skill acquisition. We then describe how effort and skill determine the level of performance that is achieved. Finally, we describe how the goal revision process is modeled.

Effort Regulation. The theory presented above suggests that during each goal striving episode the person regulates the amount of effort that they apply to the task. Effort is a dynamic variable, because it is adjusted from one time point to the next, with its value at a given time point being in part determined by its previous value. Formally, the level of effort that the person applies at point *i* within episode *j*, denoted $Effort_{i,j}$, is represented as follows:

$$Effort_{i,j} = Effort_{i-1,j} + \Delta Effort_{i,j}, \tag{1}$$

where $Effort_{i-1,j}$ is the level of effort at the previous point in time, and $\Delta Effort_{i,j}$ is the change in the level of effort. The process that produces changes in effort is represented as follows:

$$\Delta Effort_{i,j} = \alpha \times (Effort_{bl} - Effort_{i-1,j}) + \beta \times (GPD_{i-1,j} - GPD_{i-2,j}). \tag{2}$$

Our model assumes that changes in effort are governed by two factors. The first is a return to baseline process whereby effort, in the absence of any changes in goal-performance discrepancy,

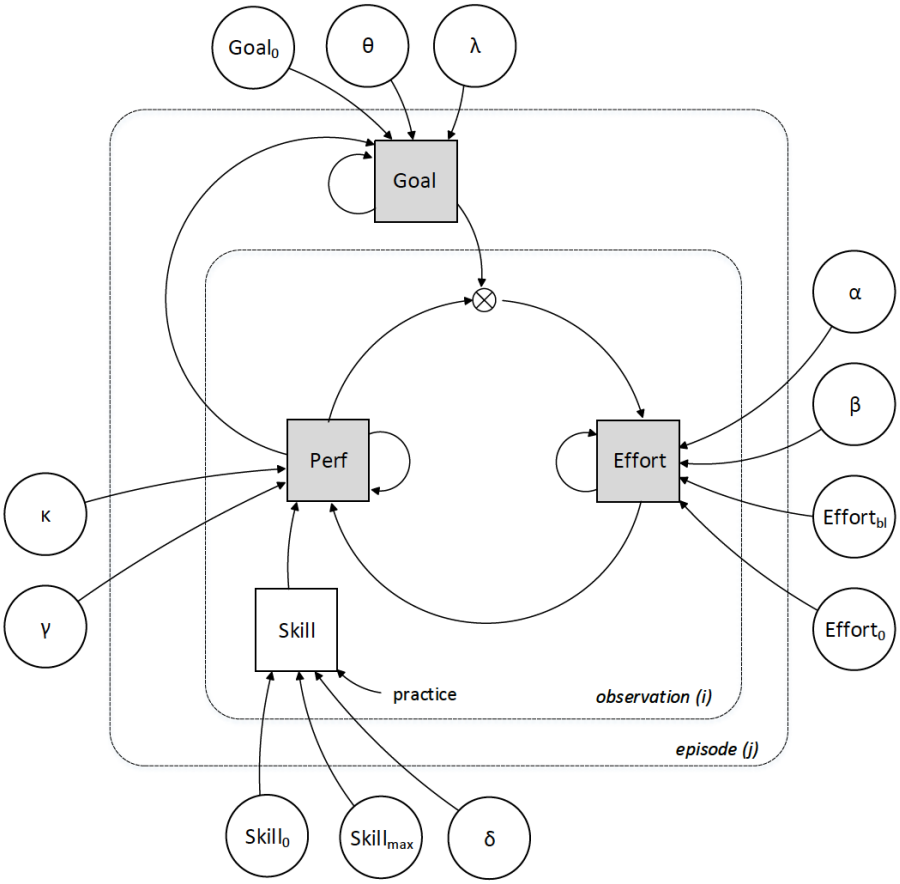

Figure 3. Summary of the computational model of dynamic self-regulation. Observed variables are represented as shaded squares, unobserved variables are represented as white squares, and estimated parameters are represented as white circles. The observed goal variable within the observation plate varies within and between goal striving episodes. Variables within the episode plate (but not the observation plate) vary only between goal striving episodes. Variables outside the episode plate do not vary over time. Variables with self-feedback loops are dynamic variables, which have levels that are in part determined by their previous values.

returns to a neutral point over time. This return to baseline process is represented in the first term of Equation 2, where $Effort_{bl}$ is the baseline level to which effort returns, and α is the rate at which effort returns to baseline. α can take on any value between 0 and 1, where higher values produce a more rapid return to baseline.

The second factor that governs changes in effort is the change in goal-performance discrepancy. The process by which effort responds to changes in goal-performance discrepancy is represented in the second term of Equation 2. $GPD_{i-1,j}$ and $GPD_{i-2,j}$ represent the goal-performance discrepancy at the previous two time points, which means that $GPD_{i-1,j} - GPD_{i-2,j}$ is the change in goal-performance discrepancy at the previous time point. β is a gain parameter that represents the person's sensitivity to changes in the goal-performance discrepancy. If β has a value of 0, the person is insensitive to changes in goal-performance discrepancy. However, based on the theory described above, we would expect that decreases in the goal-performance discrepancy should produce a decrease in effort. We therefore expect β to take on a positive value.[2]

Skill Acquisition. According to the"law of practice", skill increases over time as a function of the time spent practicing a task (Newell & Rosenbloom, 1981). The bulk of current evidence suggests that the relationship between practice and skill is exponential (Heathcote, Brown, & Mewhort, 2000; Leibowitz, Baum, Enden, & Karniel, 2010). This exponential relationship can be represented as follows:

$$Skill_{i,j} = Skill_{max} - (Skill_{max} - Skill_0) \times e^{-\delta \times Practice_{i,j}}, \tag{3}$$

where $Practice_{i,j}$ represents the total amount of practice that the person has had up to the current point in time. According to Equation 3, the person begins the task with some initial level of skill, which is denoted $Skill_0$, and skill increases over time. The relationship between practice and skill is non-linear, with large gains in the early phases of skill acquisition giving way to progressively smaller gains as the level of skill approaches the asymptote, which is denoted $Skill_{max}$. δ represents the learning rate, which controls how rapidly skill is acquired. This parameter can take on any value between 0 and 1, with higher values producing skill acquisition curves that approach the skill asymptote more rapidly.

Performance. Performance, like effort, is a dynamic variable. The total amount of output that the person has produced at the current point in time ($Perf_{i,j}$) is the sum of the total output at the previous time point ($Perf_{i-1,j}$) and the new output produced during the current time interval ($\Delta Perf_{i,j}$):

$$Perf_{i,j} = Perf_{i-1,j} + \Delta Perf_{i,j}. \tag{4}$$

The theory presented above suggests that the amount of output that the person produces during the current time interval is influenced by both skill and effort, as illustrated in Figure 2. It would

[2] An anonymous reviewer pointed out that another way of modeling effort would be to assume that effort remains at the same level, regardless of the magnitude of the goal-performance discrepancy, so long as the goal has not been achieved. We agree that this model cannot be ruled out. Distinguishing between the two models would require a more targeted experiment that is specifically designed to examined the relationship between goal-performance discrepancy and effort. For example, such an experiment might examine how effort responds to disturbances that change the goal-performance discrepancy during goal striving. These issues, however, are beyond the scope of the current paper.

be possible to approximate the family of curves shown in Figure 2 using a linear function relating effort to performance, with skill controlling the intercept of the function and effort controlling the slope. However, this would provide an incomplete representation of the underlying theory because a linear function cannot produce the performance asymptote that is proposed to arise when further increases in effort do not produce any increase in performance. For this reason, we model the effort-performance relationship using a logistic function, which is a non-linear function that allows for performance to reach an asymptote with increasing effort. The effort-performance relationship is modeled as follows:

$$\Delta Perf_{i,j} = \frac{Skill_{i,j}}{1 + e^{-\kappa \times (Effort_{i,j} - \gamma)}}. \tag{5}$$

The numerator in Equation 5 controls the performance asymptote, which determines the maximum amount of output that the person can produce at the current time point. In other words, this is the amount of output that would be produced if the person was to exert the maximum level of effort. As can be seen in the equation, the performance asymptote is determined by the person's skill level. As the person becomes more skilled, the maximum amount of output that they can produce increases.

The nature of the relationship between effort and output is determined by the κ and γ parameters. κ controls the steepness of the function. Higher values of κ mean that smaller changes in effort are necessary to achieve greater changes in performance. γ controls the location of the function. This parameter represents the amount of effort required to produce a change in performance that is equal to half of the maximum performance change. Equation 5 produces effort-performance curves like those shown in Figure 2, where the height of the asymptote increases as more skill is acquired.

Goal Revision. The final component of the model is the bottom-up process by which goal level is adjusted over time. The goal level set at the start of the goal striving episode ($Goal_j$) is determined by the goal level set at the start of the previous goal striving episode ($Goal_{j-1}$) and the extent to which they adjust their goal level ($\Delta Goal_j$):

$$Goal_j = Goal_{j-1} + \Delta Goal_j. \tag{6}$$

Fortunately, there is an existing computational model that describes how people make these adjustments (Gee, Neal, & Vancouver, 2018). There are two components to the process described by Gee et al.'s model. First, their model assumes that the person adjusts their goal level by an amount that is proportional to the discrepancy between the previous goal and the previous level of performance. This process anchors the goal on the level of performance that the person can expect to achieve, given their history of performance across previous goal striving episodes. Second, the model assumes that the adjustment of the goal around this anchor point is biased by the person's risk preference. A person with a high risk preference will tend to set a goal higher than what they can expect to achieve given previous performance, while a person with a low risk preference will tend to set a goal lower than what they can expect to achieve given previous performance. The goal adjustment can be represented as follows:

$$\Delta Goal_j = \theta \times (Perf_{j-1} - Goal_{j-1}) + \lambda. \tag{7}$$

In the above equation, $Perf_{j-1}$ represents the total amount of output produced by the end of the previous goal striving episode. θ controls the sensitivity of goal level to the goal-performance discrepancy at the end of the previous goal striving episode. θ is bounded between 0 and 1. When θ is low, the person is relatively insensitive to the goal-performance discrepancy, and so will be slow to raise their goal as performance improves over time. If theta is high, then the person will be more sensitive to the goal-performance discrepancy, and goal level will tend to change more rapidly. If theta is too high, then the person may be overly sensitive to random variation in performance from one episode to the next. λ represents risk preference, and controls the extent to which the person sets a goal above or below their anchor point.

Data

To test the model described above, we use data from the recent study by Gee et al. (2018). In their study, 60 participants each performed ten, 10-minute trials of an air traffic control simulation task. Each trial represented an independent goal striving episode. In each trial, participants had to classify ten aircraft pairs as a conflict or non-conflict based on their minimum distance of separation. Before each trial, participants set a goal for their classification performance on the upcoming trial. During the trial, the received live feedback regarding their progress toward the goal.

Each trial was broken down into five, 2-minute time windows. At the end of each window (after two, four, six, eight minutes and at the end of the trial), perceived effort was measured by asking participants to rate how hard they were trying during the previous two minutes on a scale from 0 (Not at all) to 10 (Extremely). At the end of each window, the number of decisions the participant had gotten correct up to that point in that trial was recorded. Thus, each participant provided fifty performance and effort observations (five for each of the ten goal striving episodes) and ten goal level observations (one for each episode).

Step 1: Simulating the Model

The first step in evaluating the model is to conduct a simulation study. A simulation study involves fixing the model parameters to predefined values and examining the output that emerges from the model under those conditions (Ballard et al., 2016; Ballard, Yeo, Vancouver, & Neal, 2017; Grand et al., 2016; Vancouver et al., 2018). Simulation is an important first step in evaluating a computational model for several reasons. First, it allows the researcher to demonstrate the sufficiency of the assumptions that form the model assumptions to account for the behavioral phenomena. For example, Vancouver, Weinhardt, and Schmidt (2010) developed a computational model that explained an earlier empirical result reported by Schmidt and DeShon (2007). Schmidt and DeShon observed that people tended to shift over time from prioritizing goals with larger discrepancies to goals with smaller discrepancies. Vancouver et al. (2010) showed that their model reproduced this effect, and interpreted this finding as evidence of their model's sufficiency.

Second, simulation is useful for examining the effects that individual parameters have on the output generated by the model. To do this, researchers manipulate individual parameter values whilst holding the other parameters constant and examine the effects that changes in the parameter being manipulated have on the model output. This exercise is referred to as a sensitivity analysis and is useful for understanding the process components that are responsible for a generating particular

phenomena (Davis, Eisenhardt, & Bingham, 2007; Fum et al., 2007; Vancouver & Weinhardt, 2012). Third, simulation allows the researcher to generate quantitative predictions from the model about how the process plays out over time that can be subjected to empirical examination.

We present a simulation of the above model that was conducted in R (R Core Team, 2018). The scenario we simulated mirrored the empirical protocol described above. The simulation was run over 50 time steps, with each time step representing a 2-minute window in the empirical protocol. The simulation tracked four variables over time: effort, skill, performance, and goal level. As in the empirical protocol described above, the simulation was divided up into ten unique goal striving episodes, that each lasted five time steps. Thus, goal level was updated every five time steps (i.e., at the start of each new goal striving episode). Effort, skill, and performance were updated every time step.

We chose parameter values that were reasonable based on the scale of the performance variable and the effort measure (which could each range from 0 to 10), and the predictions we made above regarding the sign of certain parameter values. These parameter values are shown in Table 1. In addition to specifying the parameter values from Equations 1-5, we also had to select initial values for the effort and goal variables. These initial values are also shown in Table 1. Note that the performance variable has an initial value of 0 due to the nature of the experimental protocol (i.e., people begin each goal striving episode with 0 correct decisions).

Table 1
Description of Model Parameters and Simulated Values

Relevant Variable	Parameter	Description	Allowable Values	Simulated Value
Effort	$Effort_{bl}$	Effort baseline	0-10	4
	α	Rate of return to baseline	0-1	0.3
	β	Sensitivity to change in GPD	any value	0.1-2.1
	$Effort_0$	Effort at start of episode	0-10	7
Skill	$Skill_0$	Skill level at start of first episode	any positive value	0.1
	$Skill_{max}$	Maximum achievable skill level	any value greater than $Skill_0$	3
	δ	Learning rate	0-1	0.1
Performance	κ	Sensitivity of performance to effort	any value	1
	γ	Effort required for half of max performance change	any value	5
Goal	θ	Sensitivity to previous episode GPD	0-1	0.5
	λ	Risk preference	any value	0
	$Goal_0$	Goal level for first episode	0-10	4

To demonstrate how sensitivity analysis can be used to examine the influence that certain model parameters on the dynamic behavior of the model, we manipulated the β parameter across six different levels (0.1, 0.5, 0.9, 1,3, 1.7, and 2.1). As described above, β is a gain parameter that controls the sensitivity of the effort adjustment process to changes in goal-performance discrepancy. Higher sensitivity means that effort reduces more rapidly in response to reductions in goal-performance discrepancy. We focus on β because, as will become apparent, this parameter has important implications for the bottom-up process by which effort regulation influences the higher-level goal revision process. In practice, researchers conducting a sensitivity analysis will typically examine a larger number of parameters. To keep our demonstration simple however, we limit our sensitivity analysis to a single parameter.

The results of the simulations are shown in Figure 4. The figure shows a 4×10 grid of panels. Each row of panels in the grid represents a different variable that is generated by the simulation. Each column of panels represents a unique goal striving episode. Within each panel, the x-axis

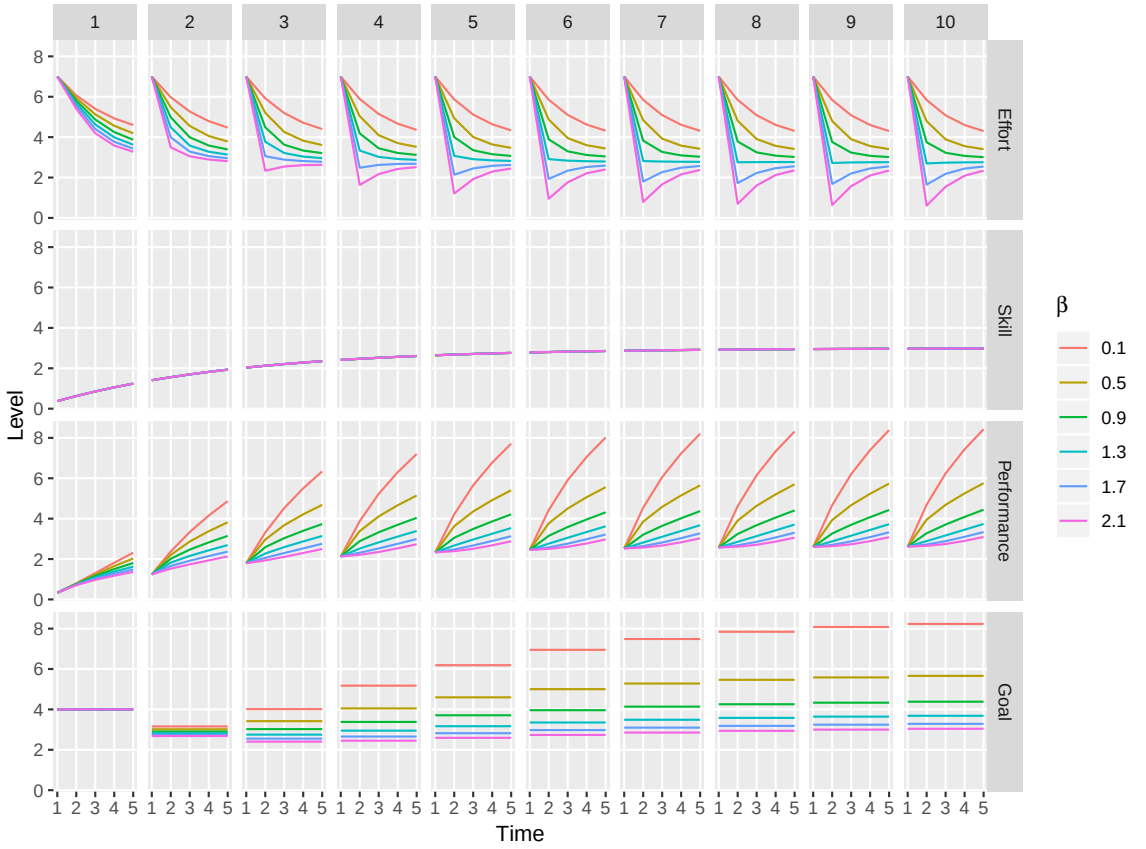

Figure 4. Results of sensitivity analysis examining the effects of sensitivity to goal-performance discrepancy on effort adjustment, performance, and goal revision. Each row of panels in the grid represents a different variable that is generated by the simulation. Each column of panels represents a unique goal striving episode. Within each panel, the x-axis represents the time point within the goal striving episode and the y-axis represents the level of the relevant variable at that time. The line color represent the different levels of the β parameter, with red signifying the lowest value (0.1) and purple signifying the highest value (2.1).

represents the time point within the goal striving episode and the y-axis represents the level of the relevant variable at that time. Finally, the line color represents the levels of the β parameter, with red signifying the lowest value (0.1) and purple signifying the highest value (2.1).

The simulations suggest that effort generally decreases over the course of each goal striving episode as progress is made toward the goal. The sensitivity of effort to changes in the goal-performance discrepancy has implications for performance within an episode and for goal revision across episodes. When sensitivity (β) is lower, performance gains during goal striving lead to only relatively small decreases in effort. This leads to more effort being applied overall, which results in a higher level of performance at the end of the episode. When sensitivity is higher, performance gains lead to more pronounced reductions in effort. This leads to less effort being applied, which results in poorer performance. At very high levels of sensitivity (e.g., when $\beta > 1.5$), performance gains lead to such pronounced reductions in effort that effort dips below its baseline level and then must

increase in order to return to the baseline. This scenario produces the worst performance outcomes.

The effort adjustment process exerts bottom-up effects on the dynamics of goal revision that emerge across the series of goal striving episodes. The goal for the first episode is not achieved under any of the six levels of the sensitivity manipulation. This can be seen by the fact that the maximum performance for any of the six levels in the first episode (approximately 2; shown in the first column, third row) is lower than the goal for the first episode (4; shown in the first column, fourth row). The failure to achieve the goal in the first episode leads to the goal for the second episode being set at a lower level than the first (see second column, fourth row). However, the trajectory of the goal variable for the third episode onward depends on effort sensitivity. When sensitivity is lower, a pattern of upward goal revision emerges. This happens because lower effort sensitivity leads to better overall performances in each episode, which means that the goal tends to be exceeded and increased from one trial to the next. On the other hand, when sensitivity is higher, little or no upward goal revision is observed. This happens because higher effort sensitivity leads to poorer performances in each episode, which means that goals are typically not exceeded.

To conclude this section, we conducted a preliminary test of the theory by examining simulated data from the model. The plausibility of the assumptions that form the model were evaluated by assessing the output produced using the parameter values in Table 1. The results presented in Figure 4 suggests the model produces a plausible pattern of effort adjustment and goal revision. We can interpret this result as preliminary evidence for sufficiency of our model to account for the dynamic self-regulatory processes described by the theory. We also carried out a sensitivity analysis to examine the effects of the effort sensitivity parameter on the model output. This analysis revealed that the relationship between goal-performance discrepancy and effort within each goal striving episode has bottom-up effects on goal revision that occurs between episodes. Finally, the simulation revealed a novel prediction for future empirical research. Specifically, the model predicts that the sensitivity of effort to changes in goal-performance discrepancy might be negatively related to the emergence of upward goal revision.

Step 2: Fitting the Model to Empirical Data

The second step in testing the computational model is to fit the model to the empirical data (i.e., data generated from observation of human participants). The goal of model fitting is to provide a quantitative test of model. This is done by estimating the values of the model parameters based on the data and examining the correspondence between the data and the predictions of the model when simulated using those parameter values. Model fitting allows for a more convincing demonstration of sufficiency, because it can be used to examine not only whether the model can reproduce the presence of particular effect (which we refer to as *qualitative* sufficiency), but also how accurately the model captures the magnitude of the effect (*quantitative* sufficiency). Another advantage of model fitting is that it allows the researcher to quantify the extent to which the predictions of the model correspond to the empirical data. This is particularly important for model comparison, which we discuss in the next section.

Similar to most statistical models, estimating the parameters of a computational model usually requires the help of a computer algorithm. There are several software platforms available that provide general-purpose optimization algorithms that have been used for estimating model parameters. Examples of platforms that have been used in organizational sciences include Vensim (e.g.,

Vancouver et al., 2010), MATLAB (e.g., Ballard et al., 2016), and R (e.g., Gee et al., 2018). General-purpose optimization is useful for simpler models with only a few estimated parameters. However, many will fail to produce reliable parameter estimates when applied to relatively complex, dynamic, and non-linear models like the model described above (for example, they may arrive at different parameter estimates depending on their starting values).

Fitting our model to the data therefore requires a more robust method of parameter estimation. For this reason, we use a Bayesian parameter estimation approach. Bayesian parameter estimation involves combining information about the researcher's a priori beliefs regarding each parameter, referred to as its *prior distribution* (henceforth, prior), with information about the *likelihood* of the model given the data. This analysis results in a *posterior distribution* (henceforth, posterior) on each parameter, which represents the range of parameter values that are credible given the researcher's prior and the empirical data. A highly dispersed posterior, which covers a broad range of possible values, suggests a high degree of uncertainty regarding the true parameter value. A narrow posterior, which covers only a small range of values, suggests more certainty. Details regarding the model likelihood function can be found in Appendix A. Details regarding the priors on the model parameters can be found in Appendix B.

Implementation

Although Bayesian parameter estimation is conceptually straightforward, the implementation of this method is often complex. For most models, including the types of models likely to be of interest in organizational sciences, there is no closed-form analytic solution for calculating the posterior. The posterior must therefore be approximated using Markov chain Monte Carlo (MCMC) methods. These methods "learn" the posterior distribution over time by producing sequences, or chains, of sample parameter values that are generated by an algorithm that is known to approximate the posterior given a sufficient number of samples (see Kruschke, 2015; Van Ravenzwaaij, Cassey, & Brown, 2018).

There are several open-source platforms for implementing Bayesian models using MCMC methods. These platforms include WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), JAGS (Plummer, 2003), OpenBUGS (Thomas, O'Hara, Ligges, & Sturtz, 2006), and Stan (Carpenter et al., 2017). The advantage of programs such as these is that they are sufficiently flexible to enable the implementation of most computational models, with MCMC algorithm being implemented automatically by the program. This allows the user to focus on the development of the model itself. This is similar to programs such as MPlus (Muthen, 1997), where the user writes the model code, and from that code the program automatically implements the algorithm necessary to estimate the parameters. In this paper, we use the Stan platform. We use Stan because its MCMC algorithm, the No-U-Turn Sampler (Hoffman & Gelman, 2014), is particularly well suited for implementing more complex models like the ones required to represent dynamic, multilevel phenomena (Stan Development Team, 2017).

To run the model fitting routine, the user includes a call to the Stan model within their code. At the time of writing, Stan can be run from R, Stata, MATLAB, Python, Julia, or the command line. For these analyses, we ran Stan via the RStan package (Stan Development Team, 2016), which provides an R interface to Stan. The R and Stan code required for implementing the model can be downloaded from https://osf.io/4euhr/. Information regarding how to interpret the RStan output can

be found in Appendix C.

Evaluating the model fit

Once the model has been fit, the researcher can evaluate whether the model provides a satisfactory account of the data. One simple way to examine the correspondence between the model and the data is to plot the predictions of the fitted model against the data (Heathcote, Brown, & Wagenmakers, 2015). This approach is particularly useful for ruling out poorly performing models, because misalignment between the model and the data are often very obvious.

Because the Bayesian approach assumes the existence of uncertainty in the model parameters, there must also be uncertainty associated with the predictions of the fitted model. Thus, the predictions of the fitted model are delivered in the form of a distribution, which is commonly referred to as the *posterior predictive* distribution. This distribution represents the values of each model output that are credible given that uncertainty in the model parameters. Figure 5 shows the results from the experiment described above superimposed over the posterior predictive distributions from the computational model. The red line represent the observed means. The blue line represents the means of the posterior predictive distributions and the light blue ribbon represents the 95% credible interval (CI) of the posterior predictive distribution. The CI spans the 2.5% and 97.5% quantiles of the distribution, and provides an indication of the uncertainty around the model predictions.

As can be seen, there is a modest decrease in effort over the course of each goal striving episode. The observed effort trajectories are mostly within the CI of the model predictions. However, the posterior predictives appear to underestimate the reduction in perceived effort that occurs within some goal striving episodes. Consistent with the notion that skill increases as a function of practice, the rate at which the performance variable increased during each goal striving episode improved over the ten episodes. Goal level also increased steadily across episodes. The observed performance and goal level trajectories were generally within the CI of the model predictions.

The above analysis suggests there is general correspondence between the fitted model predictions and the empirical observations. This finding suggests that the assumptions contained in the model, when simulated using the parameters estimated by the algorithm, are sufficient to account for the magnitude of the changes in the observed self-regulatory variables. In the next section, we continue to evaluate the computational model by directly comparing this model to an alternative.

Step 3: Model Comparisons

Although the evaluations conducted in Steps 1 and 2 demonstrate that the model indeed accounts for the phenomena it was developed to explain, demonstrations of sufficiency provide only weak evidence for a model (Farrell & Lewandowsky, 2018). This is because they do not provide any information regarding the *necessity* of the assumptions that form the model. Based on the approaches used in Steps 1 and 2, it is impossible to rule out whether the empirical phenomena could be accounted for by a simpler alternative model. As such, it is important that any empirical test of a computational model involve comparing the model to alternatives (e.g., see Ballard et al., 2016). By demonstrating that the hypothesized model not only accounts for the empirical effects but also provides a better description of the data than the alternative model(s), the researcher can provide more compelling evidence in favor of their model.

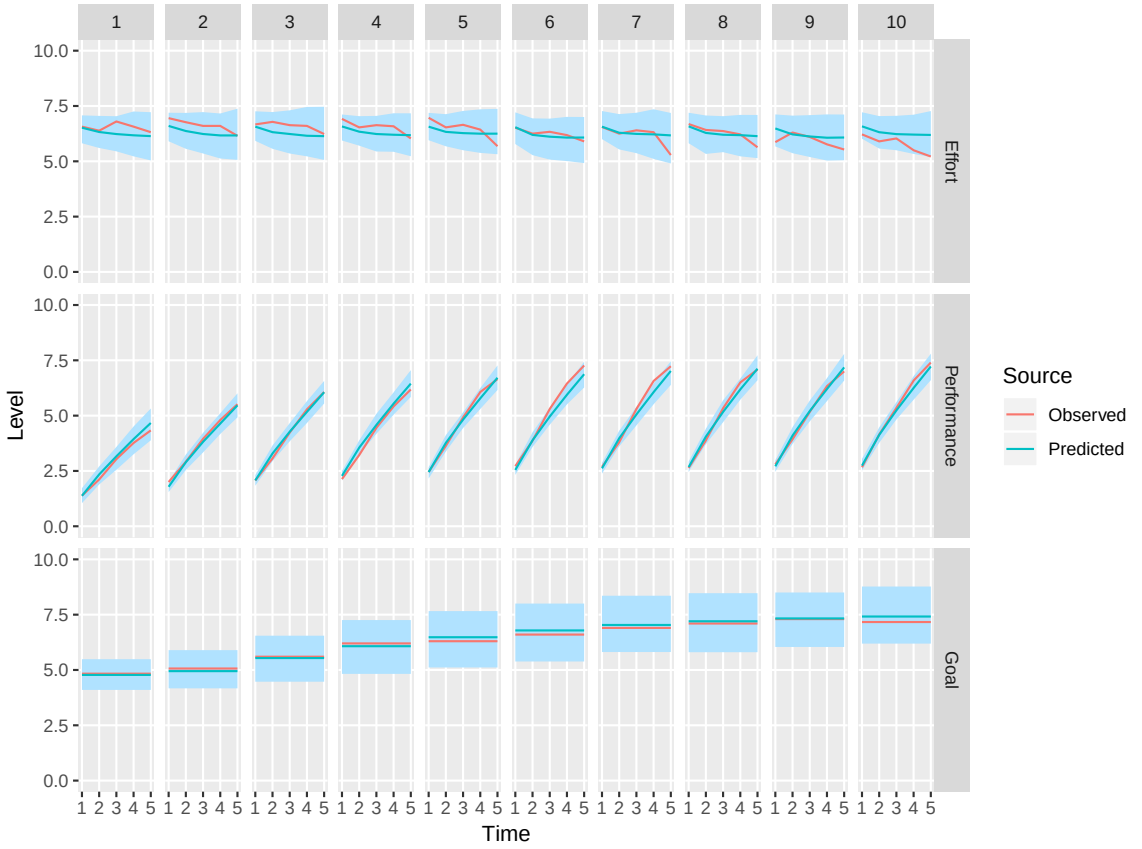

Figure 5. Observed empirical trends and posterior predictive distributions from the computational model. The lines represent the observed means. The dark blue line and light blue ribbon indicate the mean and 95% credible interval of the posterior predictive distribution.

We compare the model described above, which we now refer to as the *hypothesized model*, to an alternative model in which each the four sub-processes represented in the hypothesized model–effort regulation, skill acquisition, performance accrual, and goal revision–are described using linear functions. Table 2 summarizes the changes that were made to the hypothesized model to create the alternative model. Support for the alternative model would suggest that the non-linear dynamics included in the hypothesized model (e.g., the exponential model of learning and the logistic model of performance) may not be necessary to account for the self-regulatory processes that were at play in the experiment described above. Support for the hypothesized model would suggest that these processes are likely more complex than what can be represented using a linear systems model, and that the non-linear dynamics might be necessary to account for these processes.

When conducting model comparisons, the researcher will typically repeat Steps 1 and 2 with each alternative model. The process of simulating and fitting the alternative model(s) can help assess the sufficiency of the alternative explanation(s) for accounting for the phenomenon. However, these steps are usually not sufficient for discriminating between the alternatives. Differences in model-data fit may be too subtle to be picked up visually or may be washed out when averaged over multiple data points. Moreover, evaluation based on visual fit alone ignores other key considera-

Table 2
Summary of alternative model

Sub-process	Equation(s) replaced	Equation in alternative model
Effort Regulation	Equations 1 & 2	$Effort_{i,j} = b_0 + b_1 \times GPD_{i-1,j}$
Skill Acquisition	Equation 3	$Skill_{i,j} = b_0 + b_1 \times Practice_{i,j}$
Performance Accrual	Equation 5	$\Delta Perf = b_0 + b_1 \times Effort_{i,j} + b_2 \times Skill_{i,j}$
Goal Revision	Equations 6 & 7	$Goal_j = b_0 + b_1 \times Perf_{j-1}$

tions such as model complexity and generality (Myung & Pitt, 1997; Vandekerckhove, Matzke, & Wagenmakers, 2015). A more complex model has more flexibility, allowing it account for a wider range of observations. Generality refers to the extent to which the model applies across different settings. These two considerations are related. Models that are more complex may end up capturing sampling error in one dataset that will not generalize to others. This can reduce the ability of the model to predict new observations (this problem is referred to as *overfitting*). There is therefore a need to balance model-data fit with complexity, in order to ensure that the predictions of the model generalize.

Bayesian approaches to model comparison have some useful features when it comes to quantifying model complexity. Standard approaches to model comparison usually define complexity based on the number of estimated parameters, with more parameters being taken to mean more complexity. This is true for commonly used indices such as the AIC (Akaike, 1973), BIC, (Schwarz, 1978), and RMSEA (Steiger, 1990), which all include terms that penalize models with higher numbers of parameters. However, the complexity of a model is determined by more than just the number of parameters. One often unrecognized factor is the functional form of the model (Lee & Wagenmakers, 2013; Myung & Pitt, 1997). Models with the same number of parameters can still differ in complexity if the functional form of one model allows it more flexibility than its competitor[3]. Functional form complexity is important to take into account when assessing our computational model of self-regulation, because the functional forms of the various sub-processes (e.g., the exponential model skill acquisition and the logistic model of performance) are theoretically meaningful.

Another often overlooked factor that influences model complexity is the range of possible values that the model parameters can take on (i.e., the size of the parameter space). For example, consider two models that share the same functional form and estimated parameters, but one model requires that a certain parameter only take on positive values, whereas the other model allows that parameter to take on any value. In this case, the latter model is more flexible and can therefore account for a broader range of observations (Myung & Pitt, 1997). In our computational model, the allowable range of parameters is meaningful, because certain parameters theoretically must fall within certain ranges (e.g., the rate at which effort returns to baseline, α, is theoretically bounded between 0 and 1; the maximum achievable skill level, $Skill_{max}$, must be greater than the initial skill level, $Skill_0$). As such, it is important that we take into account this form of complexity as well.

Bayesian approaches to model comparison take into account all three of these dimensions

[3]For example, Steven's psychophyiscal power law of subjective stimulus intensity ($\Psi(x) = k \times x^a$) is more complex than Fechner's logorithmic law ($\Psi(x) = k \times ln(x + a)$), even though both laws have two parameters (k and a). The reason Steven's law is more complex is that its functional form allows for subjective stimulus intensity ($\Psi(x)$) to be a negatively or positively accelerating function of objective intensity (x), whereas under Fechner's law the function can only be negatively accelerating (Myung & Pitt, 1997).

of model complexity. They do so because each parameter in a Bayesian model must have a prior distribution associated with it. The existence of the prior "levels the playing field" so to speak because more complex models will have priors that are spread over more possible combinations of parameter values. This means that, in a more complex model, each possible combination of parameter values will tend to have lower prior probability. By contrast, simpler models have prior distributions that are more concentrated, resulting in combinations of possible parameter values that tend to have higher prior probabilities. This decrease in average prior probability as models become more complex trades off against any increase in fit that is achieved. As a result, empirical observations that are highly consistent with the predictions of a simpler model will tend to provide more evidence for the simpler model, because those observations will correspond to parameter values that have higher prior probabilities. This is true even if the empirical observations can also be accounted for by a more complex model. By contrast, empirical observations that are less consistent with the simpler model will tend to provide evidence in favor of the more complex model.

The standard solution for Bayesian model comparison is the Bayes factor (Jeffreys, 1935; Kass & Raftery, 1995). The Bayes factor refers to the ratio of the probabilities of the data under each model, and provides an index of the relative evidence delivered by the data for one model against an alternative that takes into account both fit and model complexity (Kruschke & Liddell, 2018). There are some challenges associated with the use of the Bayes factor however (Vandekerckhove et al., 2015). The Bayes factor can be difficult to obtain when comparing complex models. Calculating the Bayes factor requires the likelihood of the data under the model to be integrated across the entire parameter space, which can be computationally cumbersome for models with many parameters. This is especially true for models that need to be estimated using MCMC methods. In recent years, methods have been introduced for approximating Bayes factors for models estimated via MCMC methods (e.g., Evans, Steyvers, & Brown, 2018; Gronau et al., 2017; L. Wang & Meng, 2016). However, these methods are very computationally demanding.

Another issue associated with use of the Bayes factor is prior sensitivity. The Bayes factor is strongly influenced by the researcher's choice of priors (Rouder, Speckman, Sun, Morey, & Iverson, 2009). This means that if the priors are not well-informed (e.g., by previous empirical data), the resulting Bayes factor will be difficult to interpret. This is less problematic for standard statistical tests, where considerable thought has gone into the recommended "default" priors (e.g., Wagenmakers et al., 2018). However, when working with models that are novel and/or more complex, the researcher may be less confident in their choice of priors because a default specification will likely not exist. In these cases, Bayes factors are often difficult to interpret.

Because the model we have presented is both novel and relatively complex, we opt for another commonly-used approach to model comparison: leave one out cross validation (Geisser & Eddy, 1979). Leave one out cross validation is a method for estimating the predictive accuracy of a model by examining the model's ability to predict new data. The method involves fitting the model to all but one observation, and then examining then model's ability to reproduce the observation that had been left out. This process is extremely cumbersome. Fortunately, there are indicies available that approximate the results of the cross validation. One such index, that can be easily computed using the RStan package is the leave one out information criterion (LOO-IC). The LOO-IC is interpreted such that a lower value indicates a better trade off between fit and complexity. Our analysis revealed that the LOO-IC for the hypothesized model (estimate = 22396.9, SE = 222.2) was lower for than the

LOO-IC for the alternative model (estimate = 24529.2, SE = 148.6)[4]. This result suggests that the hypothesized model indeed strikes a better balance between fit and complexity than the alternative, linear model.

It is important to emphasize that any quantitative index should be interpreted in the context of theoretical plausibility (Jacobs & Grainger, 1994; Myung & Pitt, 1997). Plausibility refers to the degree to which the assumptions represented in the model are theoretically justifiable. Researchers should strive to ensure that the model's assumptions are consistent with previous findings and to avoid structures in the model that are not necessary to explain the phenomenon (Murphy & Russell, 2017). No quantitative index will be able to measure theoretical plausibility. The onus is on the researcher to determine whether a model is an accurate approximation of the process the model purports to represent. These considerations should be carefully weighed against the ability of the model to parsimoniously account for the data.

Step 4: Interpreting the Parameter Estimates

Once the researcher has established that the model accurately captures the empirical results and provides a better explanation of the data than competing models, they will likely wish to interpret the model's parameters. For example, the researcher may wish to determine the extent to which changes in goal-performance discrepancy during goal striving lead to subsequent changes in effort. In this case, they would interpret the β parameter. However, there are different ways that the model can be parameterized, and the choices that are made regarding the parameterization of the model impact the interpretations that can be made. One option is to estimate parameters at the sample-level, such that every participant is described by a single set of parameters. The researcher may also wish to model differences in the parameters between specific groups of participants (e.g., conditions of an experimental manipulation) in order to ascertain whether parameters differ across groups. We begin this section by interpreting the parameters from the hypothesized model in the analyses above, in which the parameters were estimated at the sample level. We then show how this model can be extended to capture differences between experimental conditions.

Sample-level model

The simplest way to parameterize a computational model is to assume that every participant in the sample can be described by the same set of parameters. Such a model would assume, for example, that every participant responds to changes in goal-performance discrepancy with the same adjustment in effort, acquires skill at the exact same rate, and so on. We refer to models parameterized in this way as *sample-level* models, because the parameters are estimated and interpreted at the level of the sample. A sample-level model produces a posterior distribution for each estimated parameter which describes the most probable values of the relevant parameter. To illustrate, the top row of Figure 6 shows the posterior distributions for the β, δ, and λ parameters (see Appendix C for more details regarding the parameter estimates). Parameter values for which the posterior density is higher are more probable. As can be seen, the most probable values of β lie within the range of

[4]Another commonly used information criterion for comparing Bayesian models is the Watanabe-Akaike Information Criterion (also known as the "Widely applicable information criterion"; Watanabe, 2010). The WAIC is an approximation to the LOO-IC, so these two indicies will generally yield similar results. In this analysis, the WAIC produced nearly identical results to the LOO-IC.

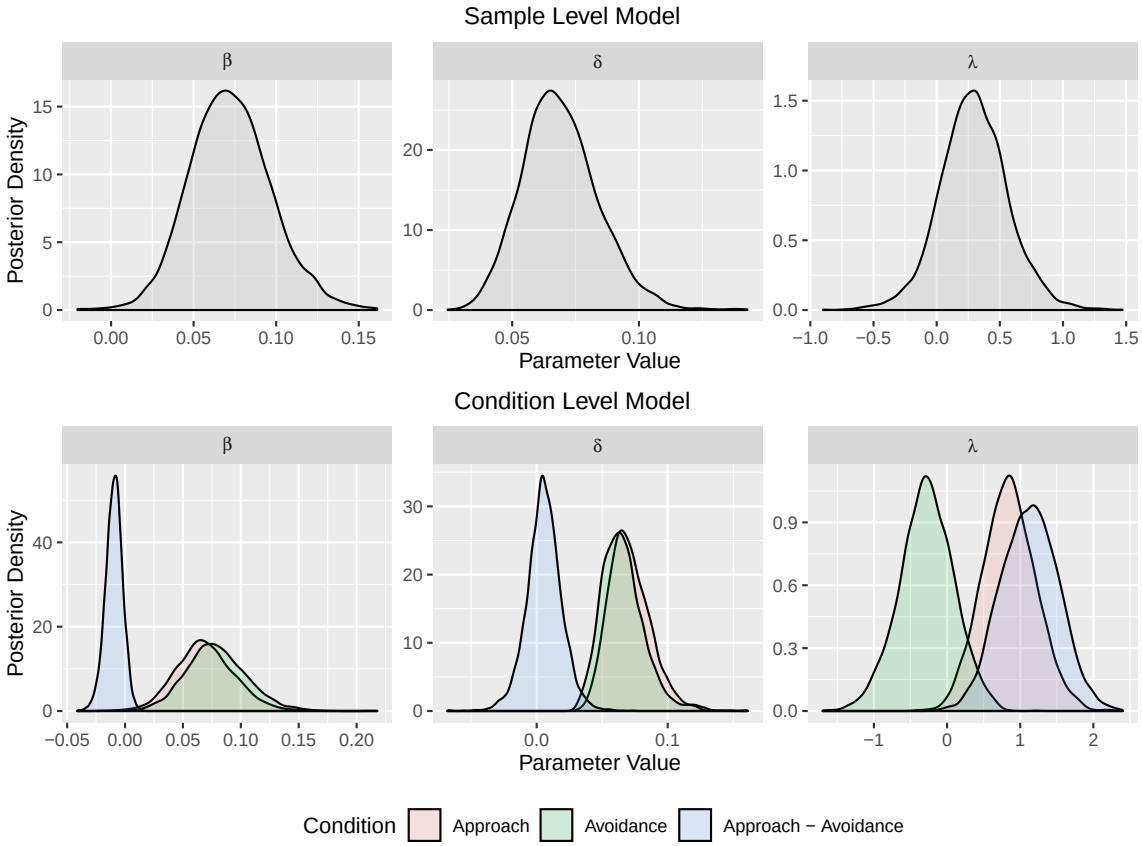

Figure 6. Posterior distributions on the β, δ, and λ parameters. The top and bottom rows display the parameter estimates from the sample-level and condition-level model respectively.

about 0 to 0.15. The most probable values of δ lie within the range of about 0.04 to 0.12. Finally, the most probable values of λ lie within the range of about -0.5 to 1.0.

One commonly used way of summarizing the most probable parameter values is by calculating the 95% credible interval for each parameter. In the above model, the CI on β is 0.026 to 0.124, the CI on δ is 0.042 to 0.102, and the CI on λ is -0.242 to 0.840. The 95% *highest density interval* (HDI; also referred to as the *highest posterior density* or HPD interval) is another commonly used interval for summarizing the posterior distribution. The 95% HDI represents the shortest possible interval containing 95% of the posterior density, and will therefore always include the most probable parameter values (Kruschke, Aguinis, & Joo, 2012). In the above model, the HDI on β is 0.029 to 0.125, the HDI on δ is 0.040 to 0.098, and the HDI on λ is -0.241 to 0.843. The CI and HDI each has advantages and disadvantages (e.g., see Kruschke, 2015), but for symmetrical distributions (e.g., β and λ) the two intervals will be very similar.

The CI or HDI is often used to decide whether particular parameter values should be accepted or rejected. For example, intervals that exclude zero are often taken as evidence that the parameter value is different from zero (Kruschke et al., 2012). This approach can be extended to any parameter value the researcher wishes to consider. For example, one could use this method to examine whether a weight parameter (i.e., a multiplier) is different from 1 or whether a probability parameter is

different from 0.5. Another simple way to quantify the evidence for a parameter being either greater than or less than a particular value is to determine the percentage of the posterior density that is above or below that value (Kruschke, 2015). For example, 99.8% of the posterior distribution on β is greater than 0. This, combined with the fact that the CI and HDI both exclude 0, would suggest that, on average, people indeed respond to reductions in goal-performance discrepancy by withdrawing effort. Contrast this result with the λ parameter, for which both the CI and HDI include 0, and for which only 87.8% of the posterior distribution is greater than 0. On the basis of these findings, we would not reject 0 as a plausible value for the λ parameter. This means that we cannot conclude that participants have an overall bias that leads them to set either risk-seeking or risk-averse goals. Although interpreting the posterior distributions as we have here is informative, it should be noted that some have advocated against making inferences based only on a parameter's posterior distribution and instead have advocated for inferences about parameters to be based on nested model comparisons (e.g., Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016).

Condition-level model

Although the sample-level model described above is useful for understanding the behavior of a single group of participants, many research questions require the consideration of multiple groups. For example, in the Gee et al. (2018) study, goal framing was manipulated between participants. For half of the participants, the goal was framed in approach terms. For these participants, the goal was expressed as the minimum number of correct decisions the participant needed to achieve. For the other half, the goal was framed in avoidance terms. In this case, the goal was expressed as the maximum number of incorrect decisions the participant was allowed to make. One of Gee et al.'s aims was to examine whether the goal revision process differs between approach and avoidance contexts.

Answering this type of question requires a model in which unique parameters are estimated separately for each experimental condition, which we refer to as a *condition-level* model. Fortunately, the sample-level model described above can easily be extended to estimate separate parameters for each condition. The bottom row of Figure 6 shows the approximate posterior distributions from a model in which β, δ, and λ were estimated at the condition level.

As can be seen, the posterior distributions on β for the approach group (illustrated by the pink distribution) and the avoidance group (the green distribution) are almost completely overlapping. The same is true for the δ parameter. The fact that the approach and avoidance distributions overlap suggests that there are virtually no differences between approach and avoidance groups in terms of the degree to which people respond to reductions in goal-performance discrepancy by decreasing effort (β) and the rate at which skill is acquired (δ). However, there is some separation between the posteriors on λ between the approach and avoidance groups, suggesting that risk preference may not be the same across these conditions.

To make inferences about differences in parameter values between conditions, we need to examine the posterior distribution on the *difference* between parameters in the approach and avoidance conditions. This is done by calculating, for each posterior sample, a variable that is equal to the sampled parameter value for the approach condition minus the sampled value for the avoidance condition. This yields a difference score for each posterior sample, which can be used to generate a posterior distribution of the difference score. The posterior on the difference scores for each param-

eter is illustrated by the blue distribution in Figure 6. As can be seen, the posteriors on the difference scores for the β and δ parameters indicate that 0 is a plausible value for each of these parameters. For the β difference score, the CI is -0.025 to 0.004, the HDI is -0.025 to 0.004, and the percentage of the distribution greater than 0 is 9.2 (i.e., 90.8% of the distribution falls below 0). For the δ difference score, the CI is -0.021 to 0.033, the HDI is -0.020 to 0.034, and the percentage greater than 0 is 68.8. These results indicate that it would be plausible to assume that β and δ are the same across approach and avoidance groups. By contrast, the posterior on the λ difference score suggests that there is a difference in this parameter between the two groups (CI: [0.373, 1.878]; HDI: [0.371, 1.878], % > 0: 99.9). This finding is consistent with Gee et al. (2018)'s results and suggests that people have a higher preference for setting risky goals (i.e., goals higher than one's expected performance) when pursuing approach goals compared to when pursuing avoidance goals. This analysis illustrates how the posterior distribution can be used to compare components of a dynamic process across groups of participants.

It is important to note that there are a range of more complex parameter structures that can be implemented to help answer more specific research questions. For example, researchers wishing to examine individual differences in process components can do so via a hierarchical model. A hierarchical model simultaneously captures variation in parameters between participants, but also allows for inferences to be made about the group(s) as a whole (Kruschke, 2015; Turner et al., 2013; Vincent, 2016). Alternatively, some researchers may wish to identify subgroups of participants who behave similarly, without any priors knowledge of group membership. This can be achieved using a mixture model (Bartlema, Lee, Wetzels, & Vanpaemel, 2014).

Discussion

Dynamic theories have become increasingly influential in the organisational sciences. However, such theories can be difficult to test, because they often describe complex, multilevel processes that operate reciprocally over time, with different sub-processes operating at different levels of analysis and exerting both top-down and bottom-up effects. We propose that computational modeling offers a more complete approach to testing theories, because it provides a direct integration between the theory itself, the model that is used to operationalize the theory, and the empirical data that is used to test the theory.

In this paper, we demonstrated how computational modeling can be used to test this type of theory. We began by translating a conceptual theory into a computational model. The model described the mechanisms by which effort regulation exerts bottom-up effects on the goal revision process. We then demonstrated four steps involved in testing the computational model. The first step is to simulate the model to examine whether the model can reproduce the phenomenon that they theory purports to explain and to generate predictions that can be tested empirically. The second step is to fit the model to the empirical data such as that generated by an experiment or observational study. The goal of this step is to estimate the model's parameter values and to quantify the degree to which the predictions of the model correspond to the data. The third step is to compare the model to alternatives in order to rule out competing explanations. The fourth step is to interpret the model's parameter values in order to quantify particular components of the process. In the next sections, we discuss the applicability of this approach to other areas of organizational sciences, and then address associated challenges.

A General Approach

The approach we advocate in this paper is a general one. It is not specific to the research question we investigated or the level of analysis at which the processes we examined operate. Our example explored how the effort regulation process that operates over time within an individual goal striving episode influences the higher level goal striving process that operates across a series of goal striving episodes. It may also, however, be the case that these processes influence, and are influenced by, processes that operate at the group or team level. For example, the performances of individual group members likely contribute to the formation of group norms (a bottom-up effect), which may in turn influence goals that are set by other group members (a top-down effect; Flynn & Amanatullah, 2010; Morgenroth, Ryan, & Peters, 2015). The approach we described above could be applied to examine such processes by extending the computational model so that it includes multiple group members each pursuing their own goals. The model might test different mechanisms by which the performances of individual members lead to the emergence of a shared performance norm. It also might examine different explanations for how group norms influence the goals of the individual members.

This approach can also be applied to help answer macro-level research questions. For example, one might use this approach to test McKinley et al. (2014)'s theory of how organisations respond to episodes of decline. Their theory describes a feedback process that operates at the organization level. According to their theory, the ways in which organizations pursue innovation influence the trajectory of the decline. Organisational decline impacts innovation flexibility and management rigidity, and these factors feed back and determine whether the organisational decline turns around or enters a downward spiral. This type of theory is highly amenable to refinement and testing using computational modeling because it explicitly specifies the patterns of change over time as well as the causes and consequences of these patterns.

Another example of a theoretical framework for which the approach demonstrated in this paper might be useful is Helfat and Peteraf (2003)'s model of the capability lifecycle. This model describes the organization-level process through which capabilities evolve over time. According to the model, capabilities develop over three stages–a founding stage, a development stage, and a maturity stage–with the level of capability increasing rapidly over the first two stages and then leveling off in the third stage. The model predicts several ways in which firms respond to external factors that influence the trajectory of capability development. For example, threats to capability such as a decrease in demand might lead a firm to redeploy the capability in a new market. As a first step, a computational model could be used to simulate different capability threats and examine whether a simulated firm responds in the manner that is predicted by the theory.

Future work might also use the approach demonstrated in this paper to test multilevel theories that describe so-called "micro to macro" processes, in which the actions and interactions of individual employees have emergent effects at the organization level. Such processes are relevant to research on "mircofoundations", which focuses on tracing organizational-level phenomena back to individual-level antecedents (Barney & Felin, 2013; Felin, Foss, & Ployhart, 2015). This research has become increasingly influential in the management, strategy, and organization theory literatures. For example, Nickerson and Zenger (2008) presented a theory that described how envy resulting from social comparisons between managers influences the structure of the organization and its activities. This theory therefore proposes a bottom-up process by which a dyad- or group-level

phenomenon (manager social comparison) gives rise to an organization-level phenomenon (organization structure and activities). Another example is the model introduced by Helfat and Peteraf (2015), which explains how the cognitive ability of individual managers (an individual-level phenomenon) can impact the performance of the firm as a whole (an organization-level phenomenon).

More generally, the organizational sciences have paid increasing attention to bottom-up processes where lower-level entities exert influence on the context in which they are embedded (e.g., Kozlowski et al., 2013; Kozlowski, Chao, Grand, Braun, & Kuljanin, 2016). Although bottom-up processes are often assumed by theory, they are challenging to examine quantitatively because it is difficult to represent this sort of process using conventional multilevel models. As a result, most of the work on bottom-up processes has been qualitative. Recently, computational modeling has been increasingly used as a tool to generate predictions regarding bottom-up processes (e.g., Grand et al., 2016). However, it is rarely the case that these models are tested against empirical data in the manner we have shown. The approach we have demonstrated offers a way to more directly test multilevel theories that describe bottom-up processes. This approach may open up avenues for future research to answer questions that cannot be answered using traditional multilevel methods. Such questions might involve performing quantitative model comparisons to examine which model most accurately characterizes a bottom-up process. For example, one might test different models to examine whether team performance norms more strongly dictated by the maximum performance achieved by any member, the minimum, or some representation of the prototypical performance. Other questions might rely on parameter estimation to quantify components of a bottom-up process. For example, one might use this approach to examine how rapidly team norms form, and how the rate of norm formation is influenced by factors such as managerial involvement in goal setting or the degree of competition between team members.

Challenges

Despite the many opportunities we believe that this approach offers, there are some challenges that should be considered. First, readers more accustomed to programs with graphical user interfaces such as SPSS, AMOS, or JASP may find that it takes time to become familiar with the Stan syntax. On the continuum from high level platform in which a handful of built-in analyses can be executed by button press (e.g., SPSS) to lower level programming languages that require the user to code the entire analysis from scratch (e.g., C++), Stan lies somewhere in the middle. Stan has all the functionality required to implement the MCMC algorithm that generates the parameter estimates built in to the program, but it requires the user to build the model from the ground up. This process is more involved than simply specifying a set of variables that are assumed to be related. However, it is the control over the model that makes the approach so flexible. This control gives the researcher the ability to translate a set of theoretical assumptions directly into a computational model that can be fit to data, enabling a rigorous test of the theory to be conducted.

A second challenge is the computational demands associated with this approach. Although the Bayesian approach has the useful feature of being able to quantify uncertainty in model parameters, as opposed to simply delivering a point estimate, this advantage comes at a cost. The MCMC algorithm that underlies this approach can require a long time to run. For example, the analyses presented in this paper ran for just under an hour. For models that are more complex, have more parameters, or that are being applied to larger datasets, it is not uncommon for the analysis to take

several hours to run. This time frame can be challenging for analyses that require comparing a large number of alternative models. Fortunately, Stan has several features that help speed up the process, such as built-in parallelization. Additionally, much of this process can be automated.

A third challenge that is important to consider concerns the necessity of the prior distribution. The ability to incorporate prior information into the analysis is a useful feature of the Bayesian framework. However, the prior must be carefully considered, because the researcher's choice of prior influences the results of the analysis (see Vanpaemel, 2010). Priors that are too narrowly concentrated on a particular region of the parameter space may mean that other parameter values more compatible with the data are overlooked in the analysis. Assuming that priors are reasonable, the effect of the prior on the posterior parameter estimates can be overcome by analyzing more data. However, this is not the case for many of the methods used for model comparison (e.g., the Bayes factor). Because the prior in part determines the model's flexibility, it has a pronounced impact on the degree to which the model is punished for its complexity (Vandekerckhove et al., 2015). As a result, the choice of prior can influence the results of the model comparison. For these reasons, it is important that priors are realistic, and where possible, informed by theory and previous research.

A fourth challenge is the problem of measurement. The need for valid measurement of constructs is just as important when analyzing data using computational models as it is when conducting traditional empirical research. One notable difference between the two approaches, however, is that computational models do not require a measure for every construct in the model. For example, we could implement a version of the model above in which effort is treated as an unobserved construct. Treating variables as unobserved can be a useful option for modeling constructs that are difficult to measure. The downside, however, is that this reduces the number of data points that inform the parameter estimates. Having fewer data will generally decrease the precision of the parameter estimates and reduce the ability to differentiate between competing models. Thus, there may be a trade-off between the desire to incorporate as much information as possible into the analysis and the desire to avoid less valid information such as inadequately measured constructs.

Conclusion

For cumulative theoretical progress to be made, theories must be able to be subjected to rigorous tests using models that provide an accurate representation of the process described by the theory. This makes testing dynamic, multilevel theories difficult, because the processes that these theories describe are often difficult to represent using standard multilevel models. We have presented an approach to testing this type of theory that involves simulating computational models, fitting them to empirical data, comparing alternative models, and interpreting parameter estimates. We hope that this approach helps promote a shift from thinking at the level of the variables to thinking at the level of the model as a whole. We believe such an approach will ultimately accelerate theoretical progress by enhancing our ability to develop, test, reject, and refine dynamic, multilevel theories.

References

Abell, P., Felin, T., & Foss, N. (2008). Building Micro-foundations for the Performance Links. *Managerial and Decision Economics*, *92*(6), 489–502. doi: 10.1002/mde

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado. doi: 10.1007/978-1-4612-1694-0_15

Ballard, T., Yeo, G., Loft, S., Vancouver, J. B., & Neal, A. (2016). An integrative formal model of motivation and decision making: The MGPM. *Journal of Applied Psychology*, *101*, 1240–1265. doi: 10.1037/apl0000121

Ballard, T., Yeo, G., Vancouver, J. B., & Neal, A. (2017). The dynamics of avoidance goal regulation. *Motivation and Emotion*, *41*, 1–10. doi: 10.1007/s11031-017-9640-8

Barney, J., & Felin, T. (2013). What are microfoundations? *Academy of Management Perspectives*, *27*(2), 138–155. doi: 10.5465/amp.2012.0107

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150. doi: 10.1016/j.jmp.2013.12.002

Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annual Review of Psychology*, *40*, 109–131.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 435–455.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal Of Statistical Software*, *76*, 1–32. Retrieved from http://mc-stan.org/users/documentation/

Carver, C. S., & Scheier, M. F. (1998). *On the self-regulation of behavior*. New York, NY: Cambridge University Press.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, *57*, 505–528. doi: 10.1146/annurev.psych.57.102904.190146

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007, apr). Developing theory through simulation methods. *Academy of Management Review*, *32*, 480–499. Retrieved from http://amr.aom.org/cgi/doi/10.5465/AMR.2007.24351453 doi: 10.5465/AMR.2007.24351453

Donovan, J. J., & Williams, K. J. (2003). Missing the mark: Effects of time and causal attributions on goal revision in response to goal-performance discrepancies. *Journal of Applied Psychology*, *88*(3), 379–390. doi: 10.1037/0021-9010.88.3.379

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*, 41–60. doi: 10.1002/(SICI)1099-0526(199905/06)4:5<41::AID-CPLX9>3.3.CO;2-6

Evans, N. J., Steyvers, M., & Brown, S. D. (2018). Modeling the Covariance Structure of Complex Datasets Using Cognitive Models: An Application to Individual Differences and the Heritability of Cognitive Ability. *Cognitive Science*, 1–20. Retrieved from http://psiexp.ss.uci.edu/research/papers/modelling-covariance-structure.pdf{\%}0Ahttp://doi.wiley.com/10.1111/cogs.12627 doi: 10.1111/cogs.12627

Farrell, S., & Lewandowsky, S. (2018). *Computational Modeling of Cognition and Behavior*. Cambridge: Cambridge University Press.

Feldman, M. S., & Pentland, B. T. (2003). Reconceptualizing Organizational Routines as a Source of Flexibility and Change. *Administrative Science Quarterly*, *48*, 94–118. Retrieved from https://www.jstor.org/stable/3556620?origin=crossref doi: 10.2307/3556620

Felin, T., Foss, N. J., & Ployhart, R. E. (2015). The microfoundations movement in strategy and organization theory. *Academy of Management Annals*, *9*(1), 575–632. doi: 10.1080/19416520.2015.1007651

Flynn, F. J., & Amanatullah, E. T. (2010). Psyched up or psyched out? The influence of coactor status on individual performance. *Organization Science*, *23*(2), 402–415. doi: 10.1287/orsc.1100.0552

Fum, D., Del Missier, F., & Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research*, *8*, 135–142. doi: 10.1016/j.cogsys.2007.07.001

Galor, O. (2007). *Discrete Dynamical Systems*. Berlin: Springer.

Gee, P., Neal, A., & Vancouver, B. (2018). A formal model of goal revision in approach and avoidance

contexts. *Organizational Behavior and Human Decision Processes*, *146*, 51–61. doi: 10.1016/j.obhdp.2018.03.002

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*, 153–160. doi: 10.1080/01621459.1979.10481632

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511. doi: 10.1214/ss/1177013604

Grand, J. A. (2017). An examination of stereotype threat effects on knowledge acquisition in an exploratory learning paradigm. *Journal of Applied Psychology*, *102*, 115–150.

Grand, J. A., Braun, M. T., Kuljanin, G., Kozlowski, S. W. J., & Chao, G. T. (2016). The dynamics of team cognition: A process-oriented theory of knowledge emergence in teams. *Journal of Applied Psychology*, *101*, 1353–1385.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., … Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. doi: 10.1016/j.jmp.2017.09.005

Harrison, J. R., Lin, Z., Carroll, G. R., & Carley, K. M. (2007). Simulation modeling in organizational and management research. *Academy of Management Review*, *32*, 1229–1245.

Hatch, M. J. (1993). The dynamics of organizational culture. *Academy of Management Review*, *18*, 657–693. doi: 10.5465/AMR.1993.9402210154

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185–207.

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An Introduction to good practices in cognitive modeling. In B. U. Forstmann & E. J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 25–48). New York, US: Springer.

Helfat, C. E., & Peteraf, M. A. (2003). The dynamic resource-based view: Capability lifecycles. *Strategic Management Journal*, *24*, 997–1010. doi: 10.1002/smj.332

Helfat, C. E., & Peteraf, M. A. (2015). Managerial cognitive capabilities and the microfoundations of dynamic capabilities. *Strategic Management Journal*, *36*, 831–850. doi: 10.1002/smj

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A thoery of the relationship between individual differences and information processing. *Psychological Review*, *91*, 153–184.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. Special Section: Modeling visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(6), 1311–1334.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophy Society*, *31*, 203–222.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude treatment interaction approach to skill acquisition. *Journal of Applied Psychology*, *74*, 657–690. doi: 10.1037//0021-9010.74.4.657

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kennedy, D. M., & McComb, S. A. (2014). When teams shift among processes: Insights from simulation and optimization. *Journal of Applied Psychology*, *99*(5), 784–815. doi: 10.1037/a0037339

Kozlowski, S. W. J., Chao, G. T., Grand, J. A., Braun, M. T., & Kuljanin, G. (2013). Advancing multilevel research design: Capturing the dynamics of emergence. *Organizational Research Methods*, *16*, 581–615. doi: 10.1177/1094428113493119

Kozlowski, S. W. J., Chao, G. T., Grand, J. A., Braun, M. T., & Kuljanin, G. (2016). *Capturing the multilevel dynamics of emergence: Computational modeling, simulation, and virtual experimentation* (Vol. 6) (No. 1). doi: 10.1177/2041386614547955

Kruglanski, A. W., Belanger, J. J., Chen, X., Catalina, K., Pierro, A., & Manetti, L. (2012). The energetics of motivated cognition: A force-field analysis. *Psychological Review*, *119*, 1–20.

Kruschke, J. K. (2015). *Doing bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). San Diego, CA: Academic Press.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*, 722–752. doi: 10.1177/1094428112457829

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*, 155–177. doi: 10.3758/s13423-017-1272-1

Latham, G. P., & Baldes, J. J. (1975). The "practical significance" of Locke's theory of goal setting. *Journal of Applied Psychology*, *60*, 122–124. doi: 10.1037/h0076354

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. New York, NY: Cambridge University Press.

Leibowitz, N., Baum, B., Enden, G., & Karniel, A. (2010). The exponential learning equation as a function of successful trials results in sigmoid performance. *Journal of Mathematical Psychology*, *54*(3), 338–340. doi: 10.1016/j.jmp.2010.01.006

Li, W. D., Fay, D., Frese, M., Harms, P. D., & Gao, X. Y. (2014). Reciprocal relationship between proactive personality and work characteristics: A latent change score approach. *Journal of Applied Psychology*, *99*(5), 948–965. doi: 10.1037/a0036169

Lin, Z. J., Yang, H., & Demirkan, I. (2007). The performance consequences of ambidexterity in strategic alliance formations: Empirical investigation and computational theorizing. *Management Science*, *53*(10), 1645–1658. doi: 10.1287/mnsc.1070.0712

Lord, R. G., Diefendorff, J. M., Schmidt, A. M., & Hall, R. J. (2010). Self-regulation at work. *Annual Review of Psychology*, *61*, 543–568. doi: 10.1146/annurev.psych.093008.100314

Luenberger, D. G. (1979). *Introduction to Dynamic Systems: Theory, Models, and Applications*. New York: Wiley.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Marks, M. A., & Mathieu, J. E. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*, 356–376.

McKinley, W., Latham, S., & Braun, M. (2014). Organizational decline and innovation: Turnarounds and downward spirals. *Academy of Management Review*, *39*, 88–110. doi: 10.3390/met7020058

Morgenroth, T., Ryan, M. K., & Peters, K. (2015). The motivational theory of role modeling: How role models influence role aspirants' goals. *Review of General Psychology*, *19*(4), 465–483. doi: 10.1037/gpr0000059

Murphy, K. R., & Russell, C. J. (2017). Mend It or End It: Redirecting the Search for Interactions in the Organizational Sciences. *Organizational Research Methods*, *20*(4), 549–573. doi: 10.1177/1094428115625322

Muthen, B. (1997). Latent variable modeling of longitudinal and multilevel data. *Sociological Methodology*, *27*, 453–480. doi: 10.1111/1467-9531.271034

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Neal, A., Ballard, T., & Vancouver, J. B. (2017). Dynamic self-regulation and multiple-goal pursuit. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 410–423. doi: https://doi.org/10.1146/annurev-orgpsych-032516-113156

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Nickerson, J. A., & Zenger, T. R. (2008). Envy, comparison costs, and the economic theory of the firm. *Strategic Management Journal*, *29*, 1429–1449. doi: 10.1002/smj

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*, 44–64. doi: 10.1016/0010-0285(75)90004-3

Pentland, B. T., Feldman, M. S., Becker, M. C., & Liu, P. (2012). Dynamics of organizational routines: A generative model. *Journal of Management Studies*, *49*, 1484–1508. doi: 10.1111/j.1467-6486.2012

.01064.x

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leische, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing (dsc 2003).* Technische Universität Wien, Vienna, Austria. doi: 10.1.1.13.3406

Powers, W. T. (1973). *Behavior: The control of perception*. Chicago, IL: Aldine.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E. J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*, 520–547. doi: 10.1111/tops.12214

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. doi: 10.3758/PBR.16.2.225

Rudolph, J. W., Morrison, J. B., & Carroll, J. S. (2009). The dynamics of action-oriented problem solving: Linking interpretation and choice. *Academy of Management Review*, *34*, 733–756. doi: 10.5465/AMR.2009.44886170

Salas, E., Kozlowski, S. W. J., & Chen, G. (2017). A century of progress in industrial and organizational psychology: Discoveries and the next century. *Journal of Applied Psychology*, *in press*. Retrieved from `http://dx.doi.org/10.1037/apl0000206` doi: 10.1037/apl0000206

Schmidt, A. M., & DeShon, R. P. (2007). What to do? The effects of discrepancies, incentives, and time on dynamic goal prioritization. *Journal of Applied Psychology*, *92*, 928–941. doi: 10.1037/0021-9010.92.4.928

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Stan Development Team. (2016). *RStan: the R interface to Stan, Version 2.10.1.* Retrieved from `http://mc-stan.org/`

Stan Development Team. (2017). *Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0.* Retrieved from `http://mc-stan.org/users/documentation`

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180. doi: 10.1207/s15327906mbr2502_4

Tarakci, M., Greer, L. L., & Groenen, P. J. F. (2016). When does power disparity help or hurt group performance? *Journal of Applied Psychology*, *101*, 415–429. doi: 10.1037/apl0000056

Team, R. C. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS Open. *R News*, *6*, 12–17.

Tsang, P. S., & Wilson, G. F. (1997). Mental workload measurement and analysis. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2nd ed., pp. 417–449). New York, NY: Wiley.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206. doi: 10.1016/j.neuroimage.2013.01.048

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*, *25*, 143–154. doi: 10.3758/s13423-016-1015-8

Vancouver, J. B. (2005). The depth of history and explanation as benefit and bane for psychological control theories. *Journal of Applied Psychology*, *90*, 38–52. doi: 10.1037/0021-9010.90.1.38

Vancouver, J. B., & Purl, J. D. (2017). A computational model of self-efficacy's various effects on performance: Moving the debate forward. *Journal of Applied Psychology*, *102*, 599–616.

Vancouver, J. B., Wang, M., & Li, X. (2018). Translating informal theories into formal theories: The case of the dynamic computational model of the integrated model of work motivation. *Organizational Research Methods*, 1–37. doi: 10.1177/1094428118780308

Vancouver, J. B., & Weinhardt, J. M. (2012). Modeling the mind and the milieu: Computational modeling for micro-level organizational researchers. *Organizational Research Methods*, *15*, 602–623. doi: 10.1177/1094428112449655

Vancouver, J. B., Weinhardt, J. M., & Schmidt, A. M. (2010). A formal, computational theory of multiple-

goal pursuit: Integrating goal-choice and goal-striving processes. *Journal of Applied Psychology*, *95*, 985–1008. doi: 10.1037/a0020628

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. Busemeyer et al. (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–317). Oxford, UK: Oxford University Press.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498. doi: 10.1016/j.jmp.2010.07.003

Vincent, B. T. (2016). Hierarchical Bayesian estimation and hypothesis testing for delay discounting tasks. *Behavior Research Methods*, *48*, 1608–1620. doi: 10.3758/s13428-015-0672-2

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76. doi: 10.3758/s13423-017-1323-7

Wang, L., & Meng, X.-L. (2016). Warp bridge sampling: The next generation. *arXiv preprint*. Retrieved from `https://arxiv.org/pdf/1609.07690.pdf`

Wang, M., Zhou, L., & Zhang, Z. (2016). Dynamic modeling. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 241–266. doi: 10.1146/annurev-orgpsych-041015-062553

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Weinhardt, J. M., & Vancouver, J. B. (2012). Computational models and organizational psychology: Opportunities abound. *Organizational Psychology Review*, *2*, 267–292. doi: 10.1177/2041386612450455

Woodard, D. B. (2007). *Detecting poor convergence of posterior samplers due to multimodality.* Discussion Paper 2008-05, Duke University, Department of Statistical Science, Durham, N. C. Retrieved from `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.122.8344`

Wright, R. A. (2008). Refining the prediction of effort: Brehm ' s distinction between potential motivation and motivation intensity. *Social and Personality Psychology Compass*, *2*, 682–701.

Appendix A: Likelihood Function for Hypothesized Model

In order to fit the computational model described in the paper to the data, a likelihood function needs to be specified. The likelihood function describes the probability of the data having been observed given the parameter values being considered. This function is necessary because it enables the algorithm to determine how probable different parameter values are given the data. In our analysis, the likelihood is informed by three variables in the data: perceived effort, performance, and goal level. As is common among statistical models of time series data, we specify separate functions for the first observation in the time series and the change in the observed variable across subsequent observations. For goal level, the first observation in the time series is the first goal striving episode. The likelihood function for these observations is defined as follows:

$$Goal_{j=1} \sim \ Normal(Goal_0, \sigma_{Goal_0}). \tag{8}$$

According to the statement above, the observed goal level for episode 1 is normally distributed with a mean equal to the $Goal_0$ parameter and with a standard deviation parameter denoted σ_{Goal_0}. The likelihood function describing changes in goal level over subsequent episodes is specified as follows:

$$\Delta Goal_j \sim \ Normal(\Delta\widehat{Goal}_j, \sigma_{\Delta Goal}), \tag{9}$$

for episodes where $j > 1$. According to the statement above, the observed changes in goal level across subsequent episodes are normally distributed with a mean equal to the change in goal level that is predicted by the model, denoted $\Delta\widehat{Goal}_j$, and with a standard deviation parameter denoted $\sigma_{\Delta Goal}$.

For the perceived effort variable, the first observation in the time series occurs after the first time window of each goal striving episode (i.e., time 1). The likelihood function for these observations is defined as follows:

$$Effort_{i=1,j} \sim \ Normal(Effort_0, \sigma_{Effort_0}). \tag{10}$$

According to the statement above, the perceived effort observed for the first observation of goal striving episode *j* is normally distributed with a mean equal to the $Effort_0$ parameter and with a standard deviation parameter denoted σ_{Effort_0}. The likelihood function describing the changes in effort over subsequent time windows in each goal striving episode is specified as follows:

$$\Delta Effort_{i,j} \sim \ Normal(\Delta\widehat{Effort}_{i,j}, \sigma_{\Delta Effort}), \tag{11}$$

for observations where $i > 1$. According to the statement above, the observed changes in perceived effort across observations after time 1 in each goal striving episode are normally distributed with a mean equal to the change in effort that is predicted by the model, denoted $\Delta\widehat{Effort}_{i,j}$, and with a standard deviation parameter denoted $\sigma_{\Delta Effort}$.

The initial level of the performance variable is constrained by the task to equal 0. Because of this, the likelihood of the performance variable can be captured using a single function that describes the change in performance across time windows with each goal striving episode:

$$\Delta Perf_{i,j} \sim \ Normal(\Delta\widehat{Perf}_{i,j}, \sigma_{\Delta Perf}). \tag{12}$$

According to the statement above, the observed changes in performance across successive observations in each goal striving episode are normally distributed with a mean equal to the change in performance that is predicted by the model, denoted $\Delta\widehat{Perf}_{i,j}$, and with a standard deviation parameter denoted $\sigma_{\Delta Perf}$.

Appendix B: Priors

Each parameter that is estimated from the data, including the standard deviation parameters in the likelihood functions, requires its own prior distribution. The priors used for the estimated parameters of the hypothesized model are shown in Table B1. For the parameters that have both lower and upper bounds (i.e., $Effort_{bl}$, α, $Effort_0$, δ, θ, and $Goal_0$), we specify uniform distributions as priors. For the parameters that can take on any value, we specify normally distributed prior distributions. For the parameters that can only take on only positive values but have no upper bound (e.g., the standard deviation parameters), we specify zero-truncated normal prior distributions.

As can be seen, all parameters that have normal or truncated normal priors have distributions with means of 0 and standard deviations of 2, except γ. The γ parameter represents the point on the

Table B1
Priors for the Estimated Parameters

Parameter	Distribution Family	Mean	SD	Lower Bound	Upper Bound
$Effort_{bl}$	Uniform			0	10
α	Uniform			0	1
β	Normal	0	2	None	None
$Effort_0$	Uniform			0	10
$Skill_0$	Truncated Normal	0	2	0	None
$\Delta Skill$	Truncated Normal	0	2	0	None
δ	Uniform			0	1
κ	Normal	0	2	None	None
γ	Normal	5	2	None	None
θ	Uniform			0	1
λ	Normal	0	2	None	None
$Goal_0$	Uniform			0	10
σ_{Effort_0}	Truncated Normal	0	2	0	None
$\sigma_{\Delta Effort}$	Truncated Normal	0	2	0	None
$\sigma_{\Delta Perf}$	Truncated Normal	0	2	0	None
σ_{Goal_0}	Truncated Normal	0	2	0	None
$\sigma_{\Delta Goal}$	Truncated Normal	0	2	0	None

effort rating scale at which half of the maximum performance change is achieved. We set the mean of the prior on γ to 5 because this is the midpoint of the effort rating scale (which ranges from 0 to 10). The γ prior therefore represents the assumption that people can achieve half of the maximum performance change with a moderate level of effort.

Note that we do not directly estimate the maximum achievable skill level. Instead, we re-parameterized the model so that it estimates a parameter that represents the difference between the initial skill level ($Skill_0$) and the maximum achievable skill level ($Skill_{max}$). This difference parameter is denoted $\Delta Skill$ and represents the extent to which skill level changes from its initial to its maximum value. $Skill_{max}$ is calculated by adding $Skill_0$ and $\Delta Skill$ for each sample. The reason for estimating $\Delta Skill$ instead of $Skill_{max}$ is that, by constraining $\Delta Skill$ to be positive, we ensure that $Skill_{max}$ is always greater than $Skill_0$. This constraint would be much less efficient to implement if we were to estimate $Skill_{max}$ directly.

Appendix C: Summary of RStan Output

Stan returns an object that contains samples from the posterior distribution for each parameter in the model. The RStan package provides a summary of each posterior, which is shown below for the hypothesized model. As can be seen in the example output below, the summary contains the posterior mean, the standard error of that mean, and the standard deviation of the posterior. It also displays the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the distribution.

```
1                mean se_mean    sd   2.5%   25%   50%   75% 97.5%    n_eff  Rhat
2 effort_bl     6.355   0.003 0.143  6.083 6.261 6.356 6.447 6.646 2029.048 1.002
3 alpha         0.378   0.002 0.098  0.208 0.309 0.369 0.440 0.593 2158.462 1.001
4 beta          0.072   0.001 0.025  0.026 0.055 0.072 0.088 0.124 2095.059 1.002
```

```
                          mean se_mean   sd    2.5%     25%     50%     75%   97.5%    n_eff  Rhat
effort_0                 6.582   0.002 0.103   6.379   6.513   6.581   6.651   6.781 2620.726 1.001
skill_0                  3.578   0.015 0.729   2.350   3.060   3.497   4.034   5.216 2319.383 1.002
skill_change             4.185   0.019 0.869   2.688   3.562   4.098   4.715   6.143 2137.001 1.001
delta                    0.069   0.000 0.015   0.042   0.059   0.068   0.078   0.102 4761.500 1.000
kappa                    3.408   0.014 0.712   2.249   2.897   3.322   3.826   5.050 2681.321 1.001
gamma                    6.748   0.004 0.152   6.471   6.641   6.741   6.849   7.070 1725.576 1.003
theta                    0.625   0.003 0.221   0.168   0.469   0.639   0.802   0.980 4012.861 1.001
lambda                   0.294   0.005 0.270  -0.242   0.124   0.292   0.466   0.840 2791.598 1.000
goal_0                   4.835   0.005 0.270   4.292   4.651   4.834   5.013   5.363 3203.900 1.000
sigma_effort_0           2.448   0.001 0.069   2.318   2.400   2.447   2.495   2.589 5921.674 1.000
sigma_effort_change      1.689   0.000 0.024   1.640   1.673   1.689   1.705   1.738 6138.292 0.999
sigma_performance_change 0.928   0.000 0.012   0.904   0.920   0.928   0.936   0.951 5539.911 0.999
sigma_goal_0             2.119   0.003 0.192   1.781   1.986   2.103   2.237   2.547 5667.437 1.000
sigma_goal_change        1.471   0.001 0.046   1.387   1.441   1.471   1.501   1.564 5662.381 0.999
```

The output also provides some information to help assess model convergence. The number of effective samples (`n_eff`) is a measure of the information contained in the posterior samples that accounts for the autocorrelation between neighboring samples in each MCMC chain. MCMC chains tend to be autocorrelated because each sample in part depends on the previous sample. The number of effective samples represents the number of independent samples that would produce an equivalent amount of information to the set of samples obtained. If there is so little autocorrelation that samples are effectively independent, the number of effective samples will be approximately equal to the total number of samples (4000 in this example, which is the RStan default). The final piece of information provided is the potential scale reduction factor (`Rhat`; Brooks & Gelman, 1998; Gelman & Rubin, 1992). This gives an indication of the extent to which the chains converged on the same region of the parameter space. As the chains approach convergence, `Rhat` approaches one. A common rule of thumb is that `Rhat` should be less than 1.1 before any inferences are made based on the posteriors (Kruschke et al., 2012). Note that this convergence diagnostic is only an approximation, and is not without limitations. For example, it has been criticized for failing to detect non-convergence in certain types of multi-model distributions (Woodard, 2007).