# Approaches to Analysis in Model-based Cognitive Neuroscience

Brandon M. Turner[a,*], Birte U. Forstmann[b], Bradley C. Love[c], Thomas J. Palmeri[d], Leendert Van Maanen[b]

[a]*Department of Psychology, The Ohio State University*
[b]*Department of Psychology, University of Amsterdam*
[c]*Department of Psychology, University College, London*
[d]*Department of Psychology, Vanderbilt University*

**Abstract**

Our understanding of cognition has been advanced by two traditionally non-overlapping and non-interacting groups. Mathematical psychologists rely on behavioral data to evaluate formal models of cognition, whereas cognitive neuroscientists rely on statistical models to understand patterns of neural activity, often without any attempt to make a connection to the mechanism supporting the computation. Both approaches suffer from critical limitations as a direct result of their focus on data at one level of analysis (cf. Marr, 1982), and these limitations have inspired researchers to attempt to combine both neural and behavioral measures in a cross-level integrative fashion. The importance of solving this problem has spawned several entirely new theoretical and statistical frameworks developed by both mathematical psychologists and cognitive neuroscientists. However, with each new approach comes a particular set of limitations and benefits. In this article, we survey and characterize several approaches for linking brain and behavioral data. We organize these approaches on the basis of particular cognitive modeling goals: (1) using the neural data to constrain a behavioral model, (2) using the behavioral model to predict neural data, and (3) fitting both neural and behavioral data simultaneously. Within each goal, we highlight a few particularly success-

*Corresponding Author
*Email address:* `turner.826@gmail.com` (Brandon M. Turner )
This work was supported by grants NIH R01-EY021833, NSF Temporal Dynamics of Learning Center SMA-1041755, and NIH P30-EY08126. The second through fifth authors are simply listed alphabetically.

ful approaches for accomplishing that goal, and discuss some applications. Finally, we provide a conceptual guide to choosing among various analytic approaches in performing model-based cognitive neuroscience.

*Keywords:* model-based cognitive neuroscience, linking, analysis methods

---

## 1. Introduction

Our understanding of cognition has been advanced by two nearly non-overlapping and non-interacting groups. The first group, mathematical psychologists, is strongly motived by theoretical accounts of cognitive processes, and instantiates these theories by developing formal models of cognition. The models often assume a system of computations and mathematical equations intended to characterize a process that might actually take place in the brain. To formally test their theory, mathematical psychologists rely on their model's ability to fit behavioral data. A good fit is thought to reflect an accurate theory, whereas a bad fit would refute it (Roberts and Pashler, 2000). The second group, cognitive neuroscientists, rely on statistical models to understand patterns of neural activity, often without any attempt to make a connection to the computations that might underlie some hypothesized mechanism. For example, some statistical approaches (e.g., multivariate pattern analysis) explicitly condition on the neural data to determine which aspects of the data produce better predictions for behavioral outcomes. Such an analysis can tell us *which* brain regions are predictive of a particular behavior and even *by how much*, but they say nothing about neither *how* nor *why* particular brain regions produce said behavior.

Although both groups are concerned with explaining behavior, they tend to approach the challenge from different vantage points. Thinking in terms of Marr (1982)'s levels of analysis, mathematical psychologists tend to focus on the computational and algorithmic levels, whereas cognitive neuroscientists focus more on the implementation level. Although progress can be made by maintaining a tight focus, certain opportunities are missed. As a result of their single-level focus, both approaches suffer from critical limitations (Love, 2015). Without a cognitive model to guide the inferential process, cognitive neuroscientists are often (1) unable to interpret their results from a mechanistic point of view, (2) unable to address many phenomena when restricted to contrast analyses, and (3) unable to bring together results from different paradigms in a common theoretical framework. On the other hand, the cognitive models developed by mathematical psychologists are inherently abstract, and the importance of physiology and brain function is often unappreciated. After fitting a model to data, mathematical psychologists can describe an individual's behavior, but they can say nothing about the behavior's neural basis. More importantly, neural data can provide information that can help distinguish between competing cognitive models that cannot be uniquely identified based on fits to behavioral data alone (Ditterich, 2010; Mack et al., 2013; Purcell et al., 2012).

The many limitations of single-level analyses have inspired researchers to combine neural and behavioral measures in an integrative fashion. The importance of solving the integration problem has spawned several entirely new statistical modeling approaches developed through collaborations between mathematical psychologists and cognitive neuroscientists, collectively forming a new field often referred to as model-based cognitive neuroscience (e.g., Forstmann et al., 2011; van Maanen et al., 2011; Turner et al., 2013b; Mack et al., 2013; Palmeri, 2014; Boehm et al., 2014; Love, 2015; Palmeri et al., 2015; Turner et al., 2015b). We refer to these as "approaches", because they are general strategies for integrating neural and behavioral measures via cognitive models, and are neither restricted to any particular kind of neural or behavioral measure, nor any particular cognitive model. However, with each new approach comes a unique set of limitations and benefits. The approaches that have emerged in the recent years fill an entire spectrum of information flow between neural and behavioral levels of analysis, and deciding between them can be difficult. Given the overwhelming demand for these integrative strategies, we believe that an article surveying the different types of analytic approaches could be an invaluable guide for any would-be model-based cognitive neuroscientist.

Here we survey and characterize the many approaches for linking brain and behavioral data. We organize these different approaches into three general categories: (1) using the neural data to constrain a behavioral model, (2) using the behavioral model to predict neural data, and (3) modeling both neural and behavioral data simultaneously. For each specific approach within each category, we highlight a few particularly successful examples, and discuss some applications. In an attempt to draw a detailed comparison between the approaches, we then organize each of the approaches according to a variety of factors: the number of processing steps, the commitment to a particular theory, the type of information flow, the difficulty of implementation, and the type of exploration. In short, we discuss the ways in which current approaches bind data at multiple levels of analysis, and speculate about how these methods can productively constrain theory. We close with a discussion about additional considerations in model-based cognitive neuroscience, and provide an outlook toward future development.

## 2. Specific Analytic Approaches

For ease of categorization and subsequent comparison, we will hypothetically assume the presence of neural data, denoted $N$, and behavioral data, denoted $B$, which may or may not have been collected simultaneously. The neural data $N$ could be neurophysiological recordings, functional magnetic resonance imaging (fMRI), electroencephalography (EEG), or other physiological measures. The behavioral data $B$ could be response probabilities, response times, confidence ratings, or other typical behavioral data collected in a cognitive experiment. Cognitive modelers are interested in characterizing the mechanisms – specified in mathematical and computational terms – that lead to the behavior $B$ observed in a given experimental condition. Commonly, this characterization is derived from fitting a cognitive model to behavioral data, interpreting the resulting parameter estimates, and comparing (qualitatively or quantitatively) the observed behavior and the behavior predicted by the model. Cognitive neuroscientists are interested in uncovering the neural mechanisms that lead to the behavior $B$ observed in a given experimental condition. Commonly, this process involves a statistical analysis of neural data with respect to observed behaviors and experimental manipulations. However, model-based cognitive neuroscientists are interested in integrating neurophysiological information $N$ and behavioral outcomes $B$ by way of a cognitive model. The central assumption of these analyses is that information obtained from either source of data ($N$ or $B$) can tell a similar story – albeit in different languages – about some aspect of cognition, and the integration of the these measures assimilates the differences in languages across data modalities.

As model-based cognitive neuroscientists, we have many choices in deciding which story we would like to tell, and these choices depend on our research goals. In practice, there seems to be at least three general categories of approaches in the emerging field of model-based cognitive neuroscience. These three categories are illustrated in the rows of Figure 1. The first set of approaches uses neural data as auxiliary information that guides or constrains a behavioral model. There are several ways in which the neural data can

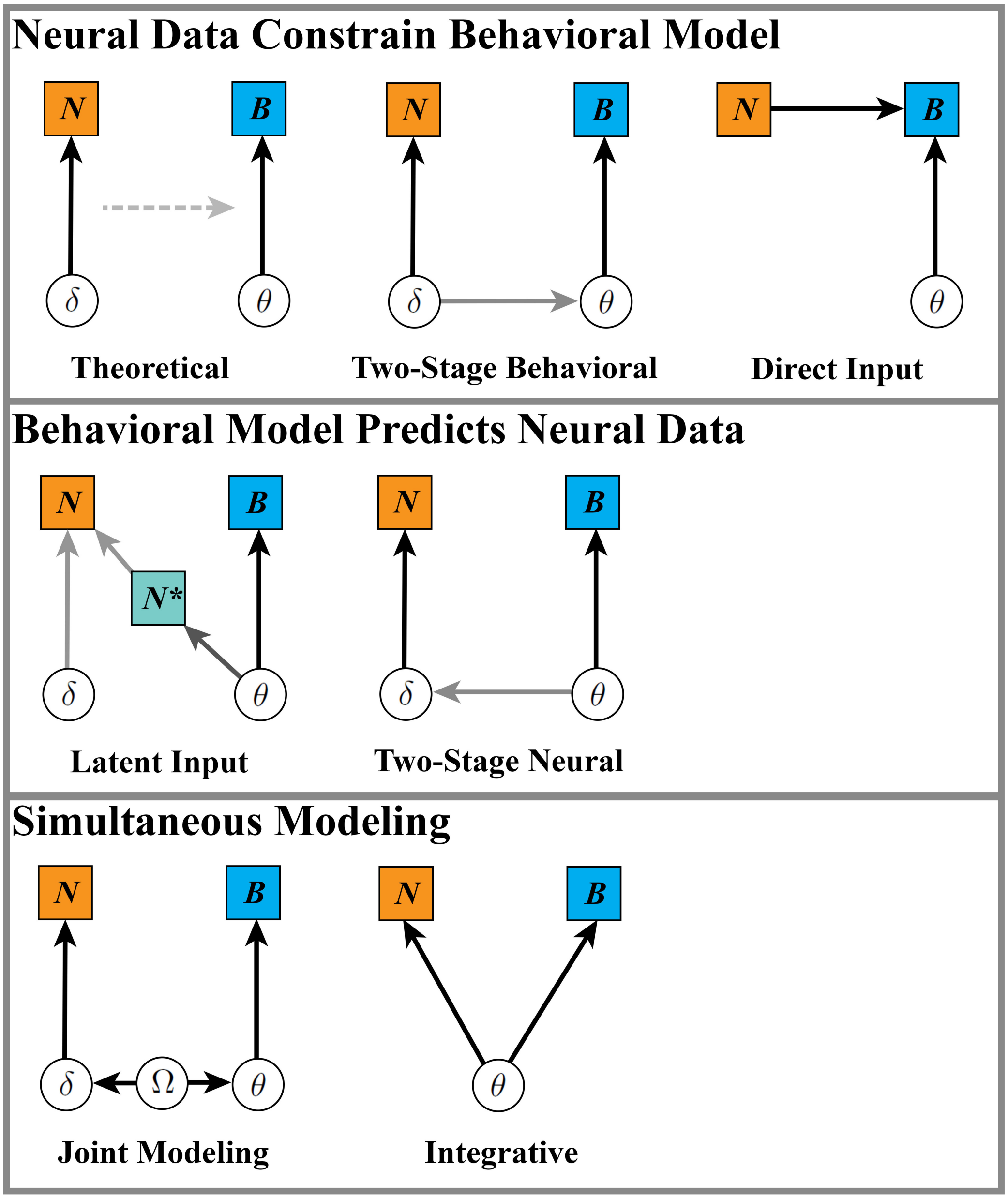


Figure 1: An illustration of several approaches used for linking neural and behavioral data, organized by specific modeling goals. $N$ represents the neural data, $B$ represents the behavioral data, $N^*$ represents simulated internal model states, and $\theta$, $\delta$, and $\Omega$ represent model parameters. When an approach is procedural, progression through processing stages is represented by arrows of decreasing darkness (e.g., the Latent Input Approach). Dashed lines indicate conceptual constraints (e.g., the Theoretical Approach), whereas solid lines indicate statistical constraints.

constrain modeling choices, and we will discuss three such approaches in the subsequent sections. The second set of approaches uses a behavioral model as a way to interpret or predict neural data. Behavioral models assume a set of mechanisms that theoretically mimic a cognitive process of interest, making them an interesting way to impose theory in data analyses. Moreover, while competing cognitive models might predict the same or similar patterns of behavioral data $B$, they might differ considerably in what they predict about neural data $N$, creating a powerful approach to model selection. We are faced with many choices in using these model mechanisms to guide our search for the interesting neural signatures. In the sections that follow, we will discuss two such approaches for accomplishing this goal. The third set of approaches builds a single model that jointly accounts for the random variation present in both the neural and behavioral data. With the proper model in place, one can simultaneously achieve constraint on the behavioral model while retaining the ability to interpret the neural data. In the sections that follow, we will discuss two approaches designed to accomplish this goal. We do not necessarily think this is a comprehensive list; in fact, we suspect that there is room for further development, and possibly the creation of entirely new analytic approaches.

Figure 1 represents the specific approaches as graphical diagrams where observable measures (i.e., data) are depicted as shaded square nodes, latent model parameters are depicted as empty circles, and arrows depict dependencies. Two of these approaches (i.e., Two-stage and Latent Input) require several processing stages, and we have represented the dependency structure of these stages as increasingly lighter shades of gray. Most of these approaches require a transformation from the data space to a (latent) parameter space, and this transformation can be unimodal (i.e., concerning only behavior data $B$ or neural data $N$) or bimodal (i.e., concerning both $B$ and $N$ simultaneously). The parameters can define a mechanistic model, like those commonly used by cognitive modelers, or they can define a statistical model, like those commonly used by cognitive neuroscientists. When an unimodal transformation is required, we denote the parameters of the neural model which predict $N$ as $\delta$, and the parameters of the behavioral model which predict $B$ as $\theta$. The neural model parameters $\delta$ might be slopes or intercept terms from a general linear model, or something more sophisticated like those used in topographic latent source analysis (Gershman et al., 2011). The behavioral model parameters $\theta$ represent things like discriminability in the signal detection theory model (Green and Swets, 1966), or the drift rate in the "diffusion decision model"[2] (Ratcliff, 1978; Forstmann et al., 2015). When a bimodal transformation is required, we generically denote the parameters as $\theta$ (e.g., the Integrative Approach in the bottom-right panel of Figure 1). For example, in the ACT-R framework (Anderson, 2007), the set of parameters $\theta$ represents a sequence of module activations, and their values have bimodal effects in the prediction of both neural and behavioral measures. Some approaches in our set require a simulation process where the parameters are used to generate synthetic data, and we will denote these data with an asterisk (e.g., $N^*$ denotes predicted neural data in the Latent Input Approach). Other approaches assume a secondary projection from a set of several parameter spaces to a group-level parameter space, such as in hierarchical modeling. We denote these higher-level parameters as $\Omega$ (e.g., the Joint Modeling Approach in the bottom-left panel of Figure 1). As an example, the joint modeling framework (Turner et al., 2013b) uses a hierarchical (Bayesian) structure for bridging the connection between neural and behavioral measures. With these general assumptions and notation in place, we can discuss how these various approaches achieve their intended analytic goal.

### *2.1. Neural Data Constrain Behavioral Model*

We begin our discussion with approaches that constrain a behavioral model with neural data. In this endeavor, the neural data are considered important, but only in the sense that they inform the mechanisms in the behavioral model. We have identified three specific approaches (i.e., see Figure 1): the Theoretical Approach, the Two-stage Behavioral Approach, and the Direct Input Approach. We now discuss each of these in turn.

#### *2.1.1. Theoretical Approach*

In the Theoretical Approach, psychological theories are developed on the basis of considerations from both neuroscience and behavioral data. The top left panel of Figure 1 illustrates the Theoretical Approach as statistically independent models of the neural and behavioral data because the link between these measures is established only through the researcher themselves (i.e., represented by the dashed arrow). In this approach, the dominant

[2] In this article, we refer to this model as the "diffusion decision model" following Forstmann et al. (2015). This same model has been called other names such as the "the diffusion model", the "drift diffusion model", and the "Wiener diffusion model."

procedure uses neural measures to inspire the development of psychological models. First, the researcher observes particular aspects of brain function, such as information about the structure (e.g., individual neurons or densely connected brain regions) or function (e.g., dorsal and ventral pathways of visual stimulus processing) of the brain. Next, the researcher develops a model of behavior that, at its core, abides by these neural observations. With an initial model structure imposed by $N$, the researcher is now able to evaluate the relative merits of nested theoretical assumptions, and make incremental adjustments in the model to provide better fits to behavioral data $B$. Unlike other approaches discussed in this article, the Theoretical Approach may draw inspiration from physiological or anatomical observations, but there is no mathematical or statistical link between the neural data $N$ and either the model architecture or the model parameters that predict the behavioral data $B$.

Although the absence of an explicit link between neural and behavioral data may seem craven, the Theoretical Approach has proven to be a powerful framework for motivating psychological theory. Perhaps the most prominent example of a Theoretical Approach is the enormous class of neural network models. Neural network models have a long history, with one classic example being Rosenblatt's Perceptron machine (Rosenblatt, 1961). In the development of the Perceptron, Rosenblatt made choices in his model that reflected operations observed in individual neurons, such as that the firing of individual neurons should be discrete (motivated by the McCullogh-Pitts neuron; McCullogh and Pitts, 1943). Although these original neural network models were heavily criticized (Minsky and Papert, 1969), pioneering work allowing for continuous activations in neuron-like units (Grossberg, 1978; Anderson, 1977; Rumelhart, 1977; McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982) evolved neural network models into more complex and successful theoretical approaches such as the parallel distributed processing (PDP; McClelland and Rumelhart, 1986) models. Superficially, these models allow for the presence of individual nodes embedded within layers of a network, and these nodes are massively interconnected across layers, resembling neural structures in the brain. Through a process known as backpropagation, PDP models can be trained on behavioral data to learn important aspects of the decision rule, facilitating further systematic explorations of representation, learning, and selective influence (i.e., by a process referred to as "lesioning").

As another example, consider the Leaky Competing Accumulator (LCA;

Usher and McClelland, 2001) model. The LCA model was proposed as a neurally plausible model for choice response time in a $k$-alternative task. The model possesses mechanisms that extend other diffusion-type models (e.g., Ratcliff, 1978) by including leakage and competition by means of lateral inhibition. These additional mechanisms have proven effective in explaining how, for example, time sensitive stimulus information can give way to differences in individual subject performance. For example, Usher and McClelland (2001) and Tsetsos et al. (2011) have shown the effects of primacy and recency for some subjects in a time-varying stimulus information paradigm. In these multi-alternative choice experiments, one response option may receive the strongest "input" (e.g., the brightness level) for the first 500 ms, but then the stimuli transition such that a different response option receives the strongest input relative to the first. In both of these studies, different parameterizations of the LCA model were used to demonstrate how primacy effects could be appreciated by having a large value for lateral inhibition relative to the strength of the input (i.e., the drift rate), and recency effects could be captured through a large leakage term relative to the input (Usher and McClelland, 2001; Tsetsos et al., 2011).

As a specific example of how the neurosciences have guided the assumptions in the LCA model, it is well known that the firing rate of individual neurons can never be negative. However, these firing rates can be attenuated by way of inhibition – a process carried out by other neurons in the system. To instantiate these neuronal dynamics, the full LCA model enforces a constraint such that if the degree of evidence for any choice alternative becomes negative, the degree of evidence for that accumulator should be reset to zero (Usher and McClelland, 2001). The floor-on-activation constraint was later found to be critical in capturing patterns of individual differences in multi-alternative choice that could not be captured by other diffusion-type models (Tsetsos et al., 2011). It is worth noting that other neurological constraints allow the LCA model to provide a unique characterization of behavioral data that would not otherwise be realized; specifically, the role of lateral inhibition relative to leakage in the model plays an interesting role in characterizing subject-specific patterns in behavioral data (Bogacz et al., 2006; van Ravenzwaaij et al., 2012; Tsetsos et al., 2011; Gao et al., 2011; Bogacz et al., 2007; Purcell et al., 2012; Teodorescu and Usher, 2013; Tsetsos et al., 2012; Ossmy et al., 2013; Turner and Sederberg, 2014; Turner et al., 2015a).

Given the highly subjective nature of the neural constraints imposed on a behavioral model, it should not be surprising that a great deal of controversy surrounds some applications of the Theoretical Approach. While neural network modelers have undoubtedly derived inspiration from the brain in building their models, the mechanistic implementation of these inspirations is often interpreted as a strong commitment, which opens the gates for scrutiny about plausibility and falsifiability (Minsky and Papert, 1969; Massaro, 1988; Roberts and Pashler, 2000). Furthermore, in some cases these additional neural mechanisms do not provide any advantage in terms of quantitative fit statistics to behavioral data over their simpler counterparts (e.g., see Ratcliff and Smith (2004), but also see Teodorescu and Usher (2013) and Turner et al. (2015a) for a different perspective). In some cases, there are also concerns centered on the level of explanation that the model provides (cf. Marr, 1982). On the one hand, the study of individual neurons constitutes an exploration of Marr's implementation level of analysis (Broadbent, 1985; Kemp and Tenenbaum, 2008; Pinker and Prince, 1988; Smolensky, 1988). On the other, the development of a cognitive model involves meandering through the computational level – Marr's highest level of analysis (Shiffrin and Nobel, 1997). To what extent should the implementation level be reflected or imposed on the computational level (e.g., Love, 2015; Frank, 2015; Teller, 1984)? For example, if we believe that individual neurons have a floor on activation or are inherently "leaky" (i.e., meaning they lose information over time), should this restriction be imposed on the dynamics of racing accumulators in a cognitive model (Zandbelt et al., 2015)? These accumulators are intended to reflect the amount of sensory evidence for each alternative – evidence that is apparently observed in many brain areas (including the lateral intraparietal area, superior colliculus, frontal eye field, and dorsolateral prefrontal cortex; Horwitz and Newsome, 1999, 2001; Kim and Shadlen, 1999; Shadlen and Newsome, 2001, 1996; Purcell et al., 2010, 2012; Hanes and Schall, 1996; Hanks et al., 2015), and so it begs the question: Which – if any – levels of decision making models should reflect the function of individual neurons? If the accumulators are to reflect the behavior of individual neurons, how might this connection be formally established (Smith, 2010; Smith and McKenzie, 2011)? Questions like this have been considered by many other scientists (e.g., Marr, 1982; Broadbent, 1985; Love, 2015; Frank, 2015; Schall, 2004; Teller, 1984), and the next two sections discuss two different ideas about how this connection should be made.

#### 2.1.2. *Two-stage Behavioral Approach*

The first formal linking approach uses neurophysiology to replace *parameters* of a behavioral model. For example, consider a model that explains some neural data $N$ with parameters $\delta$, and behavioral data $B$ with parameters $\theta$. The neural parameters $\delta$ could be divided into a set of parameters characterizing a key neural signal $\delta_1$, and a set of nuisance parameters $\delta_2$ so that $\delta = \{\delta_1, \delta_2\}$. Now suppose the behavioral model parameters could be divided into a set of parameters that are reflective of the behavioral signal $\theta_1$, and a set of parameters $\theta_2$ that are not. The structure of the Two-stage Behavioral Approach is to simply replace the set of parameters $\theta_1$ with the parameters of the neural signal $\delta_1$. We refer to this approach as the "Two-stage Behavioral" approach because the connection involves two stages, and that *behavioral* model parameters are replaced by neural parameters. This approach makes a strong commitment to how the neural signal is represented in the abstract mechanisms assumed by the behavioral model, and as a result, it is a stronger instantiation of neurophysiology than the Theoretical Approach discussed above.

The Two-stage Behavioral Approach is nicely illustrated by the work of Wang and colleagues (Wong and Wang, 2006), who developed a spiking neural network model of perceptual decision making. This model aims to account for the same kinds of behaviors as the DDM and the LCA model, but is far less abstract, with thousands of simulated spiking neurons, dense patterns of excitatory and inhibitory connections, pools of neurons associated with a single response, and the dynamics of individual neurons defined by several differential equations. While the model has dozens of potentially free parameters, most of them are defined directly by neural data. For example, the time constants of integration of different inhibitory and excitatory receptor types are based directly on physiological measures. While low-level spiking neural network models of this sort capture well many of the details of neurons and neural circuits and provide reasonable first-order predictions of behavioral data, they are difficult to simulate and quantitative fits to behavioral data are simply impossible using even state-of-the-art computer hardware (see Umakantha et al., 2015). Indeed, as a result of this additional complexity, very few efforts have been devoted to systematically studying the model's predictions for choice response time data. However, a few approximations have been developed for fitting purposes, and these approximations behave similarly to popular models in cognitive science such as the LCA model

(Wong and Wang, 2006; Bogacz et al., 2006; Roxin and Ledberg, 2008).

*2.1.3. Direct Input Approach*

The Two-stage Behavioral Approach represents one way in which the neural data can guide the behavioral model through neural model parameters, but it is easy to imagine other approaches that are more direct. For example, rather than translating the neural data $N$ to the neural model parameters $\delta$, and then using $\delta$ to constrain the behavioral model parameters $\theta$, we could instead use the neural data to directly replace dynamics of the behavioral model. This alternative approach is nicely illustrated by the Vanderbilt group (e.g., Palmeri et al., 2015; Purcell et al., 2010, 2012). They examined perceptual decision making within the sequential sampling model architecture assumed by models like the DDM (DDM; Ratcliff, 1978), and the LCA model (Usher and McClelland, 2001), among others. They specifically tested the hypothesis that different types of neurons in the frontal eye field (FEF) carry out different computations specified in accumulator models, namely that visually-responsive neurons in FEF encode the drift rate driving the decision process and that movement-related neurons in FEF instantiate the accumulation process itself. To test this linking proposition most directly (cf. Teller, 1984; Schall, 2004), they replaced the parameterized mechanisms thought to be embodied by the visually-responsive neurons, namely the time for perceptual processing and the drift rate, with the neurophysiological data recorded from visually-responsive neurons. Rather than having abstract mathematical and computational components specified by free parameters drive the decision process, the neural data ($N$) drove the decision process directly. To do this, the neural data were used to directly replace components of the model that would otherwise have been latent, and would need to be estimated from behavioral data. The only remaining free parameters were those that defined the decision making architecture (i.e., race, feedforward, lateral, or gated accumulation), and that defined speed-accuracy tradeoffs (i.e., threshold of accumulation). When constrained by neural inputs, they observed that only some of the various decision making architectures could fit the full set of behavioral data (correct and error response time distributions and response probabilities). They were then able to distinguish further between models based on how well the predicted accumulator model dynamics matched the observed neural dynamics in movement-related neurons, the neurons they hypothesized to carry out an accumulation of evidence (see Latent Input Approach below).

Although the Direct Input Approach is commonly used to feed neural data into a cognitive model, one could potentially invert the direction of influence in Figure 1 to analyze the neural data as a function of some behavioral variable, such as accuracy (e.g., Eichele et al., 2008) or response time (e.g., Weissman et al., 2006; Hanes and Schall, 1996). Once the neural data have been sorted as a function of the levels of the behavioral outcome, one might analyze the distribution of neural data between these levels (Woodman et al., 2008). Such a procedure has been the dominant analytic approach in neuroscience since its inception, but is not model-based, and so we will not consider it here. However, the model-based analogue of this analysis would be to use the model's machinery to drive the analysis of neural data. We refer to this approach as the Latent Input Approach, and will discuss it further in the next section.

### *2.2. Behavioral Model Predicts Neural Data*

Another set of analytic approaches involves searching the brain for areas that support mechanisms assumed in the behavioral model. Such a procedure allows one to interpret neural data through mechanisms in the model, which can potentially be more informative than behavioral data alone. We consider two approaches for accomplishing this goal: the Latent Input and the Two-stage Neural Approaches.

#### *2.2.1. Latent Input Approach*

The goal of the Latent Input Approach is a converse of sorts to the Direct Input Approach. In the Direct Input Approach, the goal is to use the neural data $N$ to constrain model mechanisms and parameters $\theta$ that predict behavior. In the Latent Input Approach, the cognitive model is used to guide the inference of neural data $N$, or to make predictions about $N$. To perform an analysis within this approach, one typically carries out three stages, illustrated in the middle-left panel of Figure 1. First, the parameters of a cognitive model $\theta$ are estimated by fitting the model to behavioral data $B$ alone. Second, the resulting parameter estimates are used to generate predictions about neural data $N^*$, which typically represents some "internal state" of the cognitive model in terms of the neural measure. Third, one searches for correlates of the model's internal state $N^*$ with the observed neural data $N$.

One example of an Latent Input analysis using fMRI data would be a voxel-by-voxel application of the general linear model relating the model's internal state $N^*$ to the neural data $N$ (e.g., O'Doherty et al., 2007). The typical result is a pattern of voxels representing significant correlations with the cognitive model, and these voxels are taken as the region of the brain supporting the mechanism assumed by the model. This univariate approach is commonly referred to as "model-based fMRI", but of course any neural measurement could be correlated with the model measure.

The Latent Input Approach is commonly used in reinforcement learning models to relate mechanisms of learning and prediction errors to the brain (e.g., O'Doherty et al., 2003, 2007; Gläscher and O'Doherty, 2010; Hampton et al., 2006), and has been particularly powerful in the field of clinical neuroscience (e.g., Montague et al., 2012; Wiecki et al., 2015). One simple example is the Rescorla-Wagner (RW) model that characterizes the process of learning a conditioned response through repeated presentations of a conditioned stimulus (Rescorla and Wagner, 1972). In the model, the value of the unconditioned stimulus is represented as $u$, and the value of the conditioned stimulus on Trial $t$ is represented as $v_t$. To learn the stimulus environment, the model assumes that $v_t$ is updated sequentially according to a learning rate parameter $\alpha$, and an evaluation of the prediction error $\epsilon$. Specifically, after a decision is made and the unconditioned stimulus is presented, the model's internal state of the value of the conditioned stimulus is updated according to the rule

$$v_t = v_{t-1} + \alpha\epsilon. \tag{1}$$

Eventually, the internal representation of the value $v$ converges to $u$, $\epsilon$ approaches zero, and the model "learns" the stimulus-to-response pairing. The value of $v_t$ can be directly observed by assessing the strength of the conditioned response, whereas other variables are estimated by fitting the model to behavioral data. Typically, $\alpha$ remains fixed across the trials in an experiment, allowing one to derive a trial-by-trial estimate of $\epsilon$ through Equation 1. Hence, the model produces trial-to-trial estimates of the value of the conditioned stimulus $v$ and the prediction error $\epsilon$. As outlined above, these values can be entered into an fMRI analysis as a time series by convolving them with a hemodynamic response function (HRF), and then regressing the result against the fMRI data through the general linear model. However, the estimates $v$ and $\epsilon$ are not parameters; instead, they reflect the model's internal state for value and prediction error, respectively. This distinction is important because it separates this analytic approach from other possible Two-stage approaches, such as in van Maanen et al. (2011), which we discuss below.

As the previous example makes clear, Latent Input Approaches can identify candidate neural substrates for theoretical concepts, such as prediction error, that are not directly observable but can be defined within a cognitive model. Entering latent model measures into the imaging analyses is relatively straightforward. Indeed, multiple model measures can be considered simultaneously. For example, Davis et al. (2012) simultaneously analyzed cognitive operations related to recognition and representational uncertainty by including two related measures in the imaging analysis from a cognitive model fit to trial-by-trial category learning data.

*Extensions to Model Discrimination.* One issue with what is commonly referred to as model-based fMRI is that models tend to be preferred to the extent that they correlate with many voxels in the brain. However, it is not clear that this is an appropriate criterion. Because simple cognitive models do not attempt to model every process in the brain, they should not be expected to account for the variance of every voxel. Furthermore, cognitive states may be coded by brain states that are defined by the pattern of activation over voxels. This notion of brain state is multivariate as it depends on the pattern of activity, whereas most model-based analyses focus on univariate correlations between a model measure and an individual voxel.

One approach that attempts to address these deficiencies is model decoding (Mack et al., 2013). Rather than assume a single cognitive model as the "correct" model, this generalization acknowledges that there may be competing cognitive models of the same phenomenon and uses the neural data to adjudicate between those competitors. It is well known in mathematical psychology that models assuming very different internal mechanisms can sometimes predict the same observed behavior. To the extent that different model mechanisms produce different internal model states, one way to discriminate between models predicting the same behavior is to compare those predicted internal model states to observed internal brain states. Models that predict observed behavior but cannot predict internal brain states are rejected.

Consider, for example, the work of the Vanderbilt group discussed earlier (Palmeri et al., 2015; Purcell et al., 2010, 2012). After excluding neurally-constrained models that could not fit the observed behavioral data, they were then able to distinguish further between models based on how well the predicted accumulator model dynamics matched the observed neural dynamics in movement-related neurons, the neurons they hypothesized to carry out an accumulation of evidence (see also Purcell and Palmeri, 2015, in this special issue). Only their gated accumulator model produced accumulator dynamics that matched the observed dynamics of movement-related neurons in FEF.

Consider next the recent work of Mack et al. (2013), who developed a strategy for evaluating different models of object categorization on the basis of their consistency with observed fMRI data. They specifically contrasted two well-known theories of category representation: exemplar and prototype models (see also Palmeri, 2014). Exemplar models assume that members of a category are explicitly stored in memory, and a categorical decision for a new stimulus is a function of its similarity to these remembered exemplars. Prototype models assume that category representations are abstract, averages of experienced category examples, and a categorical decision is a function of similarity to the stored category prototypes. In this sense, the prototype representation is abstract – a category could be represented in a location of feature space that is not representative of any particular known category member. These particular theories of category representation have been fiercely debated for decades (e.g., Medin and Schaffer, 1978; Minda and Smith, 2002; Zaki et al., 2003). Indeed, in their first analysis, Mack et al. (2013) showed that both exemplar and prototype models provided nearly indistinguishable fits to the observed behavioral data.

Even though the exemplar and prototype models make similar predictions about behavior, they do so by assuming very different kinds of internal representations. Indeed, the degree to which different test items activate these internal representations – similarity to stored exemplars for the exemplar model versus similarity to category prototypes for the prototype model – differs considerably between the two models. Mack et al. (2013) asked whether the pattern of brain activity elicited by different test items would be more similar to the pattern of activation of internal representations for the exemplar model or the prototype model. They specifically evaluated the mutual information shared between brain and model state using machine learning techniques like multivariate pattern analysis (MVPA) and representational similarity analysis (RSA). The patterns of brain activity across trials showed better correspondence to the internal state of the exemplar representation than the prototype representation. These findings serve as a powerful example of how the neurosciences – combined with a Latent Input Approach – allow us to draw conclusions regarding competing cognitive models that we might not otherwise reach.

These model decoding approaches represent an important departure from the Latent Input Approach discussed above. Namely, these methods do not assume that the model used to interpret the neural data is correct. Instead, they posit a set of competing models for the underlying cognitive process, and the *best* explanation is to be determined from each model's correspondence to the neural data. Once a cognitive model is selected, it can then be used as a lens on the brain data, using any existing technique, such as the aforementioned univariate approaches or representation similarity analysis (RSA). This stage of the analysis can be seen as confirmatory – the winning model has been established and is used to help interpret the neural data. Pairing model decoding with a model-based analysis approach allows for information from brain and behavior to be mutually constraining through the bridge of the cognitive model. This extra step of selecting a model based on neural data is atypical of Latent Input Approaches, and this step is not illustrated in Figure 1.

### *2.2.2. Two-stage Neural Approach*

The second approach we will discuss that uses behavior to predict neural data is related to the Two-stage Behavioral Approach discussed above, except that here, the parameters of the behavioral model $\theta$ are used to guide the analysis of the neural data $N$ instead of vice versa. While a subset of neural model parameters $\delta$ could be replaced with a subset of behavioral model parameters $\theta$ akin to the Two-stage Behavioral Approach, in practice, this is rarely done. Instead, relationships between $\theta$ and $\delta$ are formed through correlational or regression analyses. The correlational approach has been especially successful in the field of perceptual decision making (Mulder et al., 2014). For example, Forstmann et al. (2008), Forstmann et al. (2010), and Mansfield et al. (2011) show in various experimental setups that accumulator model parameters that reflect response caution correlate with averaged BOLD responses in pre-supplementary motor area and striatum, two regions in the brain that are thought to be involved in mediating cognitive control. These studies illustrate that individual differences in behavior, captured by hypothesized processes, are driven by individual differences in how the brain works. This approach thus strengthens our understanding of the role of certain brain areas in cognition, but it also adds credence to the type of cognitive model that is adopted to describe behavior.

In the regression approach, parameters of a behavioral model are used as predictors in a regression model of the neural variables. In the context of fMRI, behavioral model parameters are often entered as regressors in a general linear model that quantifies the BOLD response in certain brain areas (e.g., Mulder et al., 2012; Summerfield and Koechlin, 2010; White et al., 2014). Usually, this is done in addition to regressors that relate to the experimental manipulations, yielding statistical maps of brain activation that reflect the predicted change in neural activation (i.e., in $\delta$) for a fixed change in behavioral model parameter ($\theta$), in addition to the standard notion of a change in $\delta$ as a function of the experimental manipulation.

Some properties of behavior are difficult to cast in experimental conditions. For example, fluctuations that occur as part of a time series of observations are ideally analyzed as such (Wagenmakers et al., 2004). Moreover, these fluctuations may be related to incorrect (Dutilh et al., 2012; Eichele et al., 2008) or task-unrelated responses, for example due to attentional lapses (Weissman et al., 2006; Mittner et al., 2014). For these situations it can be useful to study fluctuations in brain and behavior over time.

To understand how the variability in brain measures from trial to trial adds to the behavioral variability, some researchers have developed models in which parameters are estimated on a trial-by-trial basis (Behrens et al., 2007; Brunton et al., 2013; Erlich et al., 2015; Hanks et al., 2015; van Maanen et al., 2011). For example, Behrens et al. (2007) used an optimal model that updates the expected reward for one of two responses on a trial-by-trial basis. The parameters of this model were also updated on a trial-by-trial basis, based on the actual trial outcome (i.e., the choice of the participant) and the expected outcome (i.e., the model prediction). Behrens and colleagues found that the level at which participants were responsive to changes in the rewards was predictive of anterior cingulate cortex activation on a trial-by-trial basis, supporting the idea that anterior cingulate cortex activation reflects changes in the environment (e.g., Rushworth et al., 2009).

A slightly different approach was taken by Van Maanen and colleagues (van Maanen et al., 2011; Ho et al., 2012; Boehm et al., 2014). Using the LBA model, these authors estimated the most likely combination of drift rate and starting point of evidence accumulation, given the distribution of these parameters across trials. The most likely combination of parameters is determined by the set of parameters that specify the response time. While powerful, this method is difficult because the most likely parameter estimates are highly uncertain, due to the large variability in the joint distribution of the model parameters, and due to the simplification of the model to include only two sources of variability. Nevertheless, van Maanen et al. (2011) showed that trial-to-trial fluctuations in BOLD in pre-supplementary motor area correlated with the trial-to-trial measure of threshold, but only for speed-stressed trials. This finding was corroborated by Boehm et al. (2014), who found a similar correlation between the trial-to-trial model parameter and a trial-to-trial estimate of the Contingent Negative Variation (CNV). The CNV is a slow rising potential, thought to represent neural activation in a cortico-basal ganglia loop including the supplementary/pre-supplementary motor areas (Nagai et al., 2004; Plichta et al., 2013).

Although the Two-stage Neural Approach has been instrumental in elucidating various mechanistic explanations of neural data, the framework neglects an important source of constraint. Namely, by analyzing the neural and behavioral data independently, the secondary analysis does not statistically guide our understanding of how these variables are related. In this way, Two-stage frameworks are not statistically reciprocal because the neural data cannot influence the parameter estimates of the behavioral model (cf. Forstmann et al., 2011). To accomplish such a goal, a framework would need to automatically learn the covariation of the neural and behavioral parameters in harmony with the analysis of the neural and behavioral data. Such a framework is the topic discussed in the next section: Simultaneous Modeling.

### *2.3. Simultaneous Modeling*

At this point, we have discussed two general analytic approaches that apply *unidirectional* statistical influence: modeling and analysis of one source of data guides the modeling and analysis of another source. The primary motivation of these approaches is that one measure is particularly well suited for answering a key theoretical question. In this way, one measure carries more "theoretical importance" than the other. However, some modeling approaches are agnostic in specifying which measure is more important, and instead posit a *bidirectional* link between the two measures. Similar to the subdivisions in other research goals above, the level at which the link is established is an important distinction between the two approaches, which we will now discuss in turn.

#### *2.3.1. Joint Modeling Approach*

The next approach we discuss is the recently developed Joint Modeling framework (Turner et al., 2013b; Turner, 2015; Turner et al., 2015b). The Joint Modeling Approach is conceptually similar to the Two-stage Neural Approach in that it attempts to relate the parameters of the behavioral model to the parameters of the neural model. However, statistically speaking, the Joint Modeling Approach is unique in the way it bridges this connection. Specifically, it assumes an overarching distribution that enforces an explicit connection between these parameters. The bottom-left panel of Figure 1 illustrates this connection via the parameters $\Omega$ that link $\theta$ to $\delta$. In this illustration, the connection enforced by $\Omega$ is clearly abstract; one must make a specific assumption about how $\theta$ and $\delta$ should coexist in their explanation of the underlying cognitive process. As an example, one simple linking function used in practice has been the multivariate normal distribution where $\Omega$ consists of the hyper mean vector and the hyper variance-covariance matrix. This connection is important because it allows the information contained in the neural data $N$ to affect the information we learn about the behavioral model parameters $\theta$.

Perhaps the greatest benefit of the Joint Modeling Approach is its flexibility – it can be applied to different modalities (e.g., fMRI or EEG data), make different assumptions about the underlying cognitive process (i.e., changing the behavioral submodel), and establish a link at any number of levels in a hierarchical model. For example, Turner et al. (2013b) used structural diffusion weighted imagining data to explain differences in patterns of choice response time data across subjects. They showed how a joint model equipped with information about the interconnectivity of important brain areas could make accurate predictions about a subject's behavioral performance in the absence of behavioral data. Turner et al. (2015b) extended this approach to build in brain state fluctuations measured with fMRI into the DDM. The problem Turner et al. (2015b) addressed centered on a lack of information about within-trial accumulation dynamics. In behavioral choice response time experiments, following the presentation of a stimulus, researchers can only observe the eventual choice and response time. These data are then used to estimate parameters of a cognitive model, following an assumption that the data observed on each of these trials arises from the same psychological process. However, this assumption – known as stationarity – is a strong one, and is seldom observed in empirical data (e.g., Peruggia et al., 2002; Craigmile et al., 2010). Turner et al. (2015b) used a multivariate model to describe the joint activation of a set of brain regions of interest, and used this description to enhance the classic DDM. In a cross validation test, they showed that their extended model could generate better predictions about behavioral data than the DDM alone, demonstrating that neurophysiology can be used to improve explanations about trial-to-trial fluctuations in behavior.

Effectively, the Joint Modeling Approach is a strategy for treating groups of parameters as covariates, and this covariation is learned through hierarchical modeling. However, one could imagine an approach for performing model-based cognitive neuroscience that is similar to the Two-stage Neural approach, but instead of correlating or regressing variables after independent analyses, the parameters of the regression equation are estimated. Such an approach can be thought of as a Joint Modeling Approach, except the linking parameters $\Omega$ are deterministic. Recently, this approach has been used in cognitive neuroscience to link decision models to neural fluctuations. For example, Nunez et al. (2015) used EEG data on a perceptual decision making experiment as a proxy for attention. They controlled the rate of flickering stimuli presented to subjects to match the sampling rate of their EEG data, a measure known as the steady-state visual evoked potential. Importantly, Nunez et al. (2015) showed that individual differences in attention or noise suppression was indicative of the choice behavior, specifically it resulted in faster responses with higher accuracy. In a particularly novel application, Frank et al. (2015) showed how models of reinforcement learning could be fused with the DDM to gain insight into activity in the subthalamic nucleus (STN). In their study, Frank et al. (2015) used simultaneous EEG and fMRI measures as a covariate in the estimation of single-trial parameters. Specifically, they used pre-defined regions of interest including the presupplementary motor area, STN, and a general measure of mid-frontal EEG theta power to constrain trial-to-trial fluctuations in response threshold, and BOLD activity in the caudate to constrain trial-to-trial fluctuations in evidence accumulation. Their work is important because it establishes concrete links between STN and pre-SMA communication as a function of varying reward structure, as well as a model that uses fluctuations in decision conflict (as measured by multimodal activity in the dorsomedial frontal cortex) to adjust response threshold from trial-to-trial.

The major limitation of the Joint Modeling Approach is its complexity, which hinders our ability to use the approach effectively in two ways. First, to estimate all of the model parameters, we must perform a sophisticated system of Markov chain Monte Carlo sampling with updates on separate blocks of model parameters (see Turner et al., 2013b; Turner, 2015; Turner et al., 2015b, 2013c, for details). This involves deriving the conditional distribution of blocks of parameters, and if desired, establishing conjugate relationships between the prior and posterior for effective estimation. One example of this has been the use of a multivariate normal assumption to link neural and behavioral submodel parameters (Turner et al., 2013b, 2015b). In this approach, an increase in any neural measure automatically scales the increase in the behavioral model parameters, and vice versa. Second, a great deal of data must be available to appreciate the magnitude of the effects of interest. This result is driven by a complexity/flexibility tradeoff we discuss below, but the basic idea is that as the number of parameters increases, the influence the data can have on the joint posterior distribution decreases. When a model is complex relative to the data, one simple approach to reduce the complexity is to reduce the number of model parameters (Myung and Pitt, 1997). In hierarchical models like the Joint Modeling Approach, one way to accomplish this is to reduce the number of levels in the hierarchy by removing its submodels (i.e., models within the Joint Model that explain one subset of the data). Such a strategy constitutes our final approach: the Integrative approach.

### *2.3.2. Integrative Approach*

In the Integrative approach, the goal is to develop a single cognitive model capable of predicting both neural and behavioral measures. This approach, illustrated in the bottom-right panel of Figure 1, uses one set of parameters $\theta$ to explain the neural $N$ and behavioral $B$ data jointly. Notice that the Integrative approach differs from the Joint Modeling Approach because the parameters $\theta$ are directly connected to the data – there is no overarching distribution $\Omega$ to intervene between the data sources. Integrative approaches allow the neural data $N$ to have a greater influence on the behavioral data $B$, a statistical property that can be measured by mutual information.

Of the approaches we have discussed, the Integrative approach is arguably the most difficult to develop. Its use requires strong commitments to both the underlying cognitive process and where this process is executed in the brain. One technical hurdle in using an Integrative approach lies in the description of random variables with different temporal properties. For example, neurophysiological measures are typically observed on a moment-by-moment basis, detailing activation in the brain throughout the trial. By contrast, behavioral data are typically observed only at the end of a trial, such as in any number of perceptual decision making tasks. So, in the instantiation of a cognitive theory that uses the Integrative approach, we would need a moment-by-moment prediction of neural data, and a trial-by-trial prediction of the behavioral data, usually assumed to be the result of a series of unobservable (i.e., latent) processes. Given the unique structure of Integrative approaches, properly fitting them to data is a difficult task, often involving sophisticated techniques such as Hidden Markov Models (e.g., Anderson et al., 2010; Anderson, 2012), or Bayesian change point analyses (e.g., Mohammad-Djafari and Féron, 2006).

Some recent applications of ACT-R have aimed for this Integrative Approach. ACT-R assumes the presence of distinct cognitive modules that are recruited sequentially during a task. The recruitment of these modules across the time course of the task can be represented as a vector of binary outcomes, such that a 1 indicates that a module is being used, and a 0 indicates it is not being used. This vector naturally lends itself to convolution with the canonical HRF in the same way as experimental design variables (i.e., called the design matrix). The result of the convolution is a model-generated BOLD signal that can be compared to empirical data. In this way, the ACT-R model can actually be used in both exploratory and confirmatory research. When used for exploration, the model-generated BOLD signal is regressed against the data in a voxel-by-voxel fashion through the general linear model (Borst et al., 2010b; Borst and Anderson, 2013). From this analysis, clusters of voxels typically emerge, and these clusters are taken to represent brain areas where the modules are physically executed. This explorative analysis more closely resembles the Latent Input Approach. However, the ACT-R model can also be used in a confirmatory fashion (Anderson, 2007; Anderson et al., 2008a,b; Borst et al., 2010a). To do this, Anderson and colleagues have identified which brain areas should become active during the recruitment of different modules (Anderson et al., 2008b; Borst et al., 2015). These brain areas were identified primarily from several exploratory analyses (Anderson, 2007), but recent work has taken these explorations to generate out-of-sample, confirmatory predictions for neural data. In these confirmatory studies, the specific pattern of module activations (i.e., the parameters $\theta$) in the model simultaneously affects the model's predictions for the BOLD response and the behavioral outcome. Although global, whole-brain predictions could be made within this framework, the strict assumption of localized module activity in the brain constitutes a fully confirmatory Integrative approach, where predictions for neural activity – as well as behavioral data – can be quantitatively evaluated.

The ACT-R framework provides an unique perspective on performing the integration between neural and behavioral measures, but actually testing these models is nontrivial. The major limitation is that one must assume a set of specific modules, and the activation of these modules in the behavioral model is latent, which makes their activation difficult to identify in behavioral data. Although neural data facilitate this identification process, current solutions rely heavily on assumptions about how modules are represented in patterns of neural activity (Anderson, 2012). Furthermore, it is unclear how one would objectively decompose other cognitive models into a discrete set of modules while preserving their key theoretical and convenient properties (for examples of cognitive models in the style of ACT-R, see van Maanen and Van Rijn, 2010; van Maanen et al., 2012, 2009). For example, the Linear Ballistic Accumulator (LBA; Brown and Heathcote, 2008) model has enjoyed widespread success due to its parsimony and remarkable mathematical tractability. Breaking the LBA model down into its constituent parts could compromise this tractability in such a way that estimation of the model's parameters would be nontrivial. Hence, it is clear that not every cognitive model can easily be transformed and prepared for an analysis using the Integrative Approach. At this point, a natural question to ask is, under what conditions should an approach be used for an analysis?

## 3. Comparing the Approaches

It is important to supplement our discussion of approaches to model-based cognitive neuroscience with a guide to how these approaches compare. This comparison is difficult and likely to be highly subjective. How should the various approaches be evaluated? Along what dimensions should they be compared and contrasted? Do these approaches cover all possible types of linkage between neural and behavioral measures? Despite our fear of improperly considering these questions, we will persist and attempt to organize the six core approaches discussed in this article along dimensions that are relevant for practical implementation (note that we have grouped both types of Two-Stage approaches together for this discussion). Table 1 provides a list of key factors that can be used to compare the strengths and weaknesses of the approaches.

### *3.1. Number of Stages*

The first factor we could compare the approaches on is the number of processing stages. The fewest number of stages occur when the approach considers both measures simultaneously. Because both the Joint Modeling Approach and the Integrative approach make formal assumptions about how

Table 1: A comparison between the six different analytic approaches on five important factors. Note that these descriptions have been formed on the work we are familiar with, and the factors represent considerations that are especially important to us.

| Factor | Theoretical | Two-stage | Direct Input | Latent Input | Joint Modeling | Integrative |
|---|---|---|---|---|---|---|
| Number of Stages | $\{2, 3, \ldots\}$ | { 2, 3 } | 1 | { 2, 3 } | 1 | 1 |
| Commitment to a Particular Theory | none | weak | medium | weak | weak | strong |
| Type of Information Flow | conceptual | one-way | one-way | one-way | two-way | two-way |
| Difficulty of Implementation | high | low | medium | medium | high | high |
| Type of Exploration | exploratory | exploratory | confirmatory | either | either | confirmatory |

both behavioral and neural measures arise, a full computational model is fit to the entire set of data in one stage. Another approach requiring only one stage is the Direct Input Approach, where the neural data replace dynamics of the behavioral model. Here, only the behavioral data are considered while fitting the model to data, but this process still only requires a single processing stage. The Latent Input and Two-stage approaches typically require the greatest number of stages at two or sometimes three. If a separate simulation stage is required to generate neural predictions $N^*$, Latent Input Approaches have three stages, whereas if the internal state of a model can be directly inferred when the behavioral model is fit to behavioral data (e.g., as in the reinforcement models described above), then the Latent Input Approach only requires two stages. In the Two-stage approach, if the parameters of the behavioral model can be regressed (or correlated with) the raw neural data, then only two stages are required. However, if some preliminary analyses of the neural data are required, then the Two-stage approach will require three stages. Finally, the Theoretical Approach can require anywhere from two to an infinite number of stages. In the simplest scenario, the first stage consists of observing some pattern or phenomena of interest in the neural data, and the second stage consists of the development of a behavioral model. However, Theoretical Approaches can also be complex to implement because they can involve an extensive, iterative process of running new experiments and refining a developing model (Shiffrin and Nobel, 1997).

### *3.2. Commitment to a Particular Theory*

The second factor involves the role of flexibility in applying new theories to the data. For example, we consider the Two-stage Approach to have weak commitment to any particular theory: one could freely use the same procedure to test any number of behavioral models on the same neural data. The commitment to a particular theory is similarly weak in the Latent Input and Joint Modeling Approaches, where behavioral models can easily be switched out and fits to data compared. We consider the Direct Input Approach to be more committed to a particular theory than these aforementioned approaches. For example, while Purcell et al. (2010) used neural data to test different assumptions about the accumulation process, they still maintained a commitment to the sequential sampling framework for these models. In this way, their analysis relies on some theoretical assumptions about the accumulation process, but not in a way that is inflexible. Going one step beyond this is the Integrative Approach, which requires strong commitments to a particular modeling framework such as in Anderson and colleagues' work (e.g., Anderson et al., 2008b; Borst et al., 2015). In this approach, it is difficult to imagine testing different models that are not contained within a similar overarching theory. Finally, the Theoretical Approach makes no commitment to any particular theory, instead it uses the data to guide the development of the theory itself.

### *3.3. Type of Information Flow*

Another factor to consider is the type of information flow. In Table 1, we consider three types: conceptual, one-way, and two-way. In the Theoretical Approach, the neural data can only guide the development of the behavioral model conceptually – there is no formal relationship between the behavioral and neural measures. At the other extreme, both the Joint Modeling and Integrative approaches use the information contained in either measure to directly constrain the estimates of the models' parameters. Hence, we refer to this type of information flow as two-way because information flows in both directions. When one source of data enforces direct constraint on the other measure, we refer to this type of information flow as one-way. All of the remaining approaches use information flow that is one-way to maximize constraint in their models.

While on the surface, a one-way information flow may seem a weakness, there are sometimes important theoretical reasons for enforcing this strict directionality. Consider, for example, the illustrated uses of the Latent Input Approach for model discrimination (Mack et al., 2013; Palmeri et al., 2015; Purcell et al., 2010, 2012; Palmeri, 2014). Here the goal was to use neural data to help discriminate between models of perceptual decision making or models of categorization that make the same behavioral predictions. The models were fit to the behavioral data in exactly the same way they might be fit if neural data were not even considered. No compromises were made in the behavioral fits to take into account the neural data, as might be the case for the Joint Modeling or Integrative Approaches. Only after the models were fit to the behavioral data were the predicted internal states of the model then compared to observed neural states in the brain. Finally, models were rejected if they could not adequately capture those observed neural states in the brain.

### 3.4. Difficulty of Implementation

From a pragmatic perspective, it is also important to consider the difficulty of performing analyses with these six approaches. Perhaps the easiest approach to implement for the readers of this special issue is the Two-stage Approach, where the parameters of a cognitive model are simply regressed against a neural signal of interest. Of medium difficulty are the Direct Input and Latent Input Approaches, because they often require model simulations or additional theoretical overhead to fit the models to data. The Joint Modeling and Integrative Approaches are considered difficult to implement because they either require sophisticated partitioning of the parameter space (e.g., Turner et al., 2015b), or estimation of hidden Markov model parameters (e.g., Anderson et al., 2010; Anderson, 2012). Perhaps the most difficult approach to implement is the Theoretical Approach, where models must be carefully constructed and iteratively fit to data as a test of specific assumptions. To make matters worse, there is no clear end point when developing a new cognitive model in the Theoretical Approach.

### 3.5. Type of Exploration

A final consideration is the type of exploration that can be used under a specific approach. Approaches can be used for exploratory or confirmatory purposes, or some mixture of the two. The Theoretical and Two-stage Approaches are considered exploratory because the general strategy involves a sequence of tests, iterating toward a solution or explanation of the data. The Direct Input Approach is considered a confirmatory approach because the neural data are used to directly replace certain mechanisms in the model, providing a test of the neural measure's plausibility in predicting the behavioral response. The Integrative Approach is also confirmatory because it makes specific assumptions about how both measures arise, where good fits to data support the assumptions of the model, and poor fits refute them. We regard the Latent Input Approach as being exploratory when used in a typical "model-based" analysis, but confirmatory when used to compare models to one another as in Mack et al. (2013) and Purcell et al. (2012). In this way, the Latent Input Approach is listed as "either" because the specific usage depends on the situation. Finally, the Joint Modeling Approach is also considered both confirmatory and exploratory, because its usage depends on the how the linking function is specified. For example, one could use a general linear model as the linking function – a confirmatory approach – or one could use ambiguous priors on hyperparameters that specify a multivariate

Gaussian linking function – an exploratory approach. Furthermore, the specific prior used on the hyperparameters allows the Joint Modeling Approach to mix between confirmatory and exploratory roles in an analysis.

## 4. Choices and Limitations

In this article, our goal was to highlight and discuss the prominent approaches to analysis in the emerging subfield of model-based cognitive neuroscience. However, we have not yet provided a guideline for choosing between them, nor have we discussed in greater detail the limitations of choosing a particular approach. In this section, we will address both of these issues.

### *4.1. Choosing Between Approaches*

Although we have described, compared, and contrasted six important approaches for analysis, we have not provided a guideline for how these approaches could be used to advance psychological theory. We believe that each of these approaches have their own utility in the pursuit and development of computational models, and the primary factor in choosing between them is the goal of the analysis. Furthermore, as a theory progresses, it is important to realize that the goals of an analysis should change. To this end, we advocate using all of these approaches to move from an exploratory analysis to a confirmatory one.

To see how this would work in practice, consider the following stages of model development. In the initial stages, one approach is to develop a cognitive theory by acknowledging patterns in the data from both the brain and the behavior. For example, knowing that the brain must first encode stimulus information in lower-level visual areas before a representation of the stimulus can be perceived and acted upon could be used to impose order in a behavioral model. Such knowledge might motivate the development of a visual encoding component of the model that precedes the development of an accurate stimulus representation. Instantiation of the encoding process in the behavioral model is an implementation of the Theoretical Approach, because the development is motivated by brain data. Here, our goal was to simply develop a model that abides by certain physiological timing restrictions as a way to establish a more constrained stimulus processing order.

After the development of the model, our goals have advanced – suppose we now wish to identify where this encoding component of our model is carried out, and specifically, which areas of the brain contribute to this process.

To accomplish this goal, we would elect to use an exploratory analysis, such as the Two-stage or Latent Input Approach. In the Two-stage analysis, we would simply fit our behavioral model to the behavioral data, and correlate the parameters regulating the encoding process of our model to say, parameters of the HRF in our neural data. Similarly, in the Latent Input analysis, we would use the timing of the encoding component in our model to search for temporally-related activations in the brain. Both of these analyses constitute searches through our neural data as a way to better understand how the brain produces behavior from a mechanistic perspective. In this way, these analyses are unidirectional and do not validate or confirm our model, but this is perfectly acceptable because it is consistent with our current goals.

Our exploratory analyses have paved the way for subsequent investigations, and now suppose we wish to use the neural data to better constrain our behavioral model. We now have well-defined hypotheses about which brain areas are involved in stimulus encoding, and we suspect that the systematic activations in these brain areas have a correspondence to the encoding phase of our model. At this point, we must reconsider our specific goals. If the goal of our analysis is to predict behavior, we might use the Direct Input Approach to map activations in the key brain areas directly to the encoding component of our model. By contrast, if our goal is to infer relationships between the neural and behavioral measures, we might use the Joint Modeling Approach to test specific impositions of brain activations to the parameters regulating the encoding process in our model. Both of these approaches are more confirmatory because they rely on specific hypotheses and assumptions that were derived from our exploratory analyses; however, they still only guide our inference. In the Direct Input analysis, because our goal was to predict the behavioral data, we have compromised our ability to evaluate the model's suitability for the neural data. We cannot make predictions about neural data that we have conditioned on, as so we cannot evaluate how well the model captures these aspects of our (neural) data. On the other hand, the Joint Modeling Approach attempts to capture both aspects of the data simultaneously, and as a result, its predictions for the behavioral data are compromised by the model's obligations to the neural data. Because the Joint Modeling Approach does not explicitly condition on either variable, it can reveal interesting *generative* properties of our model, but its *discriminative* (i.e., predictive) power is diminished (Bishop and Lasserre, 2007).

At this point, we have now developed our model and evaluated the relationships between brain and behavior in a variety of analytic approaches.

We know better than anyone in the world where the encoding part of our model is carried out in the brain, and how differences in the pattern of activation in these brain areas contribute to behavioral differences. As a final test and validation of our model, we can now move to the most confirmatory analysis we have discussed here: the Integrative Approach. To establish an integrative model, we must first make some specific assumptions about how activations in key brain areas map to the encoding component of our model. This can be a difficult process, but suppose for now that we have formally articulated this mapping in our model, derived from our previous exploratory analyses. Our goal now is to show that this integrative version of our model can produce patterns of data that match all aspects of our data. That is, adjustments of one model parameter should make specific predictions about how the pattern of neural and behavioral measures changes, and ideally, how these changes could be selectively influenced experimentally (e.g., Heathcote et al., 2015). In our opinion, this integrative analysis represents the strongest test of psychological theory, but such a test would be misguided if not first informed by the less integrative approaches.

### *4.2. Limitations of Using These Approaches*

In our working example above, we identified a few limitations of using various approaches. First, the balancing of fit between behavioral data, neural data, or both is a key consideration in model-based cognitive neuroscience. In general, to optimize predictions for say, behavior, it would be better to condition on neural data. However, if one is more interested in the joint distribution of both neural and behavioral measures, then the modeling goals are more generative than discriminative, and conditioning on one variable would introduce limitations. The authors of the present manuscript have deliberated between these three modeling goals, and arrived at only an ambiguous solution: decisions must be made on a case-by-case basis, always with the researcher's goals in mind.

Second, constraint is not always a good thing. If one does not have strong intuition about how components of a model are carried out in the brain, it would be unwise to impose strong constraints on a model. One way of autonomously carrying out justifiable constraint is to use the approaches discussed here along a continuum of increasingly more confirmatory research. As another tack, one could use some of the approaches discussed here to impose varying levels of constraint, moderating the levels of analyses between exploratory and confirmatory. For example, in the Joint Modeling Approach, one can impose a completely uninformative prior on the parameters of the linking function and specify that all parameters of the behavioral model be mapped to the neural data. Such an analysis is wildly explorative, would be difficult to implement, and would convey little information about the covariation between the measures. To move toward a more confirmatory regime, one could impose a stronger prior derived from say, previous research or investigation of the prior predictive distribution (Vanpaemel, 2010, 2011; Vanpaemel and Lee, 2012). Similarly, one could constrain the set of parameters that are related to the neural data by simply setting elements of the linking function to zero. Such an analysis would provide a greater test of the model, but would also force the model to rely more heavily on the joint distribution of the measures.

Third, in this article, we have emphasized structural connections that are largely at one level. This is a limitation because the behavioral data can be thought of as the end result of some brain process, again highlighting the mismatch between Marr's (1982) implementation and computational levels of analyses we discussed earlier. Another approach would be to impose structural connections that are multi-level, where a model uses the implementation level to drive some mechanisms, and the computational level to drive others. As a hypothetical example, the implementation level could be used to drive an evidence accumulation process that remains unaffected by experimental instructions (i.e., computational goals), whereas other mechanisms such as boundary separation or bias could be carried out by other brain areas that are systematically adjusted in response to task demands. Such a model would bridge the levels of analysis in a way that might actually be reflected in the brain (Frank, 2015).

Finally, the imposition of structure need not arise from a model of behavior. In this article, we have oriented the approaches to analysis around determining where mechanisms in the model are carried out in the brain. However, one can easily imagine reversing the orientation to determining how structural and functional differences in the brain manifest behaviorally. Such an endeavor begins with the development of a generative model of the neural data, usually formed by observing the interconnectedness of key brain regions (Ratcliff and Frank, 2012; Frank, 2006; Wong and Wang, 2006; Cavanagh et al., 2011), and ends in mapping the systematic activations of these brain areas to a model of the behavioral data. These models can be difficult to implement and test in the traditional cognitive modeling way (e.g., Lee and Wagenmakers, 2013; Shiffrin et al., 2008; Heathcote et al., 2015; Busemeyer and Diederich, 2010), because they rely on many parameters and complex simulations to validate them. However, new methods have been developed to better elucidate simulation-based models (for applications in psychology, see Turner and Van Zandt, 2012; Turner and Sederberg, 2012; Turner et al., 2013a; Turner and Sederberg, 2014; Turner and Van Zandt, 2014; Turner et al., 2015a), and as a result, we may gain new insight and interest in these network-style models in the coming years.

### *4.3. Other Approaches*

Although the approaches we have presented here encompass the most prevalent approaches to model-based cognitive neuroscience, other approaches have been used to gain a better understanding of how the brain produces a behavior. One structural example is to use some experimental variable that hypothetically affects the neural data to split the behavioral data into different levels. Once the behavioral data is divided, the data can be fit and evaluated on the basis of differences in parameter values. One example of this is in Parkinson's Disease, where drug therapy is commonly administered to compensate for decreased levels of dopamine. Frank (2006) make predictions for behavioral data for subjects on and off medication in a Go/NoGo task, and a probabilistic learning task. They used a computational neural network model to make concrete predictions for differences in task behavior based on activation of the subthalamic nucleus. Frank (2006) found that their model accurately captured the dynamics of activity in areas of the basal ganglia, and how this pattern of activity related to dynamic adjustments in response thresholds. A similar mechanism was later found in impulse control for Parkinson's patients with deep brain stimulation using a similar analysis design (Cavanagh et al., 2011).

The examples above illustrate an analytic approach where experimental variables guide the analysis of the behavioral data on the basis of how those variables affect the neural data. Another type of analysis takes the effects of the neural data one step further (e.g., Ratcliff et al., 2003, 2007, 2009, 2011; Kiani et al., 2008; Mazurek et al., 2003). For example, Ratcliff et al. (2009) used single-trial amplitude measures of EEG activity in a perceptual decision making experiment to divide their behavioral data into separate groups. Next, Ratcliff et al. fit the DDM to the data from each of these separate groups and used estimates of the drift rate parameter to show early component EEG signals were not reflective of the decision process, whereas late component EEG signals showed a positive correlation to the stimulus evidence (i.e., the drift rate). This type of analysis is similar to the Latent Input Approach, but with the flow of information moving from the neural measures to the behavioral ones. By using the neural data to guide the search for differences in behavioral model parameters, we can better understand the mechanistic properties of these neural features by interpreting them in the native language of the decision model.

## 5. Conclusions

The field of cognitive science has only begun to realize the full potential of combining brain and behavior as a way to study the mind. However, the field relies on the various approaches developed by different groups of methodological experts. Due to the seemingly disjoint ways to study cognition, many neuroscientists and cognitive modelers are unaware of their modeling options, as well as the benefits and limitations of different approaches. In this article, we have described the currently prominent general methods for integrating neural and behavioral measures, while providing some examples of their use in cognitive neuroscience. We then attempted to organize these approaches on the basis of a variety of factors: the number of stages, the commitment to a particular theory, the type of information flow, the difficulty of implementation, and the type of exploration. We concluded with a discussion of limitations and further considerations in approaching the integration problem. Our comparison of the approaches (see Figure 1, and Table 1) highlights that a broad spectrum of methods exist for performing model-based cognitive neuroscience, and there are important considerations and limitations of each approach. In the end, we conclude that model-based approaches in cognitive neuroscience are extremely important (cf. Schall, 2004; Forstmann et al., 2011, 2015; Mulder et al., 2014; White and Poldrack, 2013), and the choice of analysis strongly depends on the research goal. It seems to us that having a clearly articulated analytic goal in mind serves as the impetus for successful integration between neuroscientific measures and cognitive theory.

## 6. References


Anderson, J. A., 1977. Neural models with cognitive implications. In: LaBerge, D., Samuels, S. J. (Eds.), Basic processes in reading: Perception and comprehension. Erlbaum, Hillsdale, NJ, pp. 27–90.

Anderson, J. R., 2007. How can the human mind occur in the physical universe? Oxford University Press, New York, NY.

Anderson, J. R., 2012. Tracking problem solving by multivariate pattern analysis and hidden markov model algorithms. Neuropsychologia 50, 487–498.

Anderson, J. R., Betts, S., Ferris, J. L., Fincham, J. M., 2010. Neural imaging to track mental states. Proceedings of the National Academy of Sciences of the United States 107, 7018–7023.

Anderson, J. R., Byrne, D., Fincham, J. M., Gunn, P., 2008a. Role of prefrontal and parietal cortices in associative learning. Cerebral Cortex 18, 904–914.

Anderson, J. R., Fincham, J. M., Qin, Y., Stocco, A., 2008b. A central circuit of the mind. Trends in Cognitive Science 12, 136–143.

Behrens, T., Woolrich, M., Walton, M., Rushworth, M., 2007. Learning the value of information in an uncertain world. Nature Neuroscience 10, 1214–1221.

Bishop, C. M., Lasserre, J., 2007. Generative or discriminative? Getting the best of both worlds. Bayesian Statistics 8, 3–24.

Boehm, U., Van Maanen, L., Forstmann, B., Van Rijn, H., 2014. Trial-by-trial fluctuations in CNV amplitude reflect anticipatory adjustment of response caution. NeuroImage 96, 95–105.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., Cohen, J. D., 2006. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. Psychological Review 362, 1655–1670.

Bogacz, R., Usher, M., Zhang, J., McClelland, J. L., 2007. Extending a biologically inspired model of choice: Multi-alternatives, nonlinearity and value-based multidimensional choice. Theme issue on modeling natural action selection. Philosophical Transactions of the Royal Society: B. Biological Sciences 362, 1655–1670.

Borst, J. P., Anderson, J. R., 2013. Using model-based functional MRI to locate working memory updates and declarative memory retrievals in the fronto-parietal network. Proceedings of the National Academy of Sciences of the United States 110, 1628–1633.

Borst, J. P., Nijboer, M., Taatgen, N. A., Van Rijn, H., Anderson, J. R., 2015. Using data-driven model-brain mappings to constrain formal models of cognition. PLoS ONE 10, e0119673.

Borst, J. P., Taatgen, N. A., Stocco, A., Van Rijn, H., 2010a. The neural correlates of problem states: Testing fMRI predictions of a computational model of multitasking. PLoS ONE 5, e12966.

Borst, J. P., Taatgen, N. A., Van Rijn, H., 2010b. The problem state: A cognitive bottleneck in multitasking. Journal of Experimental Psychology: Learning, Memory, & Cognition 36, 363–382.

Broadbent, D., 1985. A question of levels: Comment on McClelland and Rumelhart. Journal of Experimental Psychology: General 114, 189–192.

Brown, S., Heathcote, A., 2008. The simplest complete model of choice reaction time: Linear ballistic accumulation. Cognitive Psychology 57, 153–178.

Brunton, B. W., Botvinick, M. M., Brody, C. D., Apr 2013. Rats and humans can optimally accumulate evidence for decision-making. Science 340 (6128), 95–8.

Busemeyer, J. R., Diederich, A. (Eds.), 2010. Cognitive modeling. Sage.

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., Frank, M. J., 2011. Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. Nature Neuroscience 14, 1462–1467.

Craigmile, P., Peruggia, M., Zandt, T. V., 2010. Hierarchical Bayes models for response time data. Psychometrika 75, 613–632.

Davis, T., Love, B. C., Preston, A. R., 2012. Striatal and hippocampal entropy and recognition signals in category learning: simultaneous processes revealed by model-based fmri. Journal of experimental psychology. Learning, memory, and cognition 38 (4), 821–39.

Ditterich, J., 2010. A comparison between mechanisms of multi-alternative perceptual decision making: Ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. Frontiers in Neuroscience 4, 184.

Dutilh, G., Vandekerckhove, J., Forstmann, B., Keuleers, E., Brysbaert, M., Wagenmakers, E., 2012. Testing theories of post-error slowing. Attention, Perception & Psychophysics 74, 454–465.

Eichele, T., Debener, S., Calhoun, V. D., Specht, K., Engel, A. K., Hugdahl, K., von Cramon, D. Y., Ullsperger, M., 2008. Prediction of human errors by maladaptive changes in event-related brain networks. Proceedings of the National Academy of Sciences of the United States 16, 6173–6178.

Erlich, J. C., Brunton, B. W., Duan, C. A., Hanks, T. D., Brody, C. D., 2015. Distinct effects of prefrontal and parietal cortex inactivations on an accumulation of evidence task in the rat. Elife 4.

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., Bogacz, R., Turner, R., 2010. Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. Proceedings of the National Academy of Sciences 107, 15916–15920.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., Wagenmakers, E.-J., 2008. Striatum and pre-SMA facilitate decision-making under time pressure. Proceedings of the National Academy of Sciences 105, 17538–17542.

Forstmann, B. U., Ratcliff, R., Wagenmakers, E.-J., 2015. Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions, in Press.

Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., Serences, J. T., 2011. Reciprocal relations between cognitive neuroscience an formal cognitive models: opposites attract? Trends in Cognitive Sciences 15, 272–279.

Frank, M., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., Badre, D., 2015. fmri and eeg predictors of dynamic decision parameters during human reinforcement learning. Journal of Neuroscience 35 (2), 485–494.

Frank, M. J., 2006. Hold your horses: A dynamic computational role for the subthalamic nucleus in decision-making. Neural Networks 19, 1120–1136.

Frank, M. J., 2015. Linking across levels of computation in model-based cognitive neuroscience. In: Forstmann, B. U., Wagenmakers, E.-J. (Eds.), An introduction to model-based cognitive neuroscience. Springer, New York, pp. 159–177.

Gao, J., Tortell, R., McClelland, J. L., 2011. Dynamic integration of reward and stimulus information in perceptual decision-making. PLoS ONE 6, 1–21.

Gershman, S. J., Blei, D. M., Pereira, F., Norman, K. A., 2011. A topographic latent source model for fMRI data. Neuroimage 57, 89–100.

Gläscher, J. P., O'Doherty, 2010. Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. WIREs Cognitive Science 1, 501–510.

Green, D. M., Swets, J. A., 1966. Signal detection theory and psychophysics. Wiley Press, New York.

Grossberg, S., 1978. A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In: Rosen, R., Snell, F. (Eds.), Progress in theoretical biology. Vol. 5. Academic Press, New York, pp. 233–374.

Hampton, A. N., Bossaerts, P., O'Doherty, J. P., 2006. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. Journal of Neuroscience, 8360–8367.

Hanes, D. P., Schall, J. D., 1996. Neural control of voluntary movement initiation. Science 274, 427–430.

Hanks, T. D., Kopec, C. D., Brunton, B. W., Duan, C. A., Erlich, J. C., Brody, C. D., Apr 2015. Distinct relationships of parietal and prefrontal cortices to evidence accumulation. Nature 520 (7546), 220–3.

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

Ho, T., Brown, S., van Maanen, L., Forstmann, B. U., Wagenmakers, E.-J., Serences, J. T., 2012. The optimality of sensory processing during the speed-accuracy tradeoff. J Neurosci 32, 7992–8003.

Horwitz, G. D., Newsome, W. T., 1999. Separate signals for target selection and movement specification in the superior colliculus. Science 284, 1158–1161.

Horwitz, G. D., Newsome, W. T., 2001. Target selection for saccadic eye movements prelude activity in the superior colliculus during a direction-discrimination task. Journal of Neurophysiology 86, 2543–2558.

Kemp, C., Tenenbaum, J. B., 2008. Structured models of semantic cognition. Commentary on Rogers and McClelland. Behavioral and Brain Sciences 31, 717–718.

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

Kim, J. N., Shadlen, M. N., 1999. Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. Nature Neuroscience 2, 176–185.

Lee, M. D., Wagenmakers, E.-J., 2013. Bayesian modeling for cognitive science: A practical course. Cambridge University Press.

Love, B. C., 2015. The algorithmic level is the bridge between computation and brain. Topics in Cognitive Science 7.

Mack, M. L., Preston, A. R., Love, B. C., 2013. Decoding the brain?s algorithm for categorization from its neural implementation. Current Biology 23, 2023–2027.

Mansfield, E. L., Karayanidis, F., Jamadar, S., Heathcote, A., Forstmann, B. U., Oct 2011. Adjustments of response threshold during task switching: a model-based functional magnetic resonance imaging study. J Neurosci 31 (41), 14688–92.

Marr, D., 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. New York: Freeman.

Massaro, D. W., 1988. Some criticisms of connectionist models of human performance. Journal of Memory and Language 27, 213–234.

Mazurek, M. E., Roitman, J. D., Ditterich, J., Shadlen, M. N., 2003. A role for neural integrators in perceptual decision making. Cerebral Cortex 13, 1257–1269.

McClelland, J., Rumelhart, D., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models. MIT Press, Cambridge, MA.

McClelland, J. L., Rumelhart, D. E., 1981. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review 8, 375–40.

McCullogh, W. S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133.

Medin, D. L., Schaffer, M. M., 1978. Context theory of classification learning. Psychological Review 85, 207–238.

Minda, J. P., Smith, J. D., 2002. Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. Journal of Experimental Psychology: Learning, Memory, and Cognition 28, 275–292.

Minsky, M. L., Papert, S. A., 1969. Perceptrons. The MIT Press, Cambridge, MA.

Mittner, M., Boekel, W., Tucker, A. M., Turner, B. M., Heathcote, A., Forstmann, B. U., 2014. When the brain takes a break: A model-based analysis of mind wandering. Journal of Neuroscience 34, 16286–16295.

Mohammad-Djafari, A., Féron, O., 2006. A Bayesian approach to change point analysis of discrete time series. International Journals of Imaging Systems and Technology 16, 215–221.

Montague, P. R., Dolan, R. J., Friston, K. J., Dayan, P., 2012. Computational psychiatry. Trends in Cognitive Science 16, 72–80.

Mulder, M., van Maanen, L., Forstmann, B. U., 2014. Perceptual decision neurosciences – A model-based review. Neuroscience 277, 872–884.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., Forstmann, B. U., 2012. Bias in the brain: A diffusion model analysis of prior probability and potential payoff. Journal of Neuroscience 32, 2335–2343.

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

Nagai, Y., Critchley, H. D., Featherstone, E., Fenwick, P. B. C., Trimble, M. R., Dolan, R. J., Apr 2004. Brain activity relating to the contingent negative variation: an fmri investigation. Neuroimage 21 (4), 1232–41.

Nunez, M. D., Srinivasan, R., Vandekerckhove, J., 2015. Individual differences in attention influence perceptual decision making. Frontiers in Psychology 8 (18), 1–13.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., Dolan, R. J., 2003. Temporal difference models and reward-related learning in the human brain. Neuron 28, 329–337.

O'Doherty, J. P., Hampton, A., Kim, H., 2007. Model-Based fMRI and its application to reward learning and decision making. Annals of the New York Academy of Science 1104, 35–53.

Ossmy, O., Moran, R., Pfeffer, T., Tsetsos, K., Usher, M., Donner, T. H., 2013. The timescale of perceptual evidence integration can be adapted to the environment. Current Biology 23, 981–986.

Palmeri, T., Schall, J., Logan, G., 2015. Neurocognitive modelling of perceptual decisions. In: Busemeyer, J. R., Townsend, J., Wang, Z. J., Eidels, A. (Eds.), Oxford Handbook of Computational and Mathematical Psychology. Oxford University Press.

Palmeri, T. J., 2014. An exemplar of model-based cognitive neuroscience. Trends in Cognitive Science 18, 67–69.

Peruggia, M., Van Zandt, T., Chen, M., 2002. Was it a car or a cat I saw? An analysis of response times for word recognition. Case Studies in Bayesian Statistics VI, 319–334.

Pinker, S., Prince, A., 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition 28, 73–193.

Plichta, M. M., Wolf, I., Hohmann, S., Baumeister, S., Boecker, R., Schwarz, A. J., Zangl, M., Mier, D., Diener, C., Meyer, P., Holz, N., Ruf, M., Gerchen, M. F., Bernal-Casas, D., Kolev, V., Yordanova, J., Flor, H., Laucht, M., Banaschewski, T., Kirsch, P., Meyer-Lindenberg, A., Brandeis, D., Sep 2013. Simultaneous eeg and fmri reveals a causally connected subcortical-cortical network during reward anticipation. J Neurosci 33 (36), 14526–33.

Purcell, B., Heitz, R., Cohen, J., Schall, J., Logan, G., Palmeri, T., 2010. Neurally-constrained modeling of perceptual decision making. Psychological Review 117, 1113–1143.

Purcell, B., Schall, J., Logan, G., Palmeri, T., 2012. Gated stochastic accumulator model of visual search decisions in FEF. Journal of Neuroscience 32, 3433–3446.

Purcell, B. A., Palmeri, T. J., 2015. Relating accumulator model parameters and neural dynamics, manuscript under review, for this special issue.

Ratcliff, R., 1978. A theory of memory retrieval. Psychological Review 85, 59–108.

Ratcliff, R., Cherian, A., Segraves, M., 2003. A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of simple two-choice decisions. Journal of Neurophysiology 90, 1392–1407.

Ratcliff, R., Frank, M. J., 2012. Reinforcement-based decision making in corticostriatal circuits: Mutual constraints by neurocomputational and diffusion models. Neural Computation 24, 1186–1229.

Ratcliff, R., Hasegawa, Y. T., Hasegawa, Y. P., Childers, R., Smith, P. L., Segraves, M. A., 2011. Inhibition in superior colliculus neurons in a brightness discrimination task? Neural Computation 23, 1790–1820.

Ratcliff, R., Hasegawa, Y. T., Hasegawa, Y. P., Smith, P. L., Segraves, M. A., 2007. Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. Journal of Neurophysiology 97, 1756–1774.

Ratcliff, R., Philiastides, M. G., Sajda, P., 2009. Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. Proceedings of the National Academy of Sciences of the United States 106, 6539–6544.

Ratcliff, R., Smith, P. L., 2004. A comparison of sequential sampling models for two-choice reaction time. Psychological Review 111, 333–367.

Rescorla, R. A., Wagner, A. R., 1972. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A. H., Prokasy, W. F. (Eds.), Classical Conditioning II: Current Research and Theory. Appleton Crofts. New York, pp. 64–99.

Roberts, S., Pashler, H., 2000. How persuasive is a good fit? Psychological Review 107, 358–367.

Rosenblatt, M., 1961. Principles of neurodynamics. Spartan Books, Washington, DC.

Roxin, A., Ledberg, A., 2008. Neurobiological models of two-choice decision making can be reduced to a one-dimensional nonlinear diffusion equation. PLoS Computational Biology 4, e1000046.

Rumelhart, D. E., 1977. Toward an interactive model of reading. In: Dornic, S. (Ed.), Attention and performance. Erlbaum, Hillsdale, NJ, pp. 573–603.

Rumelhart, D. E., McClelland, J. L., 1982. An interactive activation model of context effects in letter perception: Part 2. the context enhancement effect and some tests and extensions of the model. Psychological Review 89, 60–94.

Rushworth, M. F. S., Mars, R. B., Summerfield, C., Feb 2009. General mechanisms for making decisions? Curr Opin Neurobiol 19 (1), 75–83.

Schall, J. D., 2004. On building a bridge between brain and behavior. Annual Review of Psychology 55, 23–50.

Shadlen, M. N., Newsome, W. T., 1996. Motion perception: Seeing and deciding. Proceedings of the National Academy of Sciences of the United States 93, 628–633.

Shadlen, M. N., Newsome, W. T., 2001. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. Journal of Neurophysiology 86, 1916–1936.

Shiffrin, R. M., Lee, M. D., Kim, W., Wagenmakers, E.-J., 2008. A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. Cognitive Science 32, 1248–1284.

Shiffrin, R. M., Nobel, P. A., 1997. The art of model development and testing. Behavior Research Methods, Instruments, & Computers 29, 6–14.

Smith, P. L., 2010. From Poisson shot noise to the integrated Ornstein-Uhlenbeck process: Neurally-principled models of diffusive evidence accumulation in decision-making and response time. Journal of Mathematical Psychology 54, 266–283.

Smith, P. L., McKenzie, C. R. L., 2011. Diffusive information accumulation by minimal recurrent neural models of decision-making. Neural Computation 23, 2000–2031.

Smolensky, P., 1988. On the proper treatment of connectionism. Behavioral and Brain Sciences 11, 1–23.

Summerfield, C., Koechlin, E., 2010. Economic value biases uncertain perceptual choices in the parietal and prefrontal cortices. Front Hum Neurosci 4, 208.

Teller, D. Y., 1984. Linking propositions. Vision Research 24, 1233–1246.

Teodorescu, A. R., Usher, M., 2013. Disentangling decision models – from independence to competition. Psychological Review 120, 1–38.

Tsetsos, K., Gao, G., McClelland, J. L., Usher, M., 2012. Using time-verying evidence to test models of decision dynamics: Bounded diffusion vs. the leaky competing accumulator model. Frontiers in Neuroscience 6, 1–17.

Tsetsos, K., Usher, M., McClelland, J. L., 2011. Testing multi-alternative decision models with non-stationary evidence. Frontiers in Neuroscience 5, 1–18.

Turner, B. M., 2015. Constraining cognitive abstractions through Bayesian modeling. In: Forstmann, B. U., Wagenmakers, E.-J. (Eds.), An introduction to model-based cognitive neuroscience. Springer, New York, pp. 199–220.

Turner, B. M., Dennis, S., Van Zandt, T., 2013a. Bayesian analysis of memory models. Psychological Review 120, 667–678.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., Steyvers, M., 2013b. A bayesian framework for simultaneously modeling neural and behavioral data. NeuroImage 72, 193–206.

Turner, B. M., Sederberg, P. B., 2012. Approximate Bayesian computation with Differential Evolution. Journal of Mathematical Psychology 56, 375–385.

Turner, B. M., Sederberg, P. B., 2014. A generalized, likelihood-free method for parameter estimation. Psychonomic Bulletin and Review 21, 227–250.

Turner, B. M., Sederberg, P. B., Brown, S., Steyvers, M., 2013c. A method for efficiently sampling from distributions with correlated dimensions. Psychological Methods 18, 368–384.

Turner, B. M., Sederberg, P. B., McClelland, J. L., 2015a. Bayesian analysis of simulation-based models, in Press.

Turner, B. M., Van Maanen, L., Forstmann, B. U., 2015b. Combining cognitive abstractions with neurophysiology: The neural drift diffusion model. Psychological Review 122, 312–336.

Turner, B. M., Van Zandt, T., 2012. A tutorial on approximate Bayesian computation. Journal of Mathematical Psychology 56, 69–85.

Turner, B. M., Van Zandt, T., 2014. Hierarchical approximate Bayesian computation. Psychometrika 79, 185–209.

Umakantha, A., Purcell, B., Palmeri, T., 2015. Mapping between a spiking neural network model and the diffusion model of perceptual decision making (working title), manuscript in preparation.

Usher, M., McClelland, J. L., 2001. On the time course of perceptual choice: The leaky competing accumulator model. Psychological Review 108, 550–592.

van Maanen, L., Brown, S. D., Eichele, T., Wagenmakers, E.-J., Ho, T., Serences, J., 2011. Neural correlates of trial-to-trial fluctuations in response caution. Journal of Neuroscience 31, 17488–17495.

van Maanen, L., Van Rijn, H., 2010. The locus of the Gratton effect in picture-word interference. TopiCS in Cognitive Science 2, 168–180.

van Maanen, L., Van Rijn, H., Borst, J. P., 2009. Stroop and picture-word interference are two sides of the same coin. Psychonomic Bulletin and Review 16, 987–999.

van Maanen, L., Van Rijn, H., Taatgen, N. A., 2012. RACE/A: An architectural account of the interactions between learning, task control, and retrieval dynamics. Cognitive Science 36, 62–101.

van Ravenzwaaij, D., van der Maas, H. L. J., Wagenmakers, E. J., 2012. Optimal decision making in neural inhibition models. Psychological Review 119, 201–215.

Vanpaemel, W., 2010. Prior sensitivity in theory testing: An apologia for the Bayes factor. Journal of Mathematical Psychology 54, 491–498.

Vanpaemel, W., 2011. Constructing informative model priors using hierarchical methods. Journal of Mathematical Psychology 55, 106–117.

Vanpaemel, W., Lee, M. D., 2012. Using priors to formalize theory: Optimal attention and the generalized context model. Psychonomic Bulletin and Review 19, 1047–1056.

Wagenmakers, E.-J., Farrell, S., Ratcliff, R., 2004. Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition. Psychonomic Bulletin and Review 11, 579–615.

Weissman, D. H., Roberts, K. C., Visscher, K. M., Woldorff, M. G., 2006. The neural bases of momentary lapses in attention. Nature Neuroscience 9, 971–978.

White, C. N., Congdon, E., Mumford, J. A., Karlsgodt, K. H., Sabb, F. W., Freimer, N. B., London, E. D., Cannon, T. D., Bilder, R. M., Poldrack, R. A., 2014. Decomposing decision components in the stop-signal task: A model-based approach to individual differences in inhibitory control. Journal of Cognitive Neuroscience 26, 1601–1614.

White, C. N., Poldrack, R. A., 2013. Using fMRI to constrain theories of cognition. Perspectives on Psychological Science 8, 79–83.

Wiecki, T. V., Poland, J., Frank, M. J., 2015. Model-based cognitive neuroscience approaches to computational psychiatry: Clustering and classification. Clinical Psychological Science 3.

Wong, K.-F., Wang, X.-J., 2006. A recurrent network mechanism of time integration in perceptual decisions. The Journal of Neuroscience 26, 1314–1328.

Woodman, G., Kang, M., Thompson, K., Schall, J., 2008. The effect of visual search efficiency on response preparation: Neurophysiological evidence for discrete flow. Psychological Science 19, 128–136.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., Cohen, A., 2003. Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. Journal of Experimental Psychology: Learning, Memory and Cognition 29, 1160–1173.

Zandbelt, B. B., Purcell, B. A., Palmeri, T. J., Logan, G. D., Schall, J. D., 2015. Response times from ensembles of accumulators. PNAS 111, 2848–2853.