In Proceedings of ICCM - 2007- Eighth International Conference on Cognitive Modeling. 285 - 290. Oxford, UK: Taylor & Francis/Psychology Press.

Queueing Network Modeling of Mental Architecture, Response Time, and Response Accuracy: Reflected Multidimensional Diffusions

Yili Liu (yililiu@umich.edu)
Department of Industrial and Operations Engineering, University of Michigan
1205 Beal Avenue, Ann Arbor, MI USA

Abstract

Response time (RT) and response accuracy are two of the most commonly used performance measures in cognitive psychology and studies of cognitive architecture. This paper examines the relationship and establishes a bridge between two currently separated groups of mathematical models of RT: models of RT and mental architecture and models of RT and accuracy. The bridge, called QN-RMD, is established by extending the queueing network (QN) architecture model of RT (Liu, 1996), which has successfully integrated a large number of RT-architecture models as special cases, and by representing the state changes in a mental QN as Reflected Multidimensional Diffusions (RMD). More specifically, the "state" of a K-server QN mental architecture is represented as a reflected diffusion space of K dimensions, in which "reflecting barriers" represent and reveal architectural constraints, while "absorbing barriers" represent accuracy-related response criteria, analogous to diffusion models of RT. This approach moves beyond the current 1-D diffusion models that have successfully accounted for but are limited to single-stage fast responses. 1-D diffusions can only represent the "state" of a single server system in stochastic information accumulation, not multi-server architectures. This approach extends the architectural RT models to account for accuracy, brings the diffusion/accumulator models to the architectural domain, and unifies RT/accuracy/mental architecture modeling in a larger framework.

Introduction

Response time (RT) is arguably the most commonly used performance measure in cognitive psychology research; it is regarded as a reflection of the dynamic activities of an underlying *mental architecture* that transforms stimulus into response; and it is known to have a close relationship with *response accuracy*.

The large majority of existing mathematical models of RT can be classified into two groups—models of RT and mental architecture and models of RT and response accuracy. The first group of models (called RT-architecture models in this paper) focuses on using RT to infer the possible temporal and architectural structures of the underlying mental system that transforms stimulus into response. This paper uses "architecture" to refer to "macro-architecture" of processing stages. "Micro-architecture" neural network models are important but beyond of the scope of this paper. The second group of models (the large majority of which belong to the family of sequential sampling or stochastic information accumulation models) focuses on modeling the relationship between RT and accuracy. Each group has made great progress in modeling the aspects of RT it focuses on. There is, however, a substantial gap between the two groups of RT models. The architectural models have not made great progress in revealing and modeling the intrinsic relationship between RT and accuracy, while the sequential sampling models have been relatively silent about the architecture of the cognitive "black/mystery box" in which the samplings (such as random walks or diffusions) occur.

This paper describes our research that (1) extends the queueing network (QN) architectural model of RT to cover accuracy; (2) establishes a natural link between the QN and the sequential sampling/diffusion models through a modeling approach called Reflected Multidimensional Diffusions (RMD); (3) develops QN and RMD methods to use RT and accuracy together for revealing mental architecture. In short, mental architecture is represented as a QN, whose state changes can be analyzed as a RMD. More specifically, the "state" of a K-server queueing network of mental architecture is represented as a reflected diffusion space of K dimensions, in which "reflecting barriers" represent and reveal architectural constraints, while "absorbing barriers" represent accuracy-related subject-adopted response criteria, similar to diffusion models. This approach moves beyond the current 1-D diffusion models that have successfully accounted for but are limited to single-stage fast responses. 1-D diffusions can only represent the state of information accumulation of a single server, not multi-server architectures.

Mathematical Models of RT and Mental Architecture

As shown on the left side of Figure 1, RT/architecture models have focused on two issues that are central to RT modeling and theory in cognitive psychology. One is a temporal dimension distinguishing discrete from continuous information transmission models, and the other is an architectural arrangement dimension distinguishing serial stage models from network models. All of the models assume that the psychological activity that transforms stimulus into response is composed of a system of mental processes. Discrete information transmission models assume that a mental process transmits its processing output in an indivisible unit and will not make its output available to other processes until it is completed. Therefore, a process cannot begin until all of its preceding processes are

Mathematical Models of RT and Mental Structure Classified
in terms of Discrete versus Continuous Information Transmission
and Serial versus Network Architecture

(from Liu, 1996, “Queueing network modeling of
elementary mental processes,” Psychological Review,
103(1), pp. 116-136).

Mathematical Models of RT
and Response Accuracy
(sequential sampling models)

	Architectural arrangement of mental processes			
Temporal Transmission	Serial Stages	Network Configurations		State Transitions
Discrete	Subtractive Additive factors General Gamma	Critical Path Network		Counter/accumulator Random-walk
Continuous	Cascade Queueing series	Queueing Network (QN)	↔	Accumulator Diffusion Reflected Multidimensional Diffusions (RMD for state of QN)

Figure 1: Mathematical Models of RT and Mental Architecture (left side) and Mathematical Models of RT and Response Accuracy (right side)

completed. Continuous information transmission models, in contrast, assume that each process transmits it s partial output to other processes continuously as soon as they are available rather than waiting for the full completion of processing, and thus a process can begin even though its preceding processes are still active. Serial stage models assume a serial arrangement of mental processes, whereas network models assume a network configuration. The two dimensions jointly define four classes of models as shown on the left side of Figure 1 (Liu, 1996). As described in detailed in Liu (1996), a class of queuing network models for RT and mental architecture was proposed which, in its most general form, represents continuous-transmission-network models and they include the existing models in the other three cells as special cases, and thus unify them in a larger modeling framework. Liu (1996) also reexamined the logic and conclusions of the previous models. It turns out that many of the conclusions based on the previous models are open to alternative explanations. All the QN models in Liu (1996) were discussed in relation to empirical data. Furthermore, it was shown that QN models allow us to cover a broader range of possible mental structures that mental system might assume but had not been modeled by previous models, such as feedback or non-unidirectional information flow, information “overtaking and bypassing”, and process dependencies or non-selective influence of factor effects, and can be subjected to well-defined empirical tests. The QN approach to RT modeling published in Liu (1996) focuses on the use of RT to infer mental architecture and is able to broaden the scope of thinking about the possible configurations of mental systems and the possible causes for certain RT phenomena. However, some important questions remained open including how the QN models deal with response accuracy and what their relation is to the sequential sampling models described below.

Mathematical Models of RT and Accuracy

The importance of examining RT and accuracy together in RT analysis and modeling has been emphasized by many researchers (e.g., Audley, 1960; Corbett and Wickelgren, 1977; Dosher, 1979; Meyer et al., 1988; Pachella, 1976; Pew, 1969; Ratcliff, 1978; Wickelgren, 1977). A crucial requirement is that they both arise naturally from common processing mechanisms (Ratcliff, 1978; 1985). The class of mathematical models that have achieved the greatest success in this regard is the class of sequential sampling (also called stochastic information accumulation) models, including random-walk models and related diffusion models, and counter or accumulator models. Sequential sampling models have been applied most extensively to model RT and accuracy data in choice RT experiments (e.g., Audley, 1960; Audley and Pike, 1965; Ashby, 1983; Edwards, 1965; Gronlund and Ratcliff, 1991; Heath, 1981; Laberge, 1962; Laming, 1968; Link and Heath, 1975; Pike, 1966; Ratcliff, 1978, 1981, 1985, 1988; Ratcliff and McKoon, 1982; Ratcliff and Rouder, 1998,2000; Ratcliff, Van Zandt, and McKoon, 1999; Stone, 1960; Van Zandt, Colonius, and Proctor, 2000; and Vickers, 1970). They have also been applied to model simple RT data patterns (e.g., Diederich, 1995; Schwarz, 1994; Smith, 1995), and recently, in modeling decision making (Aschenbrenner, Albert, and Schmalhofer, 1984; Busemeyer and Townsend, 1993; Busemeyer and Diederich, 2002; Diederich, 1995, 1997, 2003a, 2003b; Diederich and Busemeyer, 2003; Roe, Busemeyer, and Townsend, 2001) and classification (Ashby, 2000; Cohen and Nosofsky, 2003; Nosofsky and Palmeri, 1997). All sequential sampling models share the notion that the human information processing system accumulates information over time until a preset response criterion is reached and this accumulation process evolves stochastically.

The random walk models assume that in a two-choice response situation, the information accumulation process "randomly walks" in discrete steps between two decision boundaries (also called "absorbing barriers") based on the value of a *cumulative evidence variable*, each boundary representing one of the two choices. The process generally walks to the positive or the negative boundary depending on whether the value of the evidence variable is positive or negative. The time for the process to reach one of the two boundaries for the first time (immediately terminating the process) is called the *first passage time*, which determines RT. The probability that the process terminates at one or the other boundary is called *first passage probability*, which determines the probability of the associated response. The continuous versions of the random walk models are called diffusion models, which assume that the corresponding stochastic process drifts continuously toward the positive or the negative boundary, depending on whether the mean rate of information accumulation is positive or negative.

The term "counter models" has been used to refer to models that assume discrete counting increments, while "accumulator models" are used broadly to refer to both discrete and continuous evidence accumulation. The idea of using a counter to model RT can be traced back to McGill (1963, 1967), Laberge (1962), and Audley and Pike (1965). Usher and McClelland's (2001) leaky, competing accumulator model represents the state of the art in accumulator modeling.

These sequential sampling models are very successful in modeling RT-accuracy relations for single stage fast binary responses, but relatively silent on multistage architecture issues. The challenge is to bridge the gap between the two groups of models summarized above.

Queueing Network Modeling of Mental Architecture, RT, and Accuracy: Reflected Multidimensional Diffusions

The QN architecture model of RT presented in Liu (1996) adopts the following assumptions similar to those commonly made in the RT literature: A stimulus is composed of several types of stimulus components (called customers), who arrive at various nodes of the processing network to request for service and that the sequence of customer arrival times, the sequence of customer service times, and the sequence of customer departure times are all stochastic processes. Presently, similar to all major RT models, QN models for RT assume that there is a separate response unit at the end of the processing network (after the "last" or the "exit" node), which is responsible for the actual response. QN models assume that a response is made when the response unit has accumulated N signal components, delivered from the "exit" node. RT is defined and determined by the network sojourn time of the Nth signal customer who completes all its network service requests and departs from the network.

To extend the 1996 QN-RT model to cover response accuracy, the QN-RMD research makes two extensions to the 1996 QN definition of RT: First, we assume that RT is defined and determined by the Nth "response activating" customer (rather than solely by the Nth "signal customer" in the 1996 model, which only elicits a correct response). In a binary RT task (e.g., Yes or No), the Nth response activating customer is the Nth Yes customer for a Yes response or the Nth No customer for a No response. In a RT task involving K alternatives, the Nth response activating customer refers to the Nth i-type customer for a trial with an i-type response. Second, we treat N as a parameter that is analogous to the setting of "counts" in accumulator models and the setting of boundary positions in random walk or diffusion models. A larger N is analogous to a higher preset count or wider boundary. The largest useful N could mean the "limit on the number of useful observations" (Swensson, 1972; p. 30; also in Usher and McClelland, 2001; p. 551-552).

In QN-RMD, the relationship between RT, accuracy, and mental architecture is studied by analyzing the departure process at the network exit node (again, its Nth departure is the Nth accumulation at the "dummy" response node) and examining how this departure process is affected by network architecture and subject-adopted response criterion. We adopt the common assumption that the departure process at each QN node i, $D_i(t)$, is a continuous stochastic process that has independent and stationary increments. Mathematically, this is equivalent to assume, $D_i(t_1) – D_i(t_0)$, …, $D_i(t_n) – D_i(t_{n-1})$ are independent for any $n\geq 1$ and $0\leq t_0\leq ...\leq t_n\leq\infty$ and the distribution of $D_i(t) – D_i(s)$ depends only on t-s, for all i. This assumption is the most commonly made in the QN literature. Formal mathematical limit theorems that justify the use of these assumptions for analyzing various types of queueing systems and thus the use of Brownian approximations have been proved for various types of flow system models (see e.g., Chen, 1996; Williams, 1998). The K-vector departure process of a K-server queueing network of mental architecture, $\boldsymbol{D}(t) = [D_i(t), i=1,…, K]$, can be represented as a Reflected Diffusion in K dimensions, in which "reflecting barriers" represent and reveal architectural constraints, while "absorbing barriers" represent accuracy-related response criteria. Informally, "reflecting barriers" define the space in which the diffusions can occur. Diffusion occurs "normally," in the interior of the space, but is instantaneously "pushed back" into the space when it hits one of the reflecting barriers. The reason "reflecting barriers" exist for QN is because departure processes cannot take on any arbitrary values. For example, in a 2-node serial QN system, if K1 is in front of K2, we must have $D_1(t) \geq D_2(t)$ (all departures at K2 must have departed from K1 first). Thus, the diffusion of the 2-vectored process $[D_1(t), D_2(t)]$ can only occur in the region in which $D_1(t) \geq D_2(t)$ (i.e., a reflecting barrier exists at $D_1(t) = D_2(t)$, with a reflecting direction pointing toward the allowed region). Reflecting barriers can be defined similarly if K2 is in front of K1, or for other situations. Thus reflecting barriers reveal the architectural arrangement of the mental system. "Absorbing barriers" are defined in the same sense as current diffusion models of RT, such as those of Ratcliff. Due to space limit, this paper chooses two representative cases as illustrations.

1. Single–server QN for binary responses

The simplest case of a QN is a single server system. The binary response case called binary, single-step responses, has been modeled most extensively and successfully by existing diffusion RT models. "The diffusion model was designed to explain fast, *single step*, as opposed to multistep, decision processes, …" (Ratcliff, et al., 1999; p. 262.). This case serves two purposes: important by itself to show a concrete link between QN and RW/diffusion models, and as the base or starting point for modeling more complex architectures. In this QN, binary response (Yes/No) are triggered by two types of customers, A and B, who arrive at the server with arrival rates of λ_a and λ_b, respectively, in accordance with a renewal process having an arbitrary interarrival distribution and are served by the server with service rates of μ_a and μ_b, respectively.

This method assumes that each type-A customer departing from the server carries an information amount of +1, while each type-B departing customer carries -1. This assumption is similar to, e.g., the classical RT diffusion models of Ratcliff et al., (1999) and the "two-barrier single channel model" of Smith (2000). In the following I show how this single server QN is mathematically identical to the classical RT diffusion models, but it offers a QN interpretation to it. Let S_n denote the total amount of information carried by the first N departing customers. We have, $S_n = \sum_{i=1}^{N} X_i$, $i \geq 1$; where $X_i=1$ for a departing customer of type A and $Xi = -1$ for a departing customer of type B.

Clearly this departure process S_n is a random walk (RW), and its relation to existing RW models of binary RT becomes at least intuitively apparent. If we speed up the departure process by considering smaller and smaller time intervals and letting Δt go to 0, then the total amount of information carried by all departed customers by time t, D(t), follows a diffusion process. In fact, this comes naturally also as a consequence of our general assumption of independent stationary increments mentioned earlier. In the queueing literature it is commonly regarded as a harmless assumption to treat fast discrete customer departures as a continuous flow (see, e.g., Harrison, 1985).Mathematically, D(t) can be characterized with the Kolmogorov backward equation, as follows:

$$\frac{\partial}{\partial t}p(t,x,y) = (\frac{1}{2}\sigma^2\frac{\partial^2}{\partial x^2} + \mu\frac{\partial}{\partial x})p(t,x,y)$$

where p(*t*, *x*, *y*) is the transition density, *x* is the starting state, *y* is the ending state, of time period *t*. This equation is called the Kolmogorov backward equation because the differentiation is with respect to the backward variable (the initial state) *x* on the right side of the equation above. This equation is identical to Ratcliff et al's (1999, p.299) diffusion equation, with difference only in the notations.

When $\mu \neq 0$ (diffusion with a drift), then as shown in Harrison (1985) in his analysis of diffusion approximation of stochastic flows in queueing systems, we have, $P_x\{X_t=b\}$, the probability that the process first crosses the barrier b before crossing the barrier at 0, when the starting position is at *x*, as $P_x\{X_t = b\} = \frac{1-\xi(x)}{1-\xi(b)}$, $0 \leq x \leq b$,

where $\xi(z) \equiv \exp(\frac{-2\mu z}{\sigma^2})$

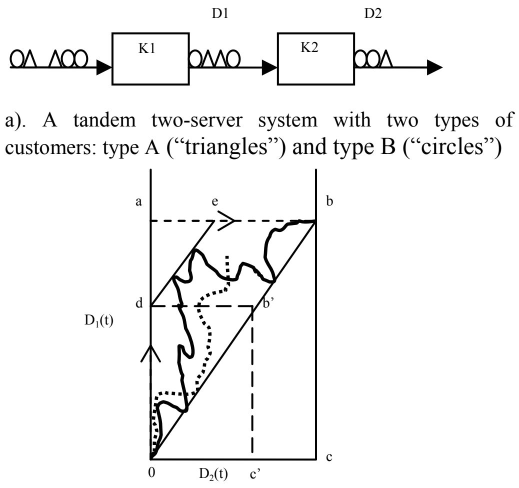
This result is the same as that in Ratcliff et al (1999, p.299), who presented $P_x\{X_t=0\}$, the probability that the process first crosses the absorbing barrier 0 before crossing the barrier at b: $Px\{Xt=0\} = 1 - Px\{Xt=b\} = 1-\frac{1-\xi(x)}{1-\xi(b)} = \frac{\xi(b)-\xi(x)}{\xi(b)-1}$

This is identical to Ratcliff et al (1999, p299), with the difference found only in the symbol notations.

The convergence of the results of the queueing literature and the RT-diffusion models shown above offers a queueing architectural explanation to the RT diffusion modeling. Since our diffusion representation of the departure process of the single-server queueing system has converged precisely with the diffusion model of Ratcliff, all the related results of Ratcliff apply. One intuitive interpretation of the diffusion parameters with the QN parameters is: The mean drift rate, μ, is determined by the difference between the two mean arrival rates (λa-λb). This is intuitive, since an RT trial with stimulus A would carry more customers (features) of A, thus produce a greater arrival rate of type A customers than a B-type RT trial. This is similar in spirit to, e.g., Ratcliff's (1978) work of using stimulus relatedness to decide drift rate. The boundary positions can be interpreted as the minimum amount of total positive or negative information carried by all the departed customer to elicit an A or B response, respectively (as in Ratcliff, we will also use b and 0 as the two boundaries). The starting point is the subject's response bias, which can be assumed as "preloaded departures" or "information preloaded in the system"—thus called a "bias." A discussion of the relationship between this single server 1-D diffusion model and the corresponding accumulator model can be found in Liu (2005). Due to space limit, we elect to discuss 2-D cases next to illustrate how to consider architecture issues in this QN-RMD framework.

2. A tandem 2-Server QN for binary responses

A basic, fundamental, and illustrative case involving all three issues: RT, accuracy, and architecture is a tandem 2-server QN as shown in Figure 2, which goes beyond single-stage RT-accuracy modeling and demonstrates the importance of considering "reflecting" barriers, in addition to "absorbing" barriers of the conventional 1-D diffusion models.



a). A tandem two-server system with two types of customers: type A ("triangles") and type B ("circles")

b). For customers' departure processes $D_1(t)$ and $D_2(t)$, if K1 and K2 form a discrete processing series, then we have two 1-d diffusions in a row, first along the border from 0 to a, then along the border from a to b. If K1 and K2 form a continuous-flow series, then we have one 2-d reflected diffusion. See text for more details.

Figure 2 A tandem 2-Server QN and its Reflected Diffusion Space

In this tandem 2-server QN shown in Figure 2, we consider a pair of 2-vectored departure processes, $\{D_{1A}(t), D_{2A}(t)\}$ and $\{D_{1B}(t), D_{2B}(t)\}$, corresponding to the type-A and type-B departures from K1 and K2, respectively. Similar to the diffusion cases for a single server, we assume each customer departing at K2 carries the same amount of information to contribute to its type of response only. A response is made when $D_{2A}(t)$ or $D_{2B}(t)$ first reaches its criterion value. Let us consider two cases, corresponding to the debates between discrete and continuous information transmission between stages.

Case A. A series of 2 discrete processing stages corresponds to a sequence of 2 1-d diffusions

The mental architecture theories of discrete processing series (e.g., Donders' and Sternberg's theories, upper-left cell of Figure 1) assume that K1 must complete all its work before K2 can start. In other words, all departures from K1 must complete before K2 starts its departure process, i.e., $D_1(t)$ must complete before $D_2(t)$ starts. In Figure 2, this can be visualized as a sequence of two 1-d diffusions for each type of customers, first along the line from 0 to a for $D_1(t)$, and then along the line from a to b for $D_2(t)$. For the simplest case of "no bias" RT tasks, both types of customers go through the same diffusion sequence (first 0-to-a, and then a-to-b).

Case B. A continuous-flow 2-server queue-series corresponds to a reflected 2-d diffusion.

The mental architecture assumption of continuous flow (e.g., McClelland's Cascade, Miller's and Liu's queue series, lower-left cell of Figure 1) does not require K1 to complete its processing before K2 can start. In other words, $D_1(t)$ and $D_2(t)$ may occur concurrently, subject to certain constraints in a queueing system. Specifically, the joint distribution of $D_1(t)$ and $D_2(t)$ can be characterized as a reflected 2-d diffusion: If K1 is in front of K2, then diffusion occurs in the upper region above the reflecting barrier shown as the diagonal line (0—b) in panel b of Figure 2. If K2 has finite waiting space, s, (in front of K2), then diffusion is further bounded by a reflecting barrier shown as line (d—e), whose vertical distance to line 0—b is s. A reflecting barrier (line a—b) exists if D_1 has an upper limit. Absorbing barrier is shown as line b—c. For simplicity of presentation and analysis, we continue to focus on the no-bias RT situation for now, meaning that the two types of customers "race in the same diffusion space," shown as a pair of trajectories (a solid and a dotted curve, where the solid one wins in this illustration) in panel b of Figure 2. Several testable performance predictions can be made with regard to RT-accuracy relation in this situation, including:

1). When K1's service rate is much larger than K2 (e.g., K1 is a super fast perceptual server) and K2 has unlimited waiting space (i.e., $s=\infty$), then the probability is almost 1 that D_1 is much greater than D_2, thus the probability of hitting the reflecting barrier at (0—b) is almost 0. In this situation, in terms of its effect on RT/accuracy, the 2-d diffusion would behave as if it is a 1-d diffusion of $D_2(t)$ along the line of (0—c), similar to the classical diffusion RT models and to the single server case discussed earlier. Informally, if K1 is so powerful, we don't have to worry about it; we just need to consider K2.

2). When K1 is not a super fast server or when an experimental factor increases the sojourn time at K1 (thus decreases the departure rate at K1), RT/accuracy will not show the same type of exponential relationship predicted by classical RT-diffusion models or single-server QN models, since now we have a true 2-d reflected diffusion, bound by the reflecting barriers, whose effect can not be ignored. Quantitatively, it would take a longer time to achieve the same level of accuracy obtained in the 1-d diffusion case.

3). Further, when there is a very low limit in the departure process of K1 (e.g., in the so-called data-limited tasks in which impoverished or significantly degraded stimuli are used), RT/accuracy will not show the same type of exponential relationship predicted by classical RT-diffusion models or single-server QN models. This can be visualized in panel b of Figure 2: when a is smaller than b (slide the line a-b downward to, say, d-b'), the 2-d diffusion will never be able to reach the absorbing barrier b-c (i.e., subject never responds) UNLESS the subject reduces the absorbing barrier, by moving it to the left to b'c', by willing to make less accurate responses. This offers an alternative explanation to the classical "infinite-RT" problem (Ashby, 1982).

4). The order of server arrangement is an architectural research question itself. The analysis above assumes K1 is in front of K2. When K2 is in fact in front of K1, diffusions would occur in the lower region. Thus, a method of using RT/accuracy together to reveal the order of K1 and K2 is to see whether an upper- or a lower-region diffusion best fits the data.

3. Other network cases

The single-server and the 2-server QNs and their diffusion representations described above are the simplest network cases. Concrete testable predictions can also be made about more complex network arrangements. Two examples are listed below:

1). A series of K discrete processing stages correspond to a series of K 1-d diffusions. This is an extension of Case A of the 2-server QN to a general series of discrete stages.

2). A series of K continuous flow servers correspond to a reflected K-d diffusion with a "lower-triangular reflection matrix." Characteristics of this reflection matrix reveal the layout of the series. This is an extension of Case B of the 2-server QN to a general queueing series.

Additional network cases and related discussions can be found in Liu (2005).

In summary, the QN-RMD (Queuing Network-Reflected Multidimensional Diffusions) represents mental architecture as a QN, whose state of operation can be represented as a multidimensional diffusion space. QN-RMD extends the QN architectural RT models to account for accuracy, brings the Random Walk/diffusion models of RT and accuracy to the multi-server architectural domain, and unifies the two currently separated schools of approaches in a larger framework. The work helps reduce the "fragmentary nature of the results" (Luce, 1986, p. 491), by "synthesizing what we know" (Newell, 1990; p. 16).

Sample References

Audley, R. J., and Pike, A. R. (1965). Some stochastic models of choice. *British Journal of Mathematical and Statistical Psychology*, 18, 207-225.

Busemeyer, J. R., and Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decisions making in an uncertain environment. *Psychological Review*, 100(3), 432-459.

Cohen, A. L., and Nosofsky, R. M. (2003). An extension of the exemplar-based random-walk model to separable-dimension stimuli. *Journal of Mathematical Psychology*, 47, 150-165.

Dai, J.G., and Harrison, J. M. (1992). Reflected Brownian motion in an orthant: Numerical methods for steady-state analysis, *Annals of Applied Probability*, 2(1), 65-86.

Harrison, J. M. (1985). *Brownian motion and stochastic flow systems*. New York: Wiley.

Liu, Y. (1996). Queueing network modeling of elementary mental processes. *Psychological Review*, 103(1), 116-136.

Liu, Y. (2005). *Queuing network modeling of mental architecture, response time, and response accuracy: Reflected multidimensional diffusions.* Tech Report 05-11 of the Dept of Industrial and Operations Engineering, University of Michigan.

Luce, R. (1986). *Response Times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

Ratcliff, R. (1985). Theoretical interpretations of speed and accuracy of positive and negative responses. *Psychological Review*, 92, 212-225.

Ratcliff, R., and McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psychological Review*, 104, 319-343.

Ratcliff, R., Van Zandt, T., and McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, 106(2), 261-300.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology,* 44, 408-463.

Townsend, J. T., & Ashby, F. (1983). *The Stochastic Modeling of Elementary Psychological Processes.* Cambridge: Cambridge University Press.

Usher, M., and McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review,* 108(3), 550-592.