Simulating Overall and Trial-by-Trial Effects in Response Selection with a Biologically-plausible Connectionist Network

Blair C. Armstrong (b.armstrong@bcbl.eu)
Basque Center on Cognition, Brain, and Language
Paseo Mikeletegi 69, San Sebastian, 20009 Spain

David C. Plaut (plaut@cmu.edu)
Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Abstract

Ratcliff, Van Zandt, and McKoon (1999, *Psych. Rev.*) claim that connectionist models fail to simulate many aspects of how individuals select one of two possible responses. Here, these claims are re-evaluated via computational and behavioral investigations of an extended version of the original numerosity judgment task. The results of the experiment indicate that some of the empirical effects that the models failed to capture do not generalize and were likely due to idiosyncratic aspects of the original methodology. The simulations show that a more biologically-plausible model captures the bulk of the new effects, including some trial-by-trial adaptive effects that are outside the scope of models tested against aggregate data, and emergent asymptotic stability that has previously required an explicit leak parameter.

Keywords: response selection, decision making, connectionism, numerosity judgment, overall and trial-by-trial effects

Understanding how one of multiple candidate responses is selected in a given task is a long-standing and critical issue in cognitive science, and is one of the earliest domains to have been investigated with computational models. To date, much of the work has focused on the sub-issue of how individuals perform in tasks in which they must rapidly select one of two possible responses (i.e., speeded two-alternative forced-choice tasks; 2AFC tasks). This has led to the development of several models that can be fit to data from 2AFC tasks with a high degree of precision (e.g., the diffusion model; Ratcliff, 1978). One of several key limitations of these models, however, is that they are highly domain-specific and are not naturally extendable to studying other intuitively related issues, such as 'closed-set' response selection tasks involving three or more pre-specified candidate responses, or 'open-set' response selection tasks which require the production of novel responses such as nonword naming. These models are also often fit to aggregated data and do not explain how the decision system adapts over time based on its past experiences.

One possible avenue for addressing these limitations is the development of a connectionist model of response selection, given the connectionist framework's grounding in domain-general learning, representation, and processing principles that are drawn from systems and cellular neuroscience. Not only might such a model be able to explain the overall and adaptive effects in 2AFC tasks, it should also be readily extendable to the other response tasks described previously. Moreover, insofar as connectionist models fail in these endeavors, this can serve to guide the development of improved principles which can, by virtue of the domain-general nature of the framework, have wide-spread implications for domains well beyond response selection (e.g., semantic cognition).

Past work by Ratcliff, Van Zandt, and McKoon (1999) provides some initial insight into the performance of connectionist models of 2AFC tasks relative to that of the diffusion model in simulating performance in a numerosity judgment task. In this task, participants were presented with a 10×10 array which was filled with a number of asterisks sampled from two overlapping distributions with 'low' and 'high' mean numbers of asterisks, and made responses indicating which distribution they believed had been sampled from to generate the stimulus. The model comparisons revealed that the connectionist models failed to capture important aspects of the behavioral data (e.g., latency-accuracy functions, trial-by-trial adaptive effects).

To address some of these limitations, Usher and McClelland (2001) introduced a revised connectionist formalism in the leaky, competing accumulator model. Changes in this model included explicit constraints on the sign of the weights between competing units and from the underlying source of evidence that drives the response units, and the use of a threshold-linear activation function that is not differentiable at all points in time. A critical implication of the latter change is that it violates the mathematical principles that underlie standard gradient descent learning algorithms such as back-propagation (Hinton, 1989). Collectively, these changes rendered the accumulator functionally analogous to the diffusion model, and generally showed identical or superior fits to that model. This notwithstanding, a fundamental issue with this type of domain-specific connectionist model is what strengths of the standard connectionist framework were given up during model development. In particular, the disconnect between these models and standard connectionist learning algorithms prevents these models from being effortlessly extended to other response selection tasks—let alone cognitive processing and learning in other domains.

An alternative approach to developing improved connectionist models of response selection is to focus, instead, on improving the domain-general assumptions of the framework. One way to do this that is independent of the particular constraints needed to simulate response selection is to more accurately instantiate the known connectivity and processing characteristics of the brain. For instance, neurons are either excitatory or inhibitory—not both, as is the case in standard connectionist models. There are also more excitatory than inhibitory neurons, which biases the type of information that can be encoded by each sub-population: inhibitory neurons serve primarily to regulate overall activation in the information-content-rich excitatory neurons. Connections between brain regions are also typically only excitatory and relatively sparse (or functionally weak), whereas there is dense (or functionally strong) connectivity among both excitatory and inhibitory neurons within a brain region. The activation dynamics of individual neurons are also better approximated by activation functions that do not possess an upper non-linearity as is the case for sigmoidal functions (Usher & McClelland, 2001), but that are nevertheless differentiable at all points in time, such as the hybrid sigmoid-linear activation function that is presented in detail later.

To date, models that are constrained by the aforementioned characteristics of systems and cellular neuroscience have been found to capture a wide range of empirical effects such as the temporal dynamics of ambiguous word comprehension (Armstrong, 2012) and the ERP correlates of word and nonword processing and of behavioral lexical decision (Laszlo & Plaut, 2012). The present work extends these investigations by evaluating whether a more biologically-plausible connectionist model can simulate overall and adaptive effects in a simple perceptual task, without abandoning key principles such as learning or adopting ad hoc connectivity constraints.

Prior to the computational work, however, an appropriate set of benchmark data must be identified. Ratcliff et al. (1999) argued that their numerosity judgment task data were representative of results of many tasks and could therefore be treated as a gold standard for model comparison. However, detailed inspection of their results suggests that idiosyncratic and esoteric aspects of their methods may have led to atypical results. For instance, participants were potentially able to adopt sophisticated response strategies beyond those that are incorporated into simple models of response selection (notwithstanding that ultimately, more complex models should account for these data). For instance, participants received extensive experience with the task (≈ 12,000 trials). This may have interacted with the fact that participants were also explicitly told that the 'low' and 'high' distributions overlapped and therefore that it would not always be beneficial to adjust their performance following responses that were labeled as 'incorrect' (in contrast to the definitional behavior of error-driven learning and to the behavior observed in Armstrong, Joordens, & Plaut, 2009). To explore adaptation in the decision system, Ratcliff and colleagues also repeatedly and somewhat predictably manipulated the likelihood of sampling from the 'low' and 'high' distributions (≈ 30 times)—effectively changing the optimal threshold for making 'low' or 'high' responses—which may have allowed participants to utilize sophisticated cognitive control mechanisms. Consequently, we first investigated the representativeness of the original data in an extension of the original study.

Numerosity Judgment Experiment

The experiment assessed the generality of the empirical findings reported by Ratcliff et al. (1999) using a slightly modified version of the original methodology. This involved (a) splitting the distribution of number of asterisks using a fixed threshold rather than two overlapping 'low' and 'high' distributions, so that the feedback provided in the task was perfectly accurate within a mega-block (400 trials split into multiple blocks) and could, in principle, provide a basis for learning to respond perfectly, (b) eliminating or minimizing participants' prior experience with potential changes in the threshold value across mega-blocks that would allow them to develop sophisticated response strategies by changing the threshold value at most two times, and (c) only presenting participants with 3 mega-blocks totaling 1,200 trials (vs. 12,000 in the original study, of which the first 1,200 were dropped) to study initial learning when adaptive effects may be larger and more readily detectable. Insofar as these modifications produce divergent results, the data from this experiment may be a superior gold standard for model evaluation.

Methods

Participants. A total of 121 right-handed Carnegie Mellon undergraduates participated in exchange for course credit.

Apparatus. The experiment was implemented using PsychoPy 1.71.01 (Peirce, 2007). Responses were recorded on standard computer keyboard using the 'z' and '/' keys.

Stimuli and design. The stimuli used in the experiment consisted of a variable number of asterisks in a 10 × 10 grid. The number of asterisks was sampled from a trimmed normal distribution (mean = 50, SD = 14, min = 28, max = 72). These stimuli were divided into 'low' and 'high' categories on the basis of whether the number of asterisks fell above or below a threshold value. The threshold that delineated a 'low' response from a 'high' response could be either 4.5 points below or above the mean of the distribution in a given mega-block. These parameters are similar to those employed in the original experiment, with the critical difference that there was no overlap between the 'low' and 'high' distributions.

Participants were presented with 10 practice trials followed by three mega-blocks of 400 experimental stimuli. The threshold for making 'low' or 'high' responses could potentially change across each of these blocks. The full set of 16 different combinations of thresholds were run across the three mega-blocks (e.g., 'low/low/low' vs. 'low/low/high' vs. 'low/high/low,' etc.). Preliminary analyses indicated that the data could be grouped based on whether the threshold remained the 'same' between adjacent blocks or 'changed' between adjacent blocks. This allowed the critical number of conditions to be reduced to four for the analyses (same–same, same–change, change–same, change–change). Trials were presented across 25 blocks of 50 trials, except for the first and last blocks, which contained 25 trials.

The frequency with which each number of asterisks would be presented was a multiple of the probability density function. The stimuli were divided into 'low' and 'high' groups on the basis of the response threshold for the mega-block. Given the threshold levels used in the experiment, 72% (36% × 2) of trials fell on the tails of the distribution and were always either 'low' or 'high.' The correct response for the remaining 28% of trials in the center of the distribution depended on the threshold for the mega-block. The positioning of the asterisks within the array and the order of the stimuli were random, with the constraint that no more than five of the same type of stimuli could be presented sequentially.

Procedure. Participants were instructed that they would have to decide whether the number of presented asterisks was either 'low' or 'high.' They would have to learn what constituted a 'low' or 'high' number of asterisks by making responses and learning from the feedback that was provided. Note that in contrast to Ratcliff et al. (1999), participants were not instructed that a given number of asterisks could be produced by either the 'low' or 'high' distributions (because the feedback in the present experiment was accurate within each mega-block), nor were they informed that the threshold that delineated between the 'low' and 'high' distribution might change during the experiment. This was predicted to increase the likelihood that participants would adapt to the changes in the characteristics of the stimuli using simple statistical learning mechanisms based on the feedback that was provided.

Participants were instructed to respond as quickly as they could without making many mistakes. To operationalize this instruction, after each block participants received a message to "try to go faster, even if it means making a few more errors", or to "try to make a few less errors even if it means slowing down." The message that a participant received depended on whether their accuracy was above or below 90%, although this was not known to them. Following the instructions, participants were presented with the practice trials followed by the 25 experimental blocks.

Each trial consisted of (1) a fixation stimulus (+) for 500–700 ms, (2) a blank screen for 50 ms, and (3) a number of asterisks, which remained on the screen until participants responded or for a maximum of 5000 ms. If the response was incorrect (4) "INCORRECT" appeared on the screen for 400 ms. The next trial began automatically.

Results

Initial proficiency. The instructions alone were sufficient for participants to configure their response system in line with the general demands of the task, as assessed via a binomial test on the accuracy of the first practice trial relative to chance (mean accuracy = 63% SE = 0.05, $p = 0.008$, $n = 104$).

Overall performance for the same–same condition. Overall performance in the experiment was slightly less accurate and slower than that reported by Ratcliff et al. (1999). This notwithstanding, the qualitative similarity between the studies on a number of metrics was quite high, including accuracy and latency as a function of distance from the response threshold, the latency distribution for correct and incorrect responses, and the latency-accuracy functions. Figures 1 and 2 plot a subset of these data (the omitted figures, which are not included because of space constraints, showed similar qualitative matches to the model data). Note that because fewer stimuli were presented as distance increased from the response threshold, the data from both the experiment and the simulation becomes increasingly unreliable as distance increases (particularly for incorrect response latency), so later comparisons between the model and the simulation focus on distance from the threshold of 10 asterisks or less.

Sequential effects for the same–same condition. Sequential effects in the same–same condition were examined as a function of the number of blocks for stimuli of different distances from the threshold. In contrast to Ratcliff et al. (1999), the data plotted in Figure 3 showed a continuous decrease in latency as a function of practice, particularly for large distances from the threshold. Similar trends were observed in the accuracy data (not shown), although performance reached an asymptotic state within the first five blocks.

Mixed-effect regression models (Baayen, Davidson, & Bates, 2008) were used to further explore the effects of a number of characteristics of the preceding trial on the current trial's accuracy and latency. Due to space constraints, only the effects of previous trial accuracy, stimulus type, and response are reported. Significant effects have p-values less than .05. For the dependent measure of accuracy, previous trial accuracy did not predict significant variance, repetitions of the same stimulus increased accuracy, and repetitions of the same response decreased accuracy. For the dependent measure of correct latency, a previous accurate response and a repetition of the same response both decreased latency, and there was no effect of stimulus repetition. Additionally, the effects of prior accuracy decreased as a function of practice.

Adaptive effects following threshold changes. Figure 4 plots correct latency as a function of trial number for different numbers of asterisks and combinations of constant or changing response thresholds across mega-blocks (participant's accuracy data, not shown, showed similar dynamics). Three groups of asterisks are presented: one fell just below the initial low threshold (37–45), one fell just above the initial low threshold and just below a later high threshold in conditions in which a threshold change occurred later (46–54), and one was well above the initial low threshold but was immediately above a high threshold if the threshold value changed (55–63). The results indicated that in both the accuracy and the latency data the adaptation that followed a threshold change occurred over an extended number of trials. Specifically, performance generally did not approach an asymptotic level until approximately 100 trials after the threshold change.

Discussion

Despite the methodological similarities, many of the effects observed in this experiment diverge from those reported by Ratcliff et al. (1999). Moreover, the analyses of early performance reported here question some of the modeling as-

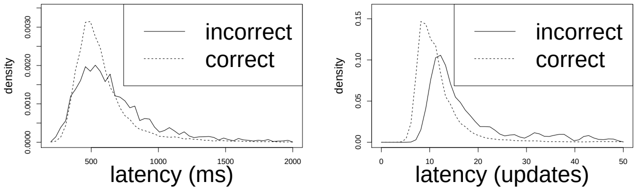

Figure 1: Latency distributions. [Left] Expt. [Right] Sim.

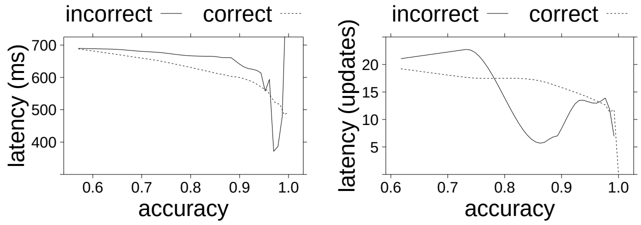

Figure 2: Latency as a function of accuracy.

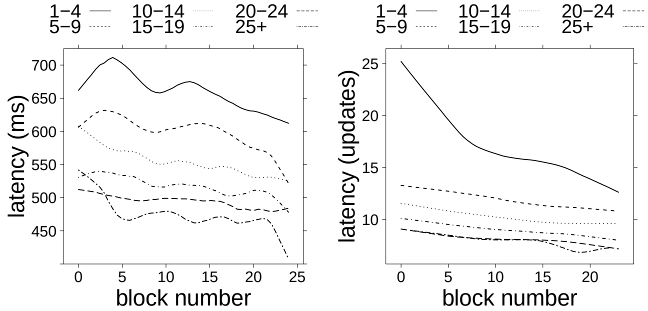

Figure 3: Correct latency as a function of block number for bins of asterisks of varying distance from the threshold.

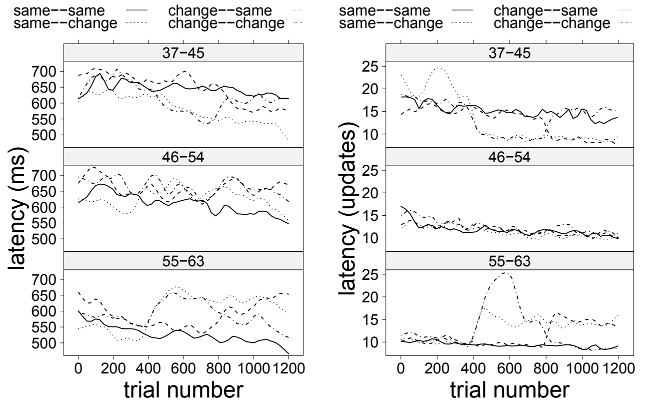

Figure 4: Correct latency as a function of trial number for the different conditions.

sumptions reported in that paper that could have contributed to the poor performance of the connectionist models that were tested. For instance, the above-chance initial accuracy during the practice blocks and the consistent performance improvements throughout the experiment suggest that a connectionist model should begin simulating the task with a basic level of proficiency which continues to improve with practice. This contrasts with the simulations conducted by Ratcliff and colleagues, for which either a trained model for which learning had been disabled or an untrained network for which learning was enabled were assessed.

Similarly, the significant effects of the accuracy of the previous trial on the accuracy and the latency of the subsequent trial that were observed in the experimented reported here were not in line with those reported by Ratcliff et al. (1999). These effects are, however, in line with the expected behavior that would result from using an error-driven learning algorithm and with the results observed in other studies (Armstrong et al., 2009). One possible reason for the discrepancy is that the participants used their explicit knowledge of the inconsistent nature of the feedback provided in the original experiment to develop adaptation strategies that override the effects of simple error-driven learning algorithms (but that may be captured by more sophisticated algorithms that do consider such factors; Hinton, 1989).

The adaptive effects observed in the present experiment following the change in the threshold also differed from those reported by Ratcliff et al. (1999). In particular, the rate of adaptation following a threshold change was relatively slow and approximately 100 additional trials were necessary to reach a new asymptotic level of performance. this contrasts with the results of the original experiment, which showed that the new asymptotic level was reached in an order of magnitude fewer trials (5-15). This discrepancy is likely due to participants' extensive experience with threshold changes at semi-predictable intervals in the original task.

Taken together, these discrepancies undermine prior claims about the representativeness of the original task and data as a gold standard for model comparison, and suggest that the present data are a more appropriate gold standard for evaluating the performance of simple models of response selection.

Numerosity Judgment Simulation

The simulation work evaluates whether a more biologically-plausible connectionist model produces the same patterns of effects that were observed in the experiment.

Methods

Participants. Two simulated participants completed each of the main conditions in the experiment (same–same, same–change, change–same, change–change).

Network architecture. The model architecture, based on the biologically-plausible connectivity principles described in the introduction, is presented in Figure 5. The visual inputs were divided into two groups of 100 units, the first of which coded for the presence of an asterisk in a given location, whereas the second coded for the absence of an asterisk in a particular location. This normalizes the overall amount of activity in a similar fashion to on-center/off-surround and off-center/on-surround visual neurons. One response unit coded for 'low' responses and the other for 'high' responses.

Arrows indicate full connectivity from one pool of units to another, with the exception that units were not connected to themselves. Outgoing connections from excitatory units were constrained to be positive and were initialized to a mean value of 0.15. Outgoing connections from inhibitory units

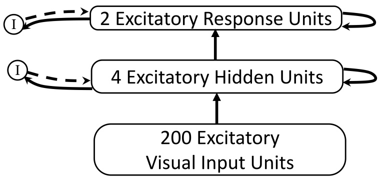

Figure 5: Network Architecture. I = Inhibitory unit. Solid arrows = excitatory connections. Dashed arrows = Inhibitory connections.

were constrained to be negative and were initialized to a mean value of -0.4. All of the units received a bias connection with a mean value of -2.19. Weights were sampled from a uniform distribution centered on the mean and with a range of 1.0, with the condition that weights below zero for excitatory units and above zero for inhibitory units were clipped at zero. Furthermore, so that the network would not need to learn positional invariance (i.e., that the same amount of excitation should arrive at the hidden layer regardless of where in the array the asterisk was presented) the weights from the units in each visual sub-group were constrained to have the same values. Finally, to reduce the difference in terms of total activation across the different pools, the output from each visual sub-group was normalized to range between 0 and 1.

All of the hidden and response units integrated their inputs over time (*dt* = 0.2). A unit's output, *o*, was a sigmoid-linear function of its net input, *i*, and of normally-distributed output variability (error), ε, per the following equation:

$$o(i) = \begin{cases} \frac{1}{1+e^{-i}} + \varepsilon, & \text{if } i <= 0 \\ 0.25i + \varepsilon, & \text{if } i > 0 \end{cases}$$

This equation approximates a threshold linear function as a continuous transition from a relatively low and stable activation state regardless of the specific amount of input, to a state wherein the activation of the unit varies linearly as a function of the input (the 0.25 value was selected because it is the derivative of the sigmoid for $i = 0$; the equation is therefore continuous and differentiable despite being defined in two parts). The error reflects the variability inherent to neural processing and had a standard deviation of 0.025 for all but the visual units. For those units, error was also intended to capture the uncertainty in an individual's estimate of the number of presented asterisks, and was set to 0.1.

Representations. The input patterns for the network were generated in the same manner as in the behavioral experiment. The target outputs for the response units were set such that a 'low' number of asterisks had a target of 1.0 for the 'low' unit and 0.0 for the 'high' unit; the complementary pattern was used for presentations of 'high' stimuli. Two sets of 1,200 patterns were created: pre-training patterns and task simulation patterns. The number of pre-training patterns was determined in pilot simulations that found that after approximately 1,200 trials, the model was about as accurate as the human participants at the beginning of the practice trials.

Pre-training and task simulation. The processing of pre-training and task-simulation trials was identical. On each trial, the net input and output of the hidden and response units were set to -2.19 and 0.1 respectively. The network was then presented with the input pattern and trained for 100 unit updates. Error was calculated for the last 95 unit updates. The error was scaled by a factor of 3.0 for the units that were supposed to be off to encourage the simulation to make slower but mostly-accurate responses. A unit's target activation was adjusted such that it was considered to be correct once it was either below 0.1 or above 0.9 for units that were supposed to be off and on, respectively. Error was calculated using a two-piece error function: cross-entropy error was used for activation values below 0.5 and sum-squared error was used for activation values above 0.5 (Hinton, 1989). Weights were adjusted after each trial using a steepest gradient descent algorithm and a learning rate of 0.01. Units were considered to have made a response when a response unit's activation exceeded 0.5. The network's response latency was how many unit updates had occurred prior to responding.

Results

Overall performance for the same–same condition. The model showed the same qualitative effects (and reasonable quantitative similarity) for accuracy and latency as a function of distance from the response threshold and the hazard functions (not shown), the latency distribution for correct responses (Figure 1), and the latency-accuracy functions up to approximately 10 asterisks, beyond which the both empirical and simulation data are not very reliable (Figure 2).

Sequential effects for the same–same condition. The model showed similar increases in accuracy (not shown) and decreases in latency (Figure 3) as a function of practice, including differential latency decreases for the slowest stimuli that were closest to the response threshold.

Mixed-effect regression analyses of the effects of the characteristics of the previous trial did show some weak patterns of disagreement with the behavioral data, however. For the dependent measure of accuracy, prior accurate responses and repetitions of the previous stimulus decreased the accuracy on the subsequent response, whereas repetitions of the response increased accuracy. For the dependent measure of latency, prior accurate responses non-significantly increased accuracy, repetitions of the previous stimulus type significantly decreased latency, and repetitions of the response non-significantly increased latency.

Adaptive effects following threshold changes. Figure 4 shows that the simulation recapitulated the main effects in the behavioral experiment following a threshold change: gradual increases in the latencies for stimuli that were suddenly closer to the new threshold, and gradual decreases in the latencies of stimuli that were further from the new threshold.

Activation trajectories. While running the simulations, an additional emergent property of this architecture was observed: despite employing an activation function for which there is no explicit upper bound, the units tended to settle to stable asymptotic activation levels (Figure 6). This was true both if the input corresponded to a number of asterisks

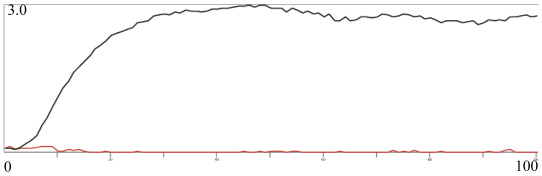

Figure 6: Activation trajectories for the 'low' (top line) and 'high' (bottom line) response units for a 'low' stimulus (32 asterisks).

that was near or far from the response threshold, although the asymptotic level of activation did differ across these two cases. This property is of interest because the leaky accumulator model required an explicit leak current to avoid runaway activation if a stimulus was presented for an extended period.

Discussion

In contrast to the claims of Ratcliff et al. (1999) the connectionist model succeeded in capturing a substantial portion of the effects observed behaviorally. The main disagreement, in terms of model-behavior mismatch, was in terms of the specific sequential effects of previous trial accuracy, stimulus repetition, and response repetition. Two causes of these discrepancies are currently being investigated: First, the current model may trade off speed and accuracy slightly differently to the human participants, which may be addressed by adjusting the scaling of the error for incorrectly activating a response unit. Second, the current simulation only instantiated memory of prior experience in the form of weight adjustments, whereas these effects may be more accurately captured in a slightly more sophisticated model wherein residual activation from processing the prior stimulus influences subsequent performance (Plaut & Booth, 2000). Nevertheless, the model did succeed in capturing the overall rates of sequential and trial-by-trial adaptation—effects that are usually outside the scope of models that are typically only evaluated by fitting aggregate data (e.g., the diffusion model) and that are not used to understand how and why model parameters (the weights in the present model) are gradually derived by learning from trial-by-trial experience.

The use of the more biologically-plausible framework also had the effect of inheriting many of the properties of the leaky accumulator model, or at least close approximations thereof, that has been tailored to response selection (e.g., bottom-up excitation, indirect competition between response units via lateral inhibition). Thus, this model gains parsimony and independent support and validity from neurobiology, while also being domain-general and suitable for studying other phenomena such as the ERPs and ambiguous word comprehension (Armstrong, 2012; Laszlo & Plaut, 2012). Moreover, as shown by the activation trajectories in Figure 6, this framework shows initial promise at generating stable asymptotic states that were not present in the leaky accumulator model without the addition of an explicit leak parameter.

General Discussion

The simulation's ability to capture most of the effects in the new behavioral data has several important implications. First, it highlights the importance of careful task analysis and a consideration of the mechanisms that drive performance in identifying an appropriate gold standard task. More generally, it also suggests that a better method for assessing model performance is through the use of a broad range of tasks that generate multiple gold standards, to avoid discrediting certain frameworks on the basis of what may ultimately be established as somewhat atypical effects (notwithstanding that more complex models with appropriate mechanisms should capture those effects). Furthermore, it highlights the value of independently-motivated biological constraints on developing an improved set of domain-general computational principles that can be readily extended to other phenomena (e.g., selecting an appropriate response when naming a nonword). Finally, this work highlights the inherent trade-off between developing tailored quantitatively-precise models versus developing domain-general models: the latter may (at least initially) produce less precise quantitative fits, but they can be extended to a much broader set of phenomena and provide not only an existence proof of model plausibility via data fitting, but a principled explanation of how and why the model performs the way it does. Thus, although much work remains in refining the biologically-plausible framework and the model of response selection, this approach promises to be of broad value to cognitive science.

Acknowledgments

The bulk of the reported work was conducted as part of BCA's PhD thesis and was supported by the CMU Psychology Department and the Center for the Neural Basis of Cognition (CNBC). We thank Kevin Mickey for discussion and for developing a preliminary simulation as part of the CNBC Summer Research Program.

References

Armstrong, B. C. (2012). *The Temporal Dynamics of Word Comprehension and Response Selection: Computational and Behavioral Studies*. Unpublished doctoral dissertation, Carnegie Mellon University Psychology Department.

Armstrong, B. C., Joordens, S., & Plaut, D. C. (2009). Yoked criteria shifts in decision system adaptation: Computational and behavioral investigations. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the* 31^{st} *Annual Conference of the Cognitive Science Society* (pp. 2130–2135). Austin, TX: Cognitive Science Society.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, *40*(1-3), 185–234.

Laszlo, S., & Plaut, D. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and Language*, *120*(3), 271–281.

Peirce, J. W. (2007). Psychopy–psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, *107*(4), 786.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

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

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