UNIVERSITY OF ROCHESTER WORKING PAPERS IN THE LANGUAGE SCIENCES – VOL. 2, NO. 1 (SPRING, 2001)

Magnuson, J. S., Dahan, D., & Tanenhaus, M. K. (2001). On the interpretation of computational models: The case of TRACE. In J. S. Magnuson and K. M. Crosswhite (Eds.), *University of Rochester Working Papers in the Language Sciences, 2 (1),* 71 - 91.

ON THE INTERPRETATION OF COMPUTATIONAL MODELS: THE CASE OF TRACE

James S. Magnuson*
Delphine Dahan+
Michael K. Tanenhaus*

* Department of Brain and Cognitive Sciences
Meliora Hall, University of Rochester, Rochester, NY, 14627

+ Max-Planck Institute for Psycholinguistics
Postbus 310, 6500 AH Nijmegen, The Netherlands

Abstract

The widespread use of neural network modeling in cognitive science has resulted in testing of explicit and sometimes novel questions based on model predictions and in turn, significant theoretical and empirical advances. Proper interpretation of computational models, however, requires that at least two key principles be observed. First, before model predictions and data can be compared, explicit hypotheses must be formulated that link the stimulus and task constraints faced by participants and measures that can be made of model performance. Second, models can be evaluated at three levels: theory, implementation, and parameters. When discrepancies are found between model predictions and data, one must determine at which level the model has failed. We will illustrate the importance of these principles with an analysis of some criticisms of the TRACE model of speech perception and spoken word recognition. Then, we review a recent study that allows a more direct linking hypothesis between TRACE and data.

Constructing an explicit model allows strong tests of theoretical assumptions, and often leads to novel predictions. Indeed, models of spoken word recognition have outstripped empirical methods in the range of predictions made by the models and the range of questions that can be addressed using conventional experimental tasks. Most notably, models of spoken word recognition such as TRACE (McClelland & Elman, 1986) make explicit predictions about the parallel activation of similar items and the time course of competition between them. Common experimental tasks typically allow single, post-processing responses to a single stimulus; predictions such as parallel activation must be tested by correlations between response measures and, for example, the number of items predicted to compete with the stimulus by a model (see Luce & Pisoni, 1998). Time course must be inferred indirectly by imposing unusual task demands (e.g., speeded response tasks; McElree 1996) or interrupting the stimulus (as in the gating task; e.g., Grosjean, 1980).

As we will discuss in detail shortly, eye tracking during spoken word recognition under certain circumscribed conditions allows a fairly transparent comparison of models and data. The mediated nature of all psycholinguistic tasks, however, makes model-data comparisons nontrivial. In order to perform such comparisons, one must formulate an explicit hypothesis linking the model to the task faced by participants in experiments.

Linking hypotheses

All experimental tasks impose some task-specific demands on participants. In spoken word recognition, common tasks include lexical decision (is what you are hearing a word or not?), shadowing or auditory naming (repeat what you hear as quickly as you can), or word or phoneme monitoring (respond whenever you hear a specified word or phoneme). Similar results are typically found with these paradigms, although different tasks show greater sensitivity to some aspects of spoken words (for example, Vitevitch and Luce [1999] found differential effects in same-different and lexical decision tasks; facilitation for high-probability phoneme sequences in nonwords in the former, inhibition in the latter), demonstrating the importance of understanding the constraints imposed by different tasks.

The goal in developing spoken word recognition models is clearly not to build a model of each task. Instead, a single model should incorporate general theoretical assumptions and provide a basis for explaining results from various tasks. Different patterns of results in different tasks indicate that a model must either include architectural features that predict the task differences, or, perhaps more plausibly, the output of the model must be passed through a model of the decision processes required for the current task demands.

Of course, a model that can account for data from a wide range of tasks on the basis of a set of theoretical principles should be preferred to one that includes task-specific mechanisms. Without a theoretical basis for model-internal mechanisms to account for task-specific phenomena, there is no basis for preferring such a model to a simpler model that accounts for task-specific differences via task-specific decision models – or linking hypotheses.

A linking hypothesis provides a quantitative link between the behavioral demands of the task faced by participants and the output of a model. For instance, given a simple case where the model output is activation over a set of lexical units (as in TRACE) and a behavioral response is hypothesized to be proportional to model activation at the time when the response is made, two things are needed to formulate an explicit linking hypothesis (beyond the simple assumption that behavioral responses should be proportional to activation). First, a transformation relating activation to the response measure (e.g., activation to probability of a correct response), which could be as simple as a one-to-one mapping, but might be a more formal decision model (e.g., the Luce [1959] choice model, as used by Allopenna, Magnuson & Tanenhaus [1998]). Second, in the case of this simple mapping, time in the model must be related to real time (see Allopenna et al. for an example).

In the absence of an explicit linking hypothesis, one could easily misinterpret a model's performance. For example, in a model in which activations are not explicitly limited by the activations of other items (as opposed, e.g., to a model in which activations must sum to 1), the difference in raw activation between an item and a potential competitor might appear far too large or small to predict the pattern seen in participant data. Model activations transformed by a choice model, however, might suggest substantially larger or smaller differences between lexical items (we will discuss choice models in more detail below). Even with an explicit linking hypothesis in place, however, care must be taken when interpreting differences between models and data.

Model failures: levels of analysis

Before a model failure can be attributed to underlying theoretical assumptions, one must try to ascertain that the failure cannot be attributed to the implementation of the model (from the representation of the input, to the levels of units [in a neural network model, for instance] and the processing dynamics [activation functions, etc.]) or simply to parameter settings (in TRACE, for example, if simulations suggest competitors are inhibited too much, one cannot conclude that lateral inhibition at the lexical level is a flawed assumption without testing different values of lexical inhibition, phoneme inhibition, bottom-up excitation, etc.).

The latter is of particular importance, as there have been suggestions that the TRACE model should only be tested with minor deviations from the original parameter set (Frauenfelder REF). We argue that this is not a tenable position. On the one hand, it is true that if different parameter sets are used to model different results, the model loses its generality; the breadth of model successes cannot be attributed to underlying theoretical assumptions if each success requires different parameters. On the other hand, equating a model with a parameter set produces a similar problem. The model now loses generality because the constraints imposed by the parameter set are, in effect, placed on a par with the underlying theoretical assumptions. The alternative is simple. Although experimenters should not limit model explorations to a "standard" parameter set, when they diverge from conventional parameter settings the onus is on the modeler to test whether the new parameters prevent the model from fitting results it was known to fit with the previous settings (this is no small burden when a model has been shown to account for a wide range of results).

We will now illustrate these principles by considering an experimental paradigm that has attracted much recent attention because of the difficulties two groups of researchers have had modeling its results using variants of the TRACE model. We will discuss replications of their simulations, along with a recent study that employs an eye-tracking variant of the task. When the principles of model interpretation we have discussed are observed, it is clear that TRACE is actually more consistent with the results than has been claimed in the past, and in fact provides a better fit to the eye tracking data than competing models.

Subcategorical mismatch effects on spoken word recognition

TRACE simulations conducted by Marslen-Wilson and Warren (1994), which failed to account for data from their subcategorical mismatch paradigm (described shortly), have provided fodder for much criticism of TRACE. Some appear to find this failure damning evidence against TRACE, and the data from the Marslen-Wilson and Warren experiment (replicated by McQueen, Norris and Cutler, 1999) has been central to recent debates over model architecture (Norris, McQueen and Cutler, 2000). We will describe the Marslen-Wilson and Warren data and simulations, and discuss what the apparent failure of their simulations actually implies for TRACE. We find that they did not adequately explore the linking hypothesis they adopted, and that they failed to examine the level at which their simulations failed. We then discuss a recent eye-tracking variant of the subcategorical mismatch paradigm, which TRACE fits quite well. With an explicit linking hypothesis between TRACE activations (or fixation data) and the lexical decision task, we find that TRACE provides a basis for the pattern of data found by Marslen-Wilson and Warren and McQueen et al. The failures of the Marslen-Wilson and Warren (1994) simulations do not provide evidence against TRACE, but demonstrate the importance of establishing linking hypotheses and distinguishing between levels of model evaluation.

Marslen-Wilson and Warren (1994)

Marslen-Wilson and Warren (1994; "MWW," hereafter) adapted the subcategorical mismatch paradigm, which had been used to study phonetic perception (e.g., Whalen, 1984), to study spoken word recognition. Subcategorical mismatches are created by splicing together two stimuli such that coarticulatory information does not match the following phonemic segment. For example, if the vowel of the vowel-consonant sequence /ud/ is excised and spliced onto the excised consonant /v/ of /uv/, the result will be recognizable as /uv/, but the vowel contains subcategorical (i.e., subphonemic) cues more consistent with /d/ than with /v/. By varying the splice point, one can manipulate the relative evidence for the two phonemes, and measure the degree to which the listener is sensitive to such subphonemic variations. MWW cross-spliced the final segments of words in order to create stimuli in which the subcategorical mismatching information specified a potential word competitor (e.g., the initial consonant and vowel of "jog" spliced with the final consonant of "job", creating a stimulus that would be most consistent with "jog" initially, due to coarticulatory information in the vowel, but would ultimately be consistent with "job") or a nonword ("jod" plus "job").

Note that the three original stimuli differed in place of articulation. MWW used a notational system specifying the lexical status of the cross-spliced stimuli (W or N for word or nonword) and the place of articulation (1, 2, or 3, arbitrarily). It is worth walking through this notational system in detail as we will use it throughout the remainder of this paper. In our example stimulus set of "job", "jog" and "jod", "job" would be W1 (word, first place of articulation), "jog" would be W2, and "jod" would be N3 (nonword, third place of articulation). From a W1/W2/N3 stimulus triplet, MWW would construct three critical stimuli, all of which would be consistent with W1 at word offset: W1W1 ("job" spliced to itself, which we will also denote as "jo(b)b", where the consonant in parentheses indicates the consonant specified by coarticulation in the vowel); W2W1 ("jog" spliced to "job", or "jo(g)b"); and N3W1 ("jod" plus "job", or "jo(d)b"). Another kind of triplet was also crucial to the MWW study, consisting of two nonwords (e.g., "smob" [N1] and "smod" [N3]) and one word ("smog" [W2]). The cross-spliced stimuli all were consistent with the nonword, N1, at offset (N1N1, W2N1, and N3N1).

These stimuli allowed MWW to test important predictions from models like TRACE, specifically, that in the mismatching word-word (W2W1) and word-nonword (W2N1) cases, the misleading coarticulatory information in the vowel should activate a word (W2), and have inhibitory effects on performance compared to cases where the mismatching information matched a nonword (N3W1 and N3N1). The basis for this prediction in TRACE is lateral inhibition between word nodes at the lexical level. If the input initially favors "jog" when hearing "jo(g)b", the corresponding node will be more strongly activated than when the input initially favors a different word ("jo(b)b") or a nonword ("jo(d)b"). The result will be that the "jog" unit will exert a stronger inhibitory influence on the "job" unit given "jo(g)b", and slow recognition. In the nonword case (W2N1), "smo(g)b" will activate "smog" such that it will still be unlikely to be recognized, but will delay a nonword response as the system, e.g., waits for "smog"'s activation to drop below the "nonword" threshold.

Counter to the lexical inhibition predictions (W1W1 < N3W1 < W2W1), MWW found that reaction times and error rates to N3W1 and W2W1 did not differ reliably (though both were slower and more errorful than responses to W1W1; note that McQueen, Norris and Cutler [2000] replicated both the word and nonword pattern reported by MWW in Dutch). MWW used simulations with TRACE to evaluate the difference between their results and actual TRACE predictions. Their simulations, as presented in the paper, appear to demonstrate a rather striking failure of TRACE to account for their data. The primary results are plotted in Figures 1 and 2 (corresponding to MWW's Figures 12 and 13, respectively; note that the values were hand-estimated from the MWW figures; the x-axis label, "Cycle / 4", reflects the fact that MWW reported tracking simulations out to 80 cycles and then down-sampling by a factor of 4). The figures plot response probabilities for W1 given the different stimuli. MWW did not report the underlying activation values, nor the details of how they

calculated response probabilities, aside from saying they used the form of the Luce choice rule used by McClelland and Elman (1986) to model phonetic decisions.

First, response strengths are calculated for each item based on activation:

$$S_{it} = e^{ka_{it}} \quad (1)$$

where S_{it} is the response strength for item i at time t, k is a free parameter controlling the advantage given to larger values in the exponential function (that is, it is a way of instantiating inhibition), and a_{it} is the activation of item i at time t. Then, the response probability of each item is calculated simply as each response strength normalized over the response strengths of all n items:

$$P_{it} = \frac{S_{it}}{\sum S_{ij}} \quad (2)$$

MWW did not formulate a formal linking hypothesis. They suggested that a threshold model applied to response probabilities would be sufficient, but because of the large differences observed in the response probabilities, relied on qualitative comparisons. In the case of the word stimuli, even qualitative inspections reveal that the expected inhibitory effect is there. The response probability is far lower for W2W1 than for W1W1 and N3W1. But recall that the pattern of lexical decision response times MWW found was W1W1 < N3W1=W2W1. Thus, their TRACE simulations fail to account for this pattern (MWW state: "this is a clear failure of TRACE to match human performance", p. 669). Indeed, from the MWW simulations, it appears that there is not a clear basis for recognizing W1 given W2W1; W1's activation asymptotes at a bit more than .3.

In the case of the nonword stimuli, the lexical decision pattern was N1N1 < N3N1 < W2N1 (although N1N1 and N3N1 did not differ reliably). This is consistent with the pattern observed in their simulations (see Figure 2). We reproduce their nonword simulation data here to make two points. First, note that the probability of W2 given W2N1 follows a time course similar to that found for W1 given W1W1. There does not appear to be a basis for rejecting W2 given W2N1, in fact. Second, this highlights the weak response probability for W1 given W2W1 in the word simulations; the probability of W2 given W2N1 is much higher, and a threshold model would have to predict that W2 be "recognized" given W2N1, and that W1 not be recognized given W2W1. As MWW conclude, this would result in a high error rate given W2N1, and therefore, their TRACE simulations also fail to account for the nonword data.

We have been careful to describe these as failures of MWW's simulations rather than failures of TRACE. This is because MWW failed to observe either of the general principles for model evaluation that we have described.

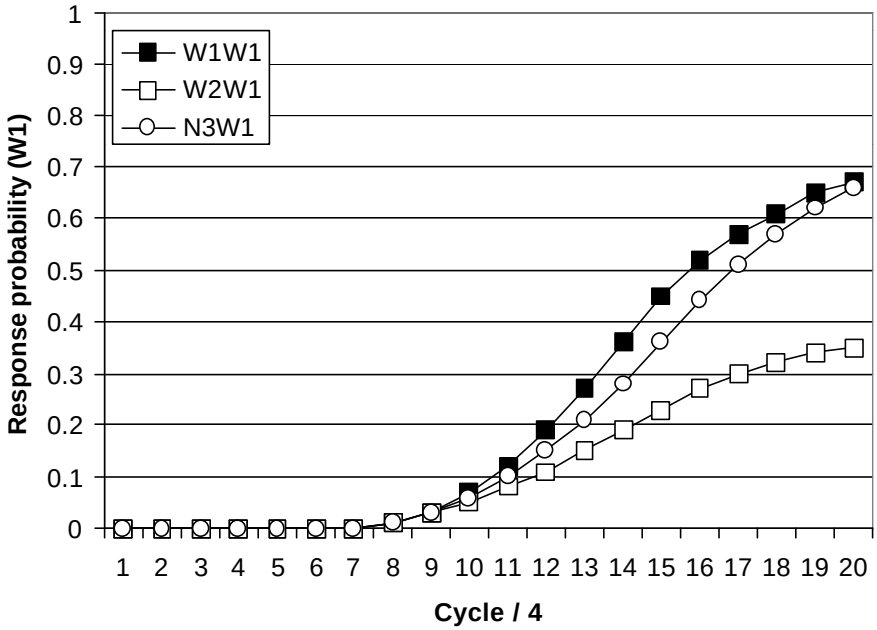

Figure 1: TRACE simulation results for word stimuli, adapted from Marslen-Wilson and Warren (1994).

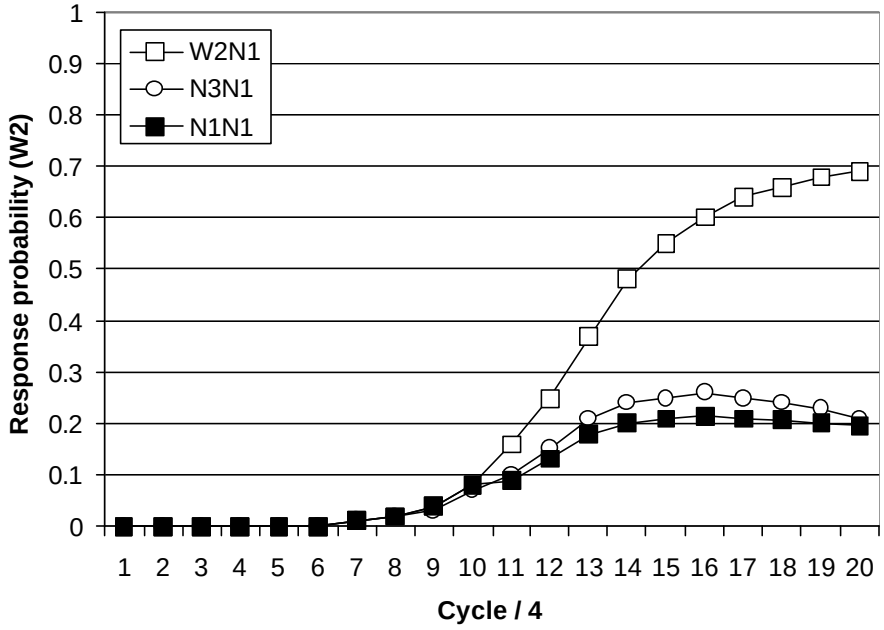

Figure 2: Nonword simulations adapted from Marslen-Wilson and Warren (1994).

Linking hypotheses. MWW fail to establish a formal linking hypothesis between their simulations and their data. As we have discussed, at the level of comparing their response probabilities to lexical decisions, it was not necessary to formalize their response threshold model of lexical decision: their TRACE simulations predicted large effects that were clearly missing from their data. However, if we back up a step at a time from their response probabilities, we will discover layers of assumptions that form an implicit linking hypothesis.

Basis for decisions. MWW assumed that in the case of the word stimuli, the only basis for a correct "yes" response would be the response probability for W1. However, a correct lexical decision does not require precise recognition; it does not matter which response probability a participant responds to on the "word" trials, as long as she responds "yes". An obvious question raised by the low response probability for W1 given W2W1 is why the probability of W1 is so low. Given the high response probability for W2 given W2N1, an obvious explanation is that the activation of W2 given W2W1 was very high and inhibited the activation of W1 and thus lowered its response probability (cf. Tanenhaus, Magnuson, McMurray & Aslin, 2000; Dahan, Magnuson, Tanenhaus & Hogan, in press). If we allowed "yes" responses whenever any item's response probability exceeded a threshold, we would expect some early "yes" decisions in response to the probability of W2 given W2W1. We will explore this possibility in detail below.

Response probabilities vs. activations. MWW report response probabilities for the items they assumed to be relevant for the word (W1) and nonword tasks (W2). Response probabilities, however, are not an inherent part of TRACE, and entail another linking hypothesis, in this case between activations and the nature of the response required of participants. The assumption is that raw perceptual evidence in the form of TRACE activations will result in choice behavior similar to that in the kinds of tasks that the Luce choice rule (1959) was designed to model (i.e., requiring competitive, nonlinear transformations of evidence levels). This issue carries over to the second general principle.

Levels of analysis. MWW also fail to consider what aspects of TRACE might be implicated in their failed simulations. They report that they found it "hard to remedy" their failure to simulate the lexical decision task (LDT) data for the word stimuli by "altering TRACE's parameters" (p. 669). As we've just discussed, however, they've added a choice model to TRACE, with its own assumptions and parameters. Changing the parameters of the underlying TRACE model might have little effect depending on the value of *k*, for example (see equation 1), or the effects of such changes, filtered through the choice model, might be amplified. It is difficult to evaluate this possibility, though, because MWW reported few details about the parameters they used for their simulations, and a paper they cited as containing the details is no longer available (a related paper sent to us by Paul Warren did not contain these details, either).

MWW attribute the failure of their simulations to lateral inhibition at the lexical level of TRACE (p. 669). As they point out, this is a central aspect of TRACE's architecture. They conclude that the assumptions about "lexical competitors and how new information affects existing activation levels" (p. 671) are to blame, and prefer models that posit bottom-up inhibition. A possibility that is not considered by MWW, and which we shall not dwell on, is that TRACE's input representations might not be adequately realistic to capture the essence of the cross-spliced stimuli. The inputs are obviously poor analogs to real speech, but as we shall see, they appear to provide a good enough analog to the experimental stimuli. However, before attributing a model failure to a relatively deep cause such as theoretical assumptions regarding inhibition, more superficial potential causes need to be eliminated.

We found the lexical decision response time patterns reported by MWW and McQueen et al. (1999) to be quite surprising, given the central theoretical assumptions underlying all current models of spoken word recognition: given a spoken stimulus, multiple candidate items are activated as a function of their goodness of fit with the input as it unfolds over time, and activated candidates compete for recognition (the most elegant demonstration being Luce and Pisoni's [1998] finding that the time to recognize a target word depends on the number of words that are phonetically similar to it). This general principle predicts that if the system is sensitive to subcategorical information, given a subcategorical mismatch that temporarily favors another word (e.g., W2 given W2W1), that word should compete (i.e., inhibit the ultimate target word) more than when the subcategorical mismatch

does not map onto an existing word. Therefore, Dahan et al. (in press) designed an eye-tracking variant of the MWW task.

Dahan et al. (in press) eye tracking experiment

Over the last several years, an eyetracking paradigm has been developed for studying spoken language processing (e.g., Tanenhaus et al., 1995). In this "visual-world paradigm," participants interact with a display of multiple objects (real objects on a table top, or two-dimensional images on a computer screen which can be manipulated using the computer mouse). With a properly constrained task (one in which participants must make a visually-guided reach in response to a spoken instruction[1]), eye movements are closely time-locked to speech.

For example, given a display containing a piece of candy, a candle, and other items with unrelated names, participants are already equally likely to be fixating the candy and candle in response to an instruction to "pick up the candle" once they have heard the initial /kae/ of "candle." Note that "equally likely" entails averaging many trials, since participants make, on average, 1.5 fixations per trial (Allopenna et al., 1998). Unlike other tasks, however, the visual world paradigm allows (1) measurement of responses to multiple items, (2) a truly on-line measurement, since most eye movements are made without conscious awareness, and eye movements are typically made incrementally prior to the availability of a conscious response, and (3) an approximation to a continuous measure, given sufficient trials. That is, if we plot the proportion of trials on which participants were fixating each possible object at each time step (e.g., each video frame sampled at 30 Hz) we find that fixation proportions map closely onto phonetic similarity over time (among other factors), with a constant lag.

Eye movements tend to follow speech with a constant lag close to the minimum time it takes to plan and launch an eye movement in simple tasks – about 150 to 200 ms. In our "candle" example, the proportion of fixations to the candle begins to diverge reliably from the proportion to the candy about 200 ms after the point of disambiguation (the onset of /l/; Allopenna et al., 1998; Spivey-Knowlton (1996); Tanenhaus et al., 1995). This task has proven sensitive to phoneme-level similarity (onset and offset competition effects, Allopenna et al., 1998), target and competitor frequency (Dahan et al., in press), and neighborhood density (Magnuson, Tanenhaus, Aslin and Dahan, submitted). The paradigm allows an explicit linking hypothesis and quantitative tests of model predictions. We will explain the linking hypothesis in the context of the subcategorical mismatch experiment.

Dahan et al. (in press) applied the visual-world eye tracking paradigm to the subcategorical mismatch questions raised by MWW using the following logic. First, the general predictions made by TRACE (W1W1 < N3W1 < W2W1) are consistent with assumptions shared by all current models of spoken word recognition. Second, formulating an explicit hypothesis linking activations in TRACE to lexical decisions is not trivial. As we discussed, a simple threshold model based on the response probability of the intended target is not tenable. The model has no way of knowing *a priori* which word is the target, and there is no guarantee that the item with the highest response probability corresponds to the target, especially with doctored stimuli with misleading coarticulatory cues. Thus, a more direct and continuous measure (phonetic similarity over time, as revealed by fixation proportions over time) might shed light on the discrepancy between model predictions and the lexical decision results.

Dahan et al. (in press) created stimuli following the procedure described by MWW (e.g., splicing stimuli at the offset of the vowel; see Dahan et al. for details), and created 15 sets of word

[1] Such tasks allow a functional interpretation of eye movements and thus avoid many of the pitfalls of eye movement interpretation discussed by Vivianni (1991). Note, however, that several experiments now suggest that eye movements remain fairly closely time-locked to speech even in passive viewing tasks (Cooper, 1974; Altmann and Kamide, 1999**;** Sussman and Sedivy, this volume).

items analogous to those used by MWW (e.g., W1=net, W2=neck, N3=*nep). The visual world paradigm imposes an additional constraint on the words – they all must be easily imageable nouns, since they must be presented to the participants visually. Participants were seated at a computer, and saw displays with four items. Multiple lists were constructed so that items were not repeated within participants. On critical trials, the picture corresponding to W1 was displayed, and one of the cross-spliced stimuli was presented in the context of an instruction to click on one of the pictures (e.g., "click on the ca(b)t"). On some critical trials, W2 was also displayed. As we will see, this allowed a crucial test of the linking hypothesis itself.

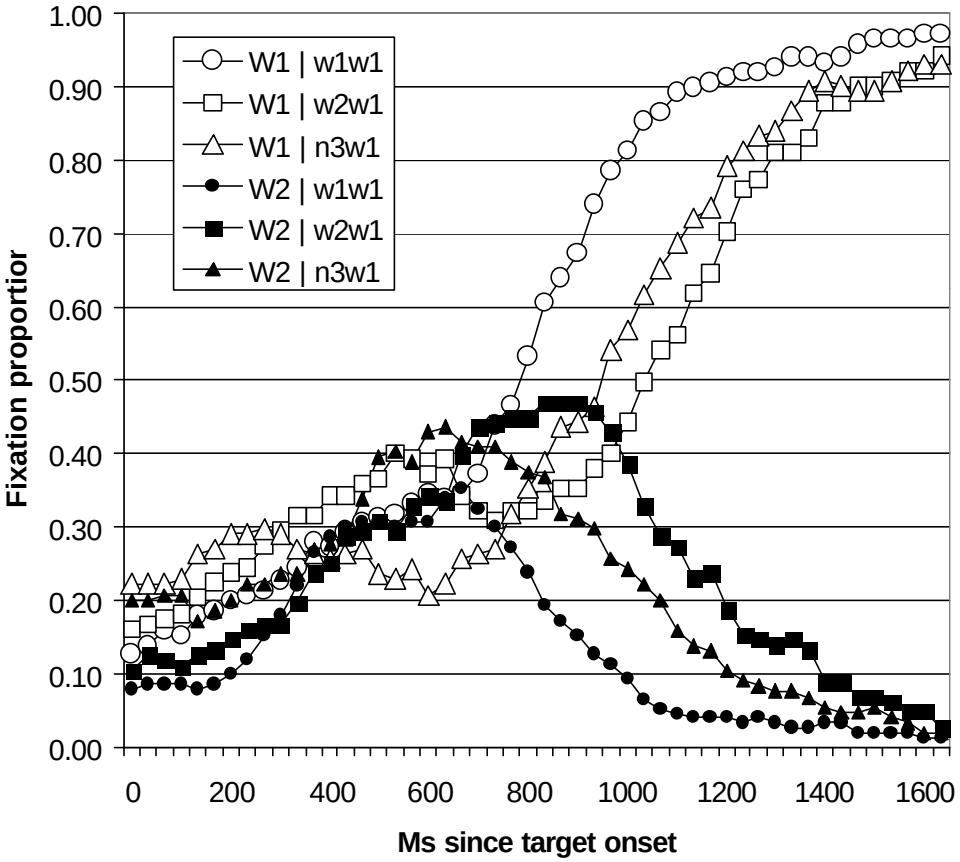

Figure 3: Fixation proportions over time from Dahan et al. (in press) in the eye tracking variant of the subcategorical mismatch paradigm when both W1 and its competitor, W2, are displayed (adapted from Dahan et al., in press).

In Figure 3, data from the conditions in which both W1 and W2 were displayed are plotted. Note that this data indicates a clear inhibitory effect of misleading coarticulatory information (the fastest rise time is for W1 given W1W1), and a larger effect when that information matches another word in the lexicon (W2W1) than when it does not (N3W1). In addition to plotting the proportion of fixations to the target, W1, we have plotted the proportion of fixations to the onset competitor, W2. While W2 is much more strongly considered than the unrelated items in all three stimulus conditions (the unrelated items are not included for the sake of clarity in the figure), the proportion of fixations to W2 follows a pattern complementary to that of W1: the largest proportion to W2 is in response to W2W1, then N3W1, and last to W1W1. Thus, the pattern of eye movement data is clearly consistent with the general prediction that competitors should be activated in proportion to their fit to the input,

and items active in parallel inhibit one another. In other words, the eye tracking data provide clear evidence for lexical competition.

We decided to replicate the MWW TRACE simulations, with the expectation that the activation and hence response probability of W2 given W2W1 incorporated into a different linking hypothesis might reduce the magnitude of the discrepancy between TRACE and the MWW LDT data. We followed the procedure described by MWW, creating analogs to our stimuli by cross-splicing the TRACE input stimuli at "vowel offset" (after the last frame containing vowel input), and presenting them to TRACE using the "standard" parameter set established by McClelland and Elman (1986). Raw activations are shown in Figure 4. To our surprise, given the MWW simulations, the underlying TRACE activations of W1, without considering W2, appeared much more consistent with the MWW LDT data than the MWW simulations suggested would be possible. Rather than the extreme differences between the W2W1 condition and the others shown in Figure 1 (and their Figure 12), we found much more modest (although substantial) differences.

To make quantitative comparisons between the eye movement data and TRACE, we used an explicit linking hypothesis (developed by Allopenna et al. [1998] and Dahan, Magnuson and Tanenhaus [2001]), which assumes eye movements in the visual world paradigm are based on two sources of input: the bottom-up speech input and the visual display. Our model for the bottom-up speech input was lexical activation in TRACE. Response strengths for the items of interest were computed using Equation 1. To incorporate the constraints of the visual display, we used a variant of the Luce choice rule (Luce, 1959). The participant was limited to four possible fixation targets – the items displayed on the screen. So instead of normalizing over the response strengths of all items in the lexicon, as in Equation 2, we normalized over the response strengths of only the four possible fixation targets (using a value of 7 for *k*, the same value used to fit data by Allopenna et al. [1998] and Dahan, Magnuson and Tanenhaus [2001]). Thus, lexical activation was based on activation and competition over all lexical items, but our choice model explicitly incorporated the choice task faced by the participant.

Simulation results for the data in Figure 3 (W1 and W2 displayed) are shown in Figure 5. The raw TRACE activations in Figure 4 provide a good qualitative fit to the data shown in Figure 3, but the simple linking hypothesis transforms the activations to provide a closer fit to the data. However, this is not a case of simply tweaking parameters to improve fit; the choice model is motivated by the explicit linking hypothesis.

The data from critical trials where W1 was displayed but W2 was not (in which case W1 was presented among three unrelated distractors) are shown in Figure 6. The linking hypothesis makes explicit predictions about the differences that should be observed when W2 is not displayed. The auditory stimulus is the same, so the contribution of the lexical model (TRACE) is predicted to be identical. The visual choices are different, though, with the result that the degree of competition with W2 that can be detected is predicted to be reduced. Although the support for W1 is less given W2W1 than W1W1, fixations cannot be made to W2, and thus a more modest inhibition effect is predicted. Simulated predictions from TRACE and our linking hypothesis (with k=7, normalized over the target and three unrelated distractors) are shown in Figure 7. As predicted, the W1W1 < N3W1 < W2W1 pattern in rise times is observed again (Figure 6), but the differences are more modest than when W2 was displayed.

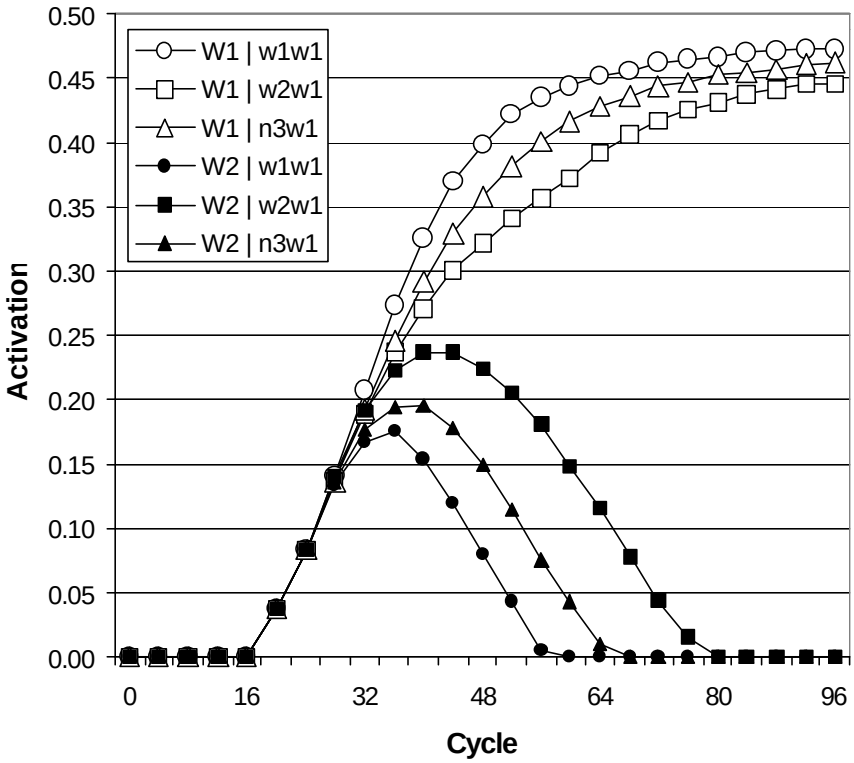

Figure 4: TRACE activations from simulations analogous to the experiment plotted in Figure 3 (adapted from Dahan et al., in press).

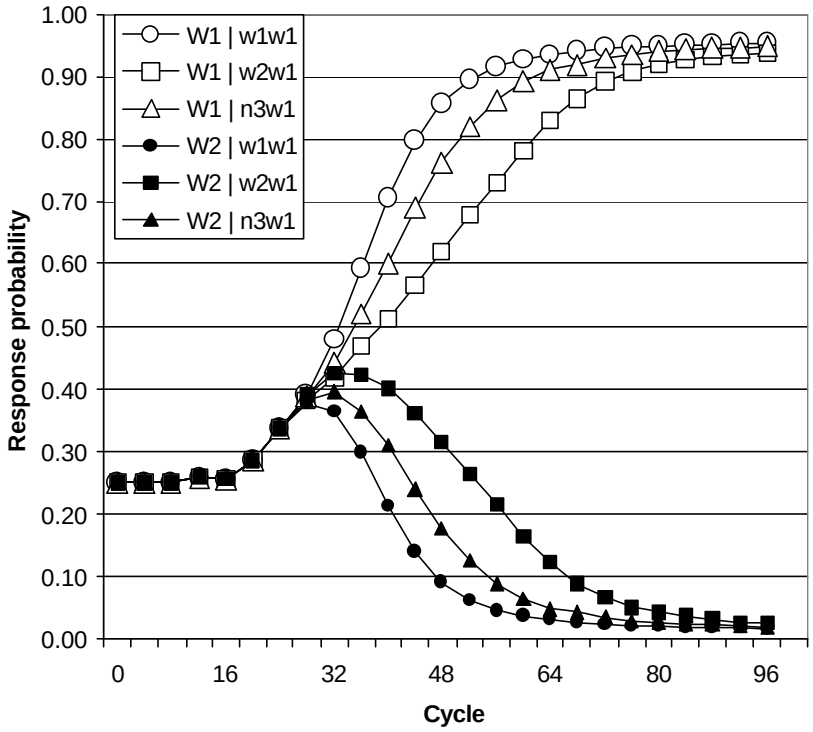

Figure 5: Response probabilities predicted from the TRACE activations displayed in Figure 4 transformed with a variant of the Luce choice rule (adapted from Dahan et al., in press).

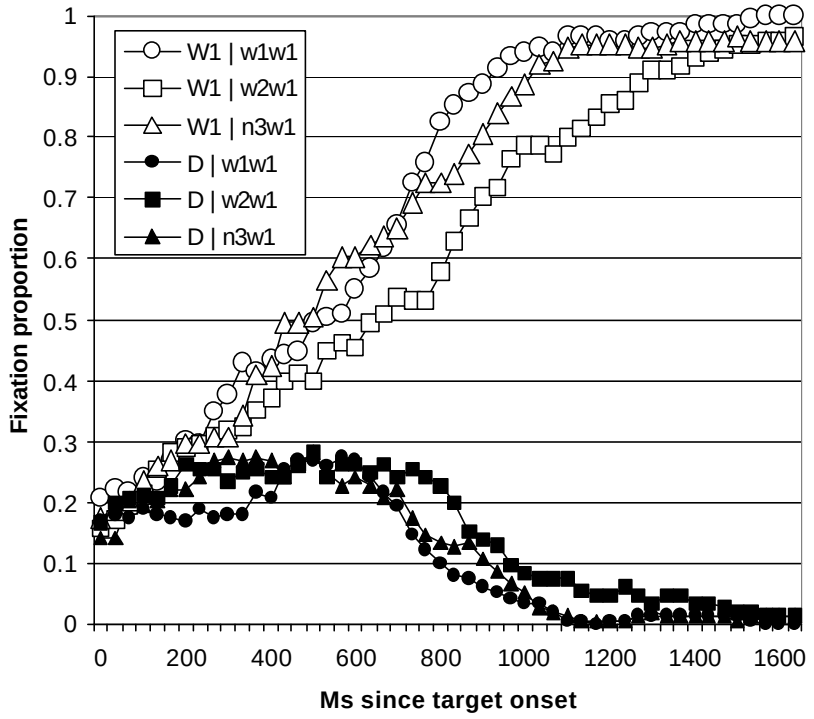

Figure 6: Fixation proportions over time from Dahan et al. (in press) in the eye tracking variant of the subcategorical mismatch paradigm when W1 was displayed, but its competitor, W2, was not (adapted from Dahan et al., in press); "D" indicates an unrelated distractor.

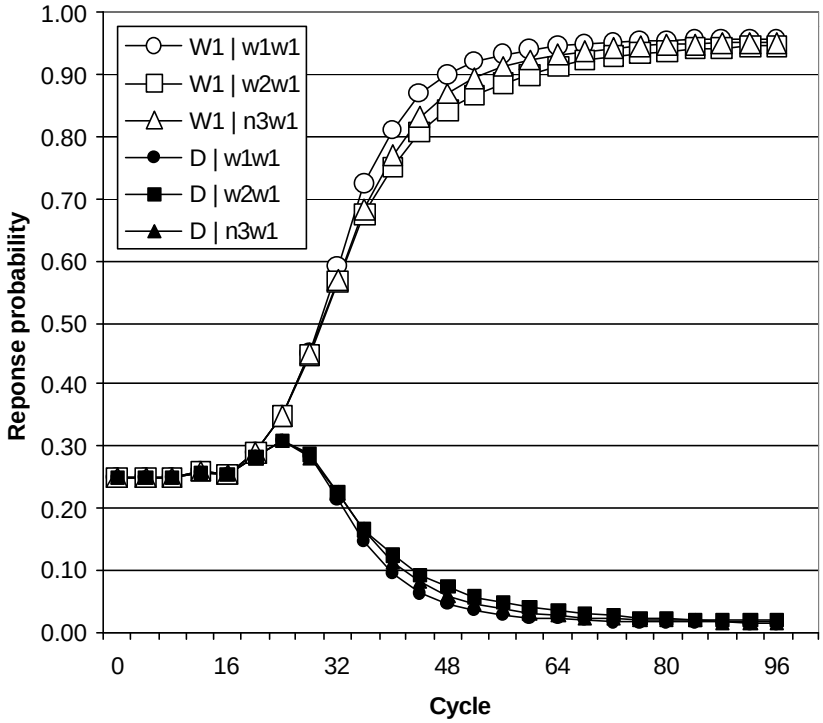

Figure 7: Response probabilities predicted from the TRACE activations displayed in Figure 4 when W2 is not included in the choice rule (adapted from Dahan et al., in press).

There are substantial discrepancies between the TRACE simulations and the data. In particular, cohort proportions rose substantially above target proportions in the data shown in Figure 3, whereas in TRACE, the cohorts never outstrip the targets. This suggests that, for example, more or less lateral inhibition, or perhaps stronger bottom-up weights, might be called for in TRACE. When W2 is not included in the choice rule, TRACE predicts a much larger weakening of the W2W1-W1W1 difference than we observe in Figure 6. This suggests our linking hypothesis may need further work. However, we did not explore parameter or linking hypothesis changes because we have used the same parameters and hypothesis to simulate competition effects (Allopenna et al., 1998) and frequency effects (Dahan et al., 2001).

Although the differences between simulations with and without W2 do not predict the differences between the data in the corresponding conditions very precisely, this result provides support for the linking hypothesis. Rather than simply fitting the data, the choice model incorporated into the linking hypothesis generates testable predictions independent of the underlying lexical model. The discrepancies may indicate that we either need to postulate a stronger influence of the visual displays (e.g., via indirect lexical activation), or a more concrete choice mechanism. We use the choice rule to predict fixation probabilities in the aggregate, but we cannot account for trial-by-trial fixation "decisions" or individual differences. A model of probabilistic fixation generation might provide a better fit and might also allow us to explore individual differences. We are currently working on such a model.

Differences between simulations. We invested considerable effort in trying to understand the differences between our simulations and those conducted by MWW. MWW provide scant details about the procedures they followed. As we mentioned above, the paper cited by MWW as containing the details about the simulations is no longer available, although Paul Warren sent a related paper with a few details, and was helpful as we attempted to replicate the MWW simulations. We will now review these attempts.

MWW used five sets of word and non-words that ended in voiced stops, and used a 390-word lexicon comprised of "all the uninflected monosyllabic words using TRACE's 15 phonemes [from] Longman's [1987] Dictionary of Contemporary English" (MWW, 1994, p. 668). Warren provided us with a 392-word version of the lexicon they used ("Monolex"). An examination of Monolex shows that there are only 21 possible bases for stimulus sets. Among TRACE's four vowels, there are 10 possible sets for /a/, 2 for /i/, 0 for /u/, and 4 for /^/. Of the three possible combinations of consonants 1 and 2 to make W1 and W2 (i.e., sets where there is a word completion with C1 (consonant 1, i.e., place of articulation 1) and C2 but not C3 [e.g., /dab/, /dad/, /dag*/]), there are 6 for /b,d/ (or /d,b/), 11 for /b,g/, and 4 for /d,g/. To make a complete set of words and non-words, there must also be at least one neighborhood in the lexicon where VC2 makes a word but VC1 and VC3 do not (e.g., */glab/, /glad/, */glag/). The number of possible sets can be doubled by rotating items through W1 and W2 assignments (e.g., W1=/dab/, W2=/dad/, as well as W2=/dab/, W1=/dad/). It was possible to make complete sets for every context with the vowels /a/, /i/ and /^/, with one exception: /id/, /ig/ was not possible (although /ig/, /id/ was) because there was no word ending in /ig/ that did not have neighbors ending in /ib/ or /id/. This left us with 41 possible base stimulus sets from Monolex (many more would be possible if we used every possible nonword set for every word set – but these would be terribly redundant, unless we had reason to suspect that the identity of the third consonant was crucial; instead, we attempted to make the sets resemble the one set of natural materials described by MWW in their Table 1 on p. 657).

We divided these sets into two groups according to whether W1 and W2 were CVCs or CCVCs. We began using the 25 CVC sets. This allowed us to avoid the problem of aligning the outputs at the vowel, as MWW did, since they used a mixture of CVC and CCVC items. Items were spliced one time slice before the center of the final C (slice 24) and one slice earlier and later. Feature spreading was set to +/- 6 slices for all features, as in the original McClelland and Elman

(1986) simulations. We again failed to replicate the patterns reported by MWW in their Figures 12 and 13: with 24 as the splice point, the activation and probability patterns closely resembled those we found in the simulations with analogs of our own experimental items; we did not observe the extreme difference between W2W1 and N3W1 or the "recognition" of W2N1 as W2. A splice point of 25 was clearly too late: neither W2W1 nor N3W1 were "recognized" as W1: the activations and probabilities of these items both peaked much lower than W1W1, and dropped to baseline levels at about the same time W1W1 peaked. A splice point of 23 was too early; there were almost no differences between the three word conditions.

Another possibility is that the CCVC items might be responsible for MWW's simulation results. The extra initial consonant might provide enough time for competitor activations to reach a level sufficient to yield the significant depression of W1's activation MWW reported. However, in simulations with the 16 CCVC items (with splice points at 30), competitors probabilities exceeded those in the CVC simulations by only negligible amounts.

An examination of each individual item showed that almost all items share the same rise time patterns: W1W1 < N3W1 < W2W1, with the difference between WW1 and N3W1 of approximately equal magnitude as the difference between N3W1 and W2W1. The activation of W1 reaches approximately the same peak (between 70 and 80) around the 70th cycle given any of the three stimuli, W1W1, N3W1, W2W1. For the nonword items, W2 is activated much more strongly by W2N1 than by N3N1, but the peak given W2N1 is about half that of W1 given W1W1, whereas MWW's simulations showed W2 | W2N1 and W1 | W1W1 as having nearly identical activation patterns. There were eight items that showed trends somewhat similar to those reported by MWW. These were /drab/, /grid/, /kub/, /kud/, /lig/, /st^d/, /s^b/, and /tab/. However, none of these really fit the trends from MWW's simulations. In particular, the probability of W1 given W2W1 is always much higher than in the MWW simulations.

Another possibility is that we used different parameter sets. We used the parameters given by McClelland and Elman (1986) for our initial simulations. MWW report using “the standard parameter settings that give TRACE the appropriate performance in normal word recognition (cf. Frauenfelder & Peeters, 1990)” (p. 668). However, Frauenfelder and Peeters only describe the values of a couple of parameters (the maximum and minimum activation levels). Frauenfelder & Peeters (1998), however, describe an alternative set of parameters that they tweaked to get good performance with their "Biglex" lexicon. Simulations with those parameters yield virtually identical results.

The value of *k* in the Luce choice rule could also have a large effect. Since MWW did not report the value they used, we tried to find the value that would yield the peak response probability level for W1 given W1W1 in their Figure 12. The best value is 15, but there is no value between 5 and 20 that gives a result much more similar to the MWW simulations (and the peak response probabilities differed more from Figure 12 as the value of *k* was varied from 15).

One striking difference between the CCVC and CVC simulations separately was that the CCVC items reach higher probability levels than the CVC items with *k* set to 10. Thus, another possible explanation is that averaging particular CCVC and CVC items could lead to the trends reported by MWW. However, our attempts at coming up with such an average have failed. We could not come up with a mixture of five CCVC and CVC items which, when averaged, would resemble the MWW simulations more closely than our own.

Finally, we considered that there might be differences in our implementations of TRACE. Paul Warren supplied us with the source code for his Macintosh-based TRACE implementation (MacTRACE), and we tested several stimuli using the default parameter MacTRACE settings (the Frauenfelder and Peeters parameters, and a value of 20 for *k*). The results are shown in the top panel of Figure 8. The results closely resemble those from all of our other simulations. The main difference is that the probabilities for W1 asymptote more quickly. This is due to the large value of *k*. Note that the probabilities also start out at 0, whereas they began at .25 in our simulations. This is because in

our simulations, only the displayed items were included in the choice rule, but for these simulations, all lexical items were included (following the procedure of MWW). However, these are not crucial differences. Including all items changes the time course only slightly compared to our simulations, but still does not yield the results MWW reported. We experimented with using smaller values of *k*. With much smaller values (around 10), we can get W1 given W1W1 and W1 given N3W1 to asymptote around .7 (as in the MWW figures), but W1 given W2W1 does not asymptote at .3 – it always asymptotes around the same value as for the other conditions, although slightly later.

If we were to down-sample by a factor of 4, and, like MWW, clip at the twentieth down-sampled cycle the results would not resemble those presented by MWW at all. If instead we down-sample by 2 and clip the results at the twentieth down-sampled cycle, we find differences at cycle 20 that resemble those shown in MWW's Figure 12, although the trends do not perfectly resemble theirs (because W1 given W2W1 always reaches too high a probability by the clip point).

This suggests that it might be possible, by down-sampling by a factor of 2 instead of 4, clipping at the twentieth down-sampled cycle (cycle 40), and using a smaller value of *k*, to obtain results comparable to theirs. However, this requires an indefensible assumption. Why should the results be clipped? The model may pass through a stage where very large differences are predicted, but how can we determine that that stage should form the basis for the lexical decision results? The differences become smaller as time goes on, and clipping the simulation data where the differences are large is simply arbitrary.

Thus, despite going to great lengths to attempt to replicate the MWW simulations, we have been unable to, even with assistance from Paul Warren and the computer programs they used to conduct the simulations. We are convinced that the basic prediction of TRACE is not so extreme as that claimed by MWW for the subcategorical mismatch data. MWW remarked that "…we cannot guarantee that there is not some version of TRACE that would do a better job of simulating the … results" (p. 671). Our simulations show that the standard version of TRACE does a better job simulating the results than whatever version they used; indeed, we simply cannot arrive at a version of TRACE that will do as poor a job as in the MWW simulations.

However, even at later cycles, in all of the simulations we have presented, TRACE still does not appear to be able to predict the MWW and McQueen et al. (1999) LDT data patterns (although it fits the eye tracking data reasonably well). TRACE always predicts a substantial difference in the response probability to W1 given W2W1 versus N3W1. As we mentioned earlier, however, alternative linking hypotheses might provide a basis for predicting the W1W1 < N3W1 = W2W1 pattern found in the LDT data.

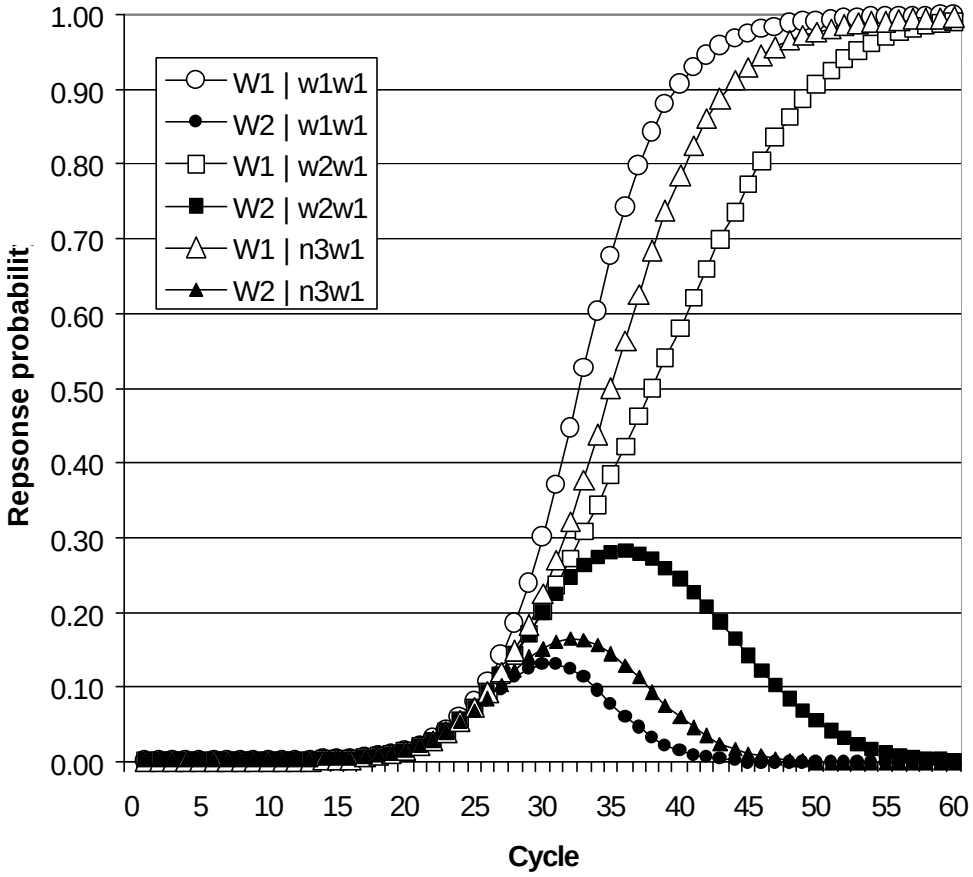

Figure 8: MacTRACE simulations.

Simulating lexical decisions. A "yes" response to W1W1, W2W1, or N3W1 does not entail that W1 was perceived when the response was made. A "yes" response merely indicates that the participant found it probable that the stimulus was a word. The activation of any word could result in a "yes" decision (of course, this is not a novel suggestion; see Grainger and Jacobs [1996], for example, and Tanenhaus et al. [2000] explicitly formulated this hypothesis). Let us follow the implications of this logic for each word stimulus.

For W1W1, we expect W1 to inhibit all other words because of the large bottom-up support for W1, and depending on the response threshold, most if not all "yes" responses will be based on the activation of W1. Given N3W1, the activation of W1 will be slowed due to the subcategorical mismatch. Because the input does not match W2 better than W1, W2 is unlikely to reach threshold. In the case of W2W1, we would expect the activation of W2 to be elevated relative to the W1W1 and N3W1 conditions. This would slow the activation of W1, but the high activation of W2 might reach the threshold with the result that responses to W2W1 would include more early responses. These "yes" responses to W2, when averaged with the later responses to W1, would result in a lower mean given W2W1 than expected.

In order to test the viability of this hypothesis, we developed a simple model to simulate probabilistic lexical decisions when either W1 or W2 reaches threshold (the activations of other words in our TRACE simulations remained low throughout stimulus presentations, and never reached the thresholds we used). We conducted simulations using both the fixation proportions and TRACE activations from the Dahan et al. (in press) study (Experiment 2 from that study). Each simulation was repeated 1000 times for each of the 15 experimental items, across a range of thresholds. The algorithm worked as follows:

1. At the current time step, randomly select W1 or W2
2. If the selected activation or proportion is greater than the threshold
   a. Compute how far above threshold (x = data – threshold)
   b. Generate r, a random number between 0 and 1
      i. If $r \leq x$, "yes"; else, start from step 2 with the other proportion or activation if it has not been checked; else, return to step 1 to check the next time step
      ii. If "yes", stop; else, if at last time step, generate "no"; else, return to step 1 to check the next time step

Figure 9 presents predicted lexical decision latencies as a function of threshold for observed fixation proportions (upper panel) and TRACE activations (lower panel). The predicted latencies given W2W1 and N3W1 are roughly identical and higher than those given W1W1 across a range of thresholds before they diverge in both simulations. This demonstrates the viability of the hypothesis, and both can therefore provide a basis for the lexical decision data patterns found by MWW and McQueen et al. (1999). The simulations based on TRACE activations provide a basis for the W1W1 < N3W1 = W2W1 lexical decision pattern. Over a modest range of thresholds, similar response times (greater than the response time for W1W1) are predicted for W2W1 and N3W1. The simulations based on the fixation proportions provide a remarkable fit to the McQueen et al. (1999) data. McQueen et al. found response times of 340 (W1W1), 478 (W2W1), and 470 ms (N3W1). The predicted response times in the upper panel of Figure 11 are ms from word onset. If we convert them to ms from word offset by subtracting 585 ms (the average duration of stimuli in the Dahan et al. study), the predicted response times at a threshold of .42 are 274, 449 and 442 ms, respectively. These absolute response times are all shorter than those found by McQueen et al., but the differences between them are quite close to those in the McQueen et al. data.

Norris et al. (2000) objected to this logic by claiming that it would predict a bimodal distribution of response times for W2W1, which they did not observe in the McQueen et al. (1999) data. However, a mixture of two distributions need not yield a bimodal distribution. Depending on the means and variability of the distributions, the possible combined distributions range from a clear bimodal distribution to one which cannot be distinguished from a unimodal distribution. Rather than a bimodal distribution, the prediction is that the variability in the W2W1 response time distribution should be greater than that for the N3W1 distribution. McQueen et al. graciously provided us with the raw data from their experiment. We analyzed standard deviation for the W1W1, W2W1, and N3W1 conditions. Consistent with our hypothesis, the mean standard deviations were 162, 201, and 168 ms, respectively, with a significant effect of condition ($F[2,88] = 11.85$, $p < .0001$, $MSE = 1700.8$). Newman-Keuls tests indicated that the standard deviation was greater in the W2W1 condition than in the N3W1 condition at the .05 level. This provides highly suggestive confirmatory evidence for our hypothesis.

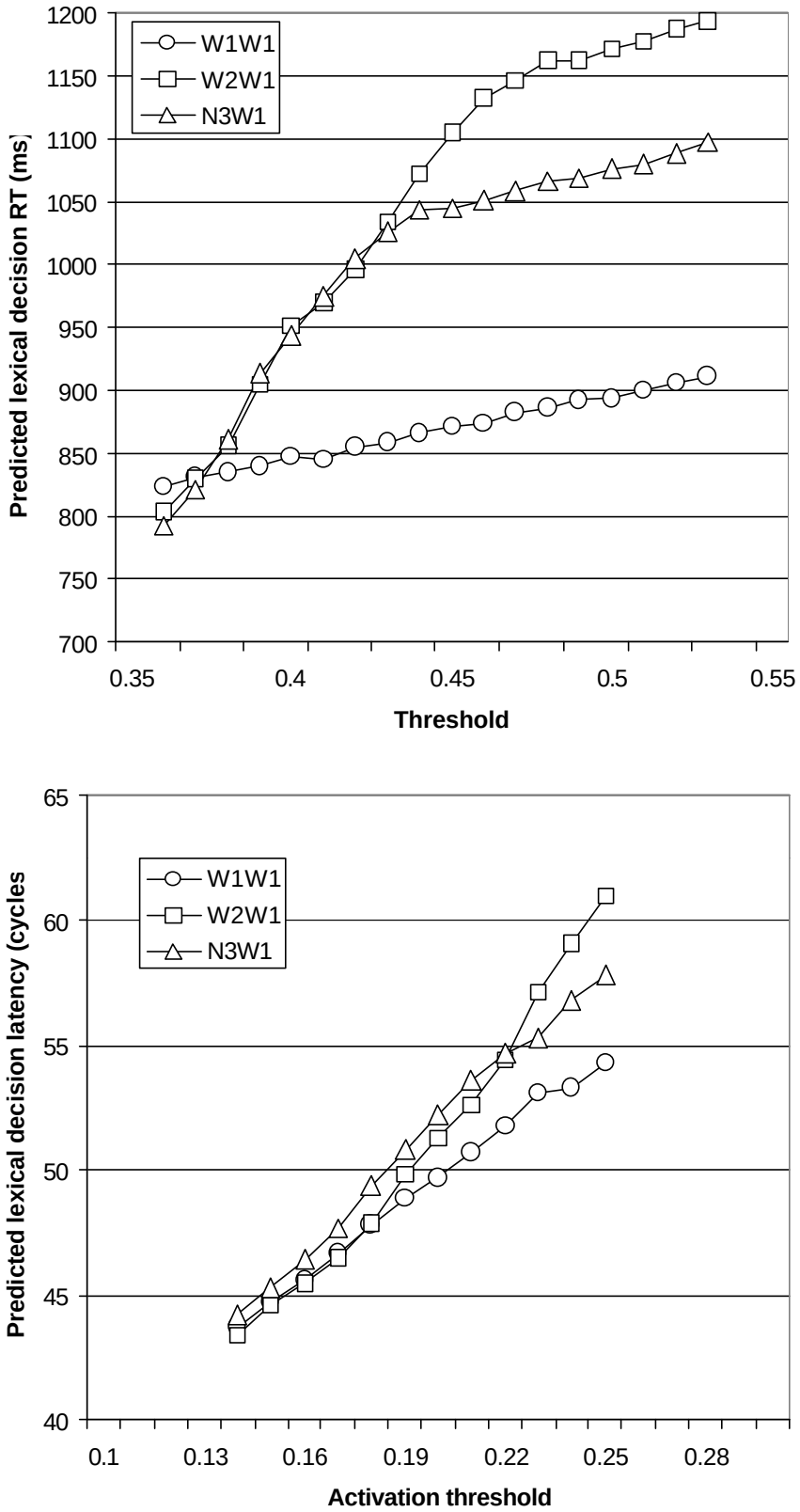

Figure 9: Lexical decision simulations. Upper panel: based on fixation proportions from Figure 3. Lower panel: based on TRACE activations.

Discussion

We have argued that model failures may be due to a number of different factors. Without explicit hypotheses linking the input and task conditions faced by experimental participants to model inputs and outputs, one cannot evaluate apparent discrepancies between models and data. Even with explicit linking hypotheses, model failures may implicate parameters, implementation or architecture, or theoretical assumptions. It is vital to distinguish between these levels, lest one reject a model unfairly.

In the example of the MWW subcategorical mismatch TRACE simulations, we demonstrated that the apparent failure of TRACE was (1) not as extreme as it appeared in the MWW simulations (we could not replicate the pattern shown in their simulation figures) and (2) dependent on the hypothesis linking TRACE activations to lexical decisions. We showed that an arguably simpler linking hypothesis than that used by MWW (one that did not have *a priori* knowledge of target identity) provides a basis for the pattern of lexical decision latencies found by MWW and McQueen et al. (1999). Furthermore, TRACE does a fair job of capturing the trends in on-line eye tracking data, and the hypothesis linking TRACE activations to fixation proportions over time makes testable predictions which we found to be generally accurate (if not precise).

Norris et al. (2000) conducted their own simulations of the subcategorical mismatch data with their "Merge" model and a mock-up of TRACE (with very few nodes, no means of representing temporal relations – "jog", "goj", "jgo", etc., would all activate "jog" equally – and a highly impoverished input representation). They were able to obtain the lexical decision latency patterns found by MWW and McQueen et al. (1999) with both models using a simple threshold linking hypothesis. However, they were only able to fit the results with their miniature TRACE model when (a) they used an algorithmic optimization procedure to set its parameters and (b) when the model was allowed to "resonate" for 15 time steps on each slice of input. They then attribute the MWW TRACE failure to the absence of such resonance in TRACE, and note that the fit obtained with mini-TRACE was "less than optimal" and "not equal to the fit" found with Merge. The implication is that because it was so difficult to set parameters for mini-TRACE, and because the results were not as good as for their Merge model with those parameters, TRACE is to be dispreferred. This conclusion is not sound, however. We have conducted simulations with a similar, though somewhat simpler, "mini-TRACE" and had little trouble finding parameters that work. But more importantly, the mock-up of TRACE they used is simply not comparable to the full version of TRACE. While we have argued that it is important to distinguish between levels of model analysis, this is not to say that these levels are independent. One cannot test theoretical assumptions without an adequate implementation and appropriate parameters.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *Journal of Memory and Language*, **38**, 419-439.

Altmann, G. T. M. and Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73, 247-264.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, **6**, 84-107.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, 42, 317-367.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., and Hogan, E. M. (in press). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cogntive Processes.*

Frauenfelder, U. H. & Peeters, G. (1990). Lexical segmentation in TRACE: an exercise in simulation. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing. Psycholinguistic and computational perspectives* (pp. 50-86). Hove: Erlbaum.

Frauenfelder, U. H. & Peeters, G. (1998). Simulating the time course of spoken word recognition: an analysis of lexical competition in TRACE. In J. Grainger and A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 101-146). Mahwah, NJ: Erlbaum.

Grainger, J. and Jacobs, A. M. (1996). Orthographic processing visual word recognition: A multiple read-out model. *Psychological Review*, **103**, 674-691.

Grosjean, F. (1980). Spoken word-recognition processes and the gating paradigm. *Perception and Psychophysics*, 28, 267-283.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The Neighborhood Activation Model. *Ear & Hearing*, **19**, 1-36.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (submitted). The time course of spoken word recognition in an artificial lexicon.

Marslen-Wilson, W., & Warren P. (1994). Levels of perceptual representation and process in lexical access. *Psychological Review*, **101**, 653-675.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.

McElree, B. (1996). The locus of lexical preference effects in sentence comprehension: A time course analysis. *Journal of Memory and Language*, **32**, 536-571.

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance*, **25**, 1363-1389.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behavioral & Brain Sciences*, **23**, 299-325.

Spivey-Knowlton, M. J. (1996). Integration of Visual and Linguistic Information: Human Data and Model Simulations. Ph.D. thesis, University of Rochester.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, **268**, 1632-1634.

Tanenhaus, M. K., Magnuson, J. S., McMurray, B., and Aslin, R. N. (2000). No compelling evidence against feedback in spoken word recognition. *Behavioral & Brain Sciences*, **23**, 348-349.

Vitevitch, M. S. and Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40, 374-408.

Viviani, P. (1990). Eye movements in visual search: Cognitive, perceptual, and motor control aspects. In E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes* (pp. 353-393). Amsterdam: Elsevier.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception and Psychophysics*, **35**, 49-64.

UNIVERSITY OF ROCHESTER WORKING PAPERS IN THE LANGUAGE SCIENCES – VOL. 2, NO. 1 (SPRING, 2001)

James S. Magnuson and Katherine M. Crosswhite, Editors
Joyce Mary McDonough, Series Editor