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

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 o Tw (Mag oLuce choice rulethe fo327 6

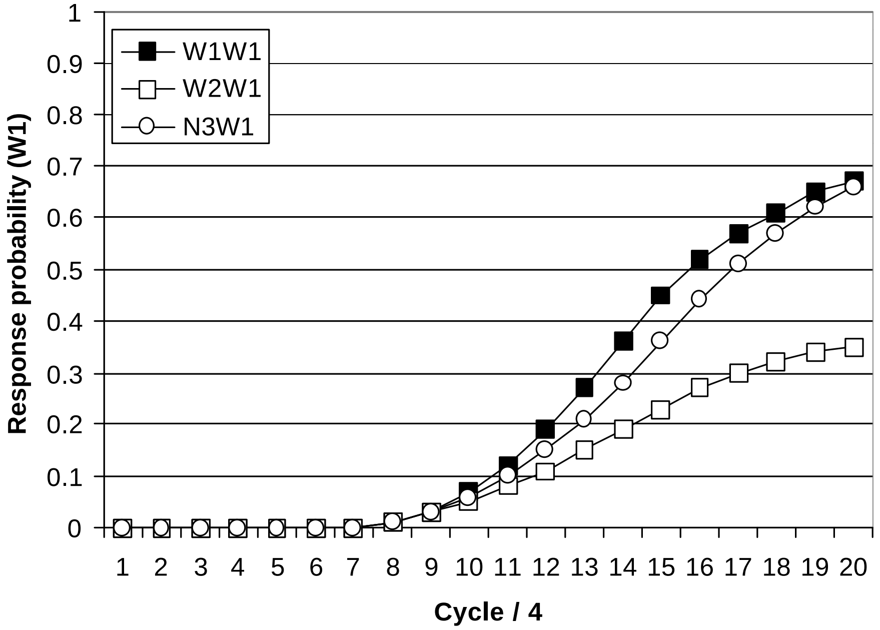


Figure 1: TRACE si

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

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 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.

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 sse ane are rotating ssTin+/D -0.0622 Tc 0.3olex shows

(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.

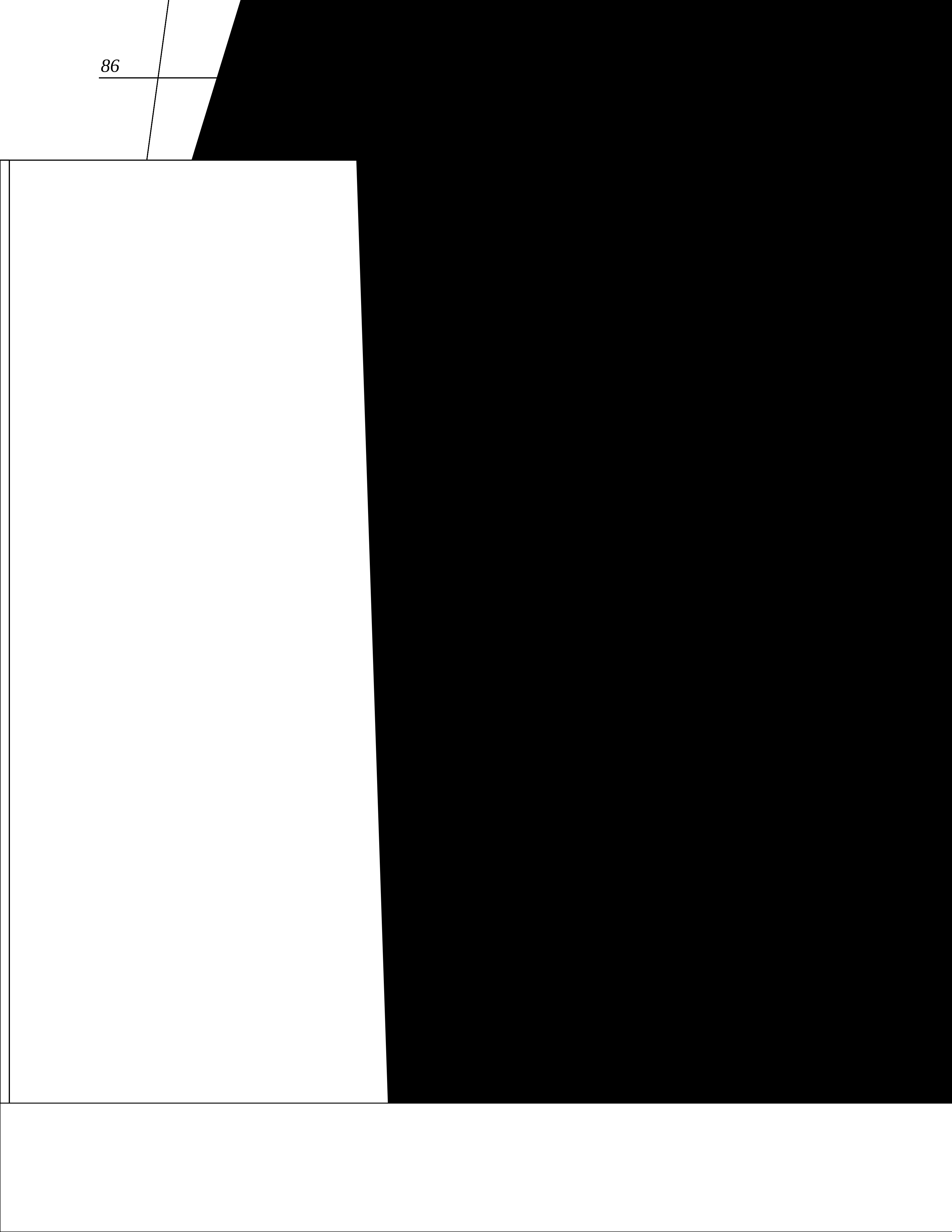
An examination of each individual item showed that almost all items share the same rise time patterns: W1W1 < N3W1 < W2W1, with the difference between WW1 and N3W1 of approximately equal magnitude as the difference between N3W1 and W2W1. The activation of W1 reaches approximately the same peak (between 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.

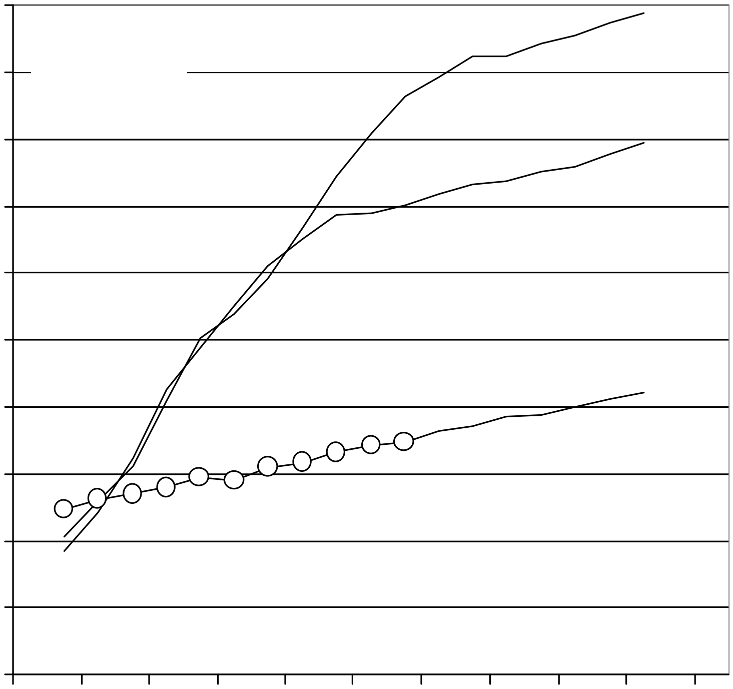
Another possibility is that we used different parameter sets. We used the parameters given by McClelland and Elman (1986) for our initial simulations. MWW report using “the standard parameter settings that give TRACE the appropriate performance in normal word recognition (cf. Frauenfelder & Peeters, 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 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


86



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 modeln of speech processing. Psychosinuiestce and*

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.