Evaluating word reading models at the item level: Matching the grain of theory and data

Mark S. Seidenberg
Departments of Psychology and Linguistics
Neuroscience Program
University of Southern California

David C. Plaut
Departments of Psychology and Computer Science
Center for the Neural Basis of Cognition
Carnegie Mellon University

June 1997

To appear in *Psychological Science*

Abstract

Spieler and Balota (1997) show that connectionist models of reading account for relatively little item-specific variance. In assessing this finding it is important to recognize two factors that limit how much variance such models can possibly explain. First, item means are affecting by several factors that are not addressed in existing models, including processes involved in recognizing letters and producing articulatory output. These limitations point to important areas for future research but have little bearing on existing theoretical claims. Second, the item data include a substantial amount of error variance that would be inappropriate to model. Issues concerning comparisons between simulation data and human performance are discussed with an emphasis on the importance of evaluating models at a level of specificity ("grain") appropriate to the theoretical issues being addressed.

Spieler and Balota (this issue, hereafter SB) correlated the mean RTs of 31 subjects naming 2820 words with performance measures for the same words derived from two connectionist models of word reading: the feedforward network developed by Seidenberg and McClelland (1989, hereafter SM89) and the attractor network developed by Plaut, McClelland, Seidenberg, and Patterson (1996, hereafter PMSP). They found that neither model accounted for much of the variance associated with individual items. This is perhaps surprising insofar as such models have been influential partly because they closely simulate empirical effects observed in numerous studies, such as the frequency-by-consistency interaction (e.g., Taraban & McClelland, 1987). SB's data raise questions about whether the models go beyond more traditional measures such as word frequency, neighborhood density, and orthographic length in contributing to understanding word reading phenomena.

In this brief response, we discuss several issues that must be addressed in assessing the performance of our models in particular and connectionist models in general. Our main point is that the ability of our models to account for item-level data is limited by factors related to both the models and the data. With regard to the models, human performance is affected by several factors beyond the scope of current implementations, including processes involved in recognizing letters and producing articulatory output, and individual differences among subjects. These factors are important to understand but have little bearing on our ac-

count of the phenomena we have already addressed, which largely concern frequency and consistency of spelling-sound correspondences. As discussed below, it is clear that our models could be extended in these additional directions utilizing existing principles. With regard to the data, there are limits on the robustness of the estimates of human performance provided by all experiments including SB's. In fact, there is a large amount of error associated with the item means which we would not expect nor want the models to capture. Finally, even within the limits of the implemented models and the behavioral data, there are important issues about how to relate the two. The simple correlations that SB report may not provide an adequate basis for assessing the models. Questions about the linking assumptions that relate model and data arise in assessing every simulation model and need to be considered carefully.

Relationships Among the Measures

SB's main focus is on the predictive value of traditional, theory-independent measures such as log frequency, Coltheart N (a measure of neighborhood density), and length, compared to measures derived from our models. They report two critical findings: first, the traditional measures together account for more of the variance in mean naming latencies for 2820 words than do either of the measures derived from our models; and second, the model measures account for little additional unique variance (also see Besner, in press; Besner & Bourassa, 1995, for similar regression analyses based on 30 subjects naming 300 words).

The measures from the models (the phonological error score in the SM89 model; settling time in the PMSP model) reflect how effectively the models generate phonological codes for words and nonwords. They are omnibus measures that they reflect the aggregate effects of all of the factors that influence performance. Unlike the traditional measures, the model measures derive directly from a theory of how words are represented and processed. The settings of the weights in the models are determined by properties of the training set, including the frequencies of words and structural relationships among them (e.g., the extent to which they share subword patterns that are pronounced similarly or differently). The models should therefore capture the effects of factors such as log frequency and Coltheart N. Length is more complicated because it affects several aspects of word reading, such as the encoding of visual display and the production of articulatory output, that are outside the scope of current models.

We recomputed the regressions against SB's item means[1], entering measures derived from the models first and then determined how much residual variance was explained by the traditional measures. We also carried out equivalent analyses with a similar dataset acquired by Seidenberg and Waters (1989, hereafter SW). In the latter study (using the same methods as SB), 30 McGill University undergraduates named aloud the 2900 words from the SM89 corpus plus some additional items. The results of both analyses are given in Table 1 (see Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995, for additional regression analyses involving a subset of these data).

Considering the SM89 model first, the phonological error scores account for most of the variance associated with log frequency, and the combination of error score and length leaves very little variance accounted for by log frequency and Coltheart N. This is particularly clear for the SW dataset, where frequency and Coltheart N do not account for significant variance once the SM89 error score and orthographic length are taken into account. The PMSP settling times exhibit a similar pattern although, as SB point out, they account for substantially less variance (see discussion below).

It should be clear why the models do not account for very much variance beyond what is attributable to the traditional factors. The performance of the models should not be *independent* of log frequency or Coltheart N; rather, we expect measures derived from the models to *subsume* these effects. The models do account for a small amount of unique variance, presumably related at least in part to spelling-sound consistency, which none of the other measures captures. Effects of this factor are likely to be small: our

[1]We thank Dan Spieler and David Balota for providing us with their individual item and subject data.

Table 1
Item-Level Variance Accounted for by Various Factors

Spieler and Balota (1997) data (n=2820)					
Factor	Total	After SM89	After SM89+Length	After PMSP	After PMSP+Length
Log Frequency	.0732	.0223	.0164	.0558	.0484
Coltheart N	.1279	.0790	.0051	.1106	.0128
Orthographic Length	.1439	.1108		.1253	
SM89 Error Score	.1007				
PMSP Settling Time	.0333				
Seidenberg and Waters (1989) data (n=2813)					
Factor	Total	After SM89	After SM89+Length	After PMSP	After PMSP+Length
Log Frequency	.0083	.0005[a]	.0000[a]	.0044	.0025
Coltheart N	.0373	.0214	.0007[a]	.0304	.0022
Orthographic Length	.0478	.0350		.0399	
SM89 Error Score	.0291				
PMSP Settling Time	.0143				

Note: Columns headed by "After" indicate amount of variance accounted for by each other factor after partialling out the effects of the named factor(s), where "SM89" refers to the SM89 Phonological Error Score, "PMSP" refers to the PMSP Settling Time, and "Length" refers to Orthographic Length.

[a] Not reliable at p=.05.

theory suggests that such effects will be limited to relatively low-frequency words; for high-frequency items, little matters other than frequency, length, and articulatory factors. Thus, spelling-sound consistency will account for small amounts of variance calculated over the entire corpus (see Treiman et al., 1995) and, in fact, the effects of this factor in behavioral experiments are also quite small (on the order of tens of milliseconds). The fact that the models accurately capture such effects and the interaction between frequency and consistency despite their small size is therefore important.

In summary, the model measures do not fully account for length effects, but when combined with length they largely account for the effects of factors such as frequency and Coltheart N. More importantly, the model measures derive from a theory of how frequency and other effects arise within the lexical system. The residual effects of length remind us that there are aspects of word recognition and pronunciation that are beyond the scope of the implemented models, and the remaining unexplained variance (in the SB dataset) indicates that there may be additional factors to consider, some of which we discuss below.

How Much Variance Is There To Account For?

One striking aspect of the SB data is how much variance is unexplained by any known factor. Entering all measures into a multiple regression, including ones related to properties of the initial phonemes, leaves less than half of the variance explained (also see Treiman et al., 1995). This fact raises questions about how much variance any model should explain given the reliability of the latency data and the robustness of the estimates of the item means.

For the 2791 items in common between the SB and SW studies, the correlation between the item means is a surprisingly low .54. Inspection of the data from the two studies shows that overall mean and standard deviation were much smaller in the SB study (468, 21.6) than in the SW study (570, 45.4). The differences between studies are illustrated in Figure 1, which presents data concerning the benchmark set of words

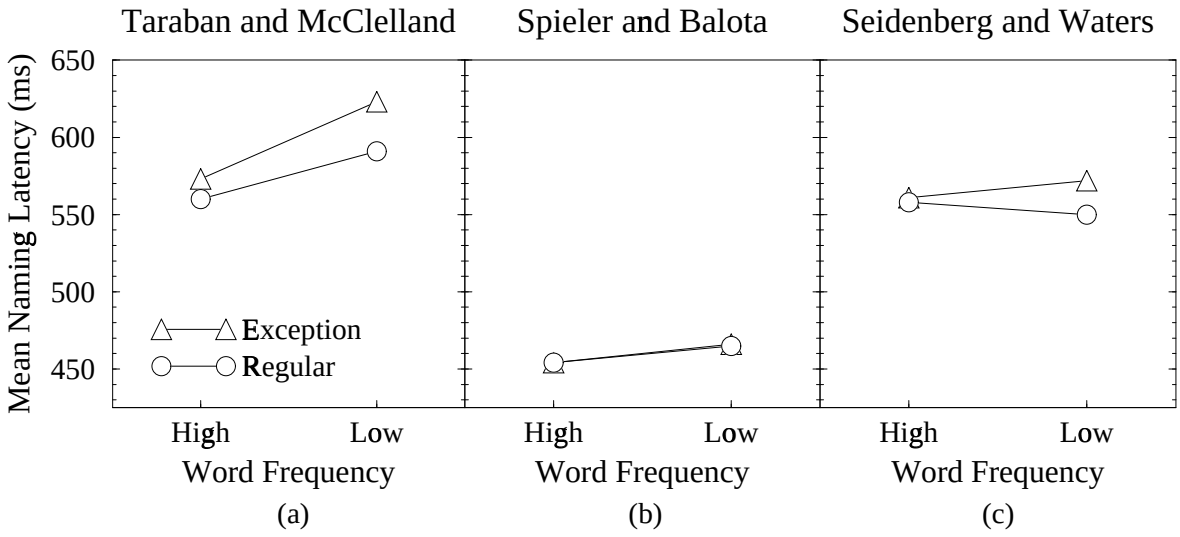

Figure 1. Mean performance on Taraban and McClelland's (1987) high- and low-frequency regular and exception words, for (a) Taraban and McClelland's (1987) subjects, (b) Spieler and Balota's (this issue) subjects, and (c) Seidenberg and Waters's (1989) subjects.

used by Taraban and McClelland (1987, hereafter TM). Whereas the SW data replicate the frequency-by-consistency interaction, the SB data do not, consistent with earlier findings indicating that frequency and consistency effects are larger for slower subjects (Seidenberg, 1985).

These differences between experiments must be due to factors other than stimulus characteristics that SB emphasized. Each model represents a single subject and therefore cannot account for between-subject and between-experiment variance. SB's analyses involve correlating data from each model and means derived from 31 subjects. In fact, if the RTs for each of SB's individual subjects are correlated with the item means for the remaining subjects, the correlations range from $-.0546$ to .222. Thus, the SM89 model correlates more highly with the item means (.317) than all of SB's subjects; the PMSP settling times yield a higher correlation (.182) than all but 6 of SB's subjects.[2]

Although individual differences cannot be captured by a single instantiation of a model, it is clear that they can be addressed within our framework. Consider, for example, the reduction of frequency and consistency effects for faster subjects, evident in Figure 1. The theory behind the SM89 and PMSP models provides an explanation for these results under the assumption that faster subjects have more reading experience; within the SM89 and PMSP models, RTs get faster and effect sizes get smaller with increased training. Effects due to speed-accuracy tradeoffs could also be explained within the existing framework. Subjects who emphasize speed begin articulating their responses earlier than subjects who emphasize accuracy. This could be captured by differences in how long an attractor network is allowed to settle (see Kawamoto & Zemblige, 1992). Thus, although the model measures that SB focused on do not themselves account for these effects, they are compatible with the broader theory that the models approximate.

PMSP also discuss individual differences in naming that are related to variation in the division of labor between pathways within the lexical system. Subjects may vary in the extent to which they rely on the

[2]One possibility is that the mean performance of multple networks would acount for much more variance than any individual network. To test this, we trained 31 different versions of a feedforward version of the PMSP network (PMSP Simulation 2, trained for 400 epochs using a square-root frequency compression), each starting with different initial random weights. After training, the error scores for each network accounted for an average of .0527 of the variance in the SB RTs (range .0450–.0791). The average of their error scores accounted for only slightly more variance: .0854. This is because, unlike the subjects, the performance measures across networks are highly correlated (mean .847, range .800–.881). Thus, initial random weights appears to be an insufficient basis for capturing individual differences among subjects.

orthography-semantics-phonology computation in naming; the strength of input from this pathway affects the behavior of the orthography-phonology computation. Differences in division of labor may also underlie the variability among subjects in studies such as SB's.

Fitting Models to Data

The final issue to consider concerns the assumptions that link statistics derived from our models and behavioral data. The SM89 model yielded a phonological error score that captured effects of frequency and consistency estimated by averaging across the multiple items and subjects employed in individual experiments. Although there was an explicit theory relating error score to latency, the model did not literally model reaction times. PMSP began the exploration of a more direct approach, using the settling times in a recurrent, attractor network to model reaction times (also see Kawamoto & Zemblige, 1992). Settling time in the PMSP attractor network was measured by the processing time required for phonological activation to fully stabilize. Using this measure as an analogue of naming latency is problematic, however, as it assumes that that subjects do not begin to initiate their responses until they have completed the computation of the stimulus' phonological code. This is almost certainly false; under speeded naming instructions subjects may begin to initiate their responses before the entire pattern has been computed. Thus, whereas naming latency measures the time to *initiate* articulation, settling time reflects the *finishing* time for the computation of phonology. This may be why this measure has somewhat less predictive power than the error scores from the SM89 model. Attempts to model time-varying articulatory mechanisms more directly are currently underway (Plaut & Kello, in press).

Clearly there are issues related to the modeling of reaction times that need to be considered further. For example, recent empirical findings (Kawamoto, Kello, Jones, & Bame, in press) suggest that subjects may use a response criterion based primarily on initial phoneme. Some preliminary modeling work (Harm & Seidenberg, 1997) suggests that applying such a response criterion to the phonemic feature representations of a spelling-sound attractor network accounts for more item-level variance (.139 for the SB data, .194 for the SW data) compared with the PMSP attractor network (.033 and .014, respectively). Although these analyses are not definitive, they clearly indicate that better fits to itemwise data can be achieved using existing models if that is the goal.

Conclusions

Our approach to understanding cognitive processes such as word reading is to attempt to articulate general computational principles that, when instantiated in specific domains, give rise to observed performance. Much of the power of the approach derives from the effectiveness of these principles across domains. Better, detailed fits to particular datasets could certainly be obtained by adopting more ad hoc and domain-specific assumptions and mechanisms, but not without a loss of explanatory power. Of course, our ultimate goal is to develop theories that are both theoretically coherent across domains and quantitatively accurate within domains. In the meantime, it is important to maintain an appropriate match between the scope of the theory and characteristics of the behavioral data to be explained. Every computational model is limited in scope and can be falsified merely by considering phenomena that have not yet been addressed. In this respect, research on word recognition is following a normal scientific progression in which the limitations of current models provide the impetus for the next generation of research. At the same time, identifying the limitations of our models and directions for future research also requires recognizing limits imposed by the quality of the data in order to avoid fitting error. Empirical challenges like those of SB provide an important impetus for advances in research, but do not constitute the sole basis on which the adequacy of models should be judged.

Acknowledgments

This research was supported financially by the National Institute of Mental Health (Grants MH47566 and KO2-01188). We thank Marlene Behrmann, Mike Harm, and Jay McClelland for helpful comments and discussions. Correspondence concerning this paper can be addressed either to Mark Seidenberg, Neuroscience Program, University of Southern California, Los Angeles, CA 90089–2520, marks@neuro.usc.edu, or to David Plaut, Mellon Institute 115–CNBC, 4400 Fifth Avenue, Pittsburgh PA 15213–2683, plaut@cmu.edu.

References

Besner, D. (in press). Basic processes in reading: Multiple routines in localist and connectionist models. In P. A. McMullen, & R. M. Klein (Eds.), *Converging methods for understanding reading and dyslexia*. Cambridge, MA: MIT Press.

Besner, D., & Bourassa, D. C. (1995, June). *Localist and parallel distributed processing models of visual word recognition: A few more words.* Paper presented at the Annual Meeting of the Canadian Brain, Behaviour, and Cognitive Science Society, Halifax, N.S., Canada.

Harm, M., & Seidenberg, M. S. (1997). *Phonological representations, reading, and dyslexia: Insights from a connectionist model.* Manuscript submitted for publication.

Kawamoto, A. H., Kello, C., Jones, R., & Bame, K. (in press). Initial phoneme versus whole word criterion to initiate pronunciation: Evidence based on response latency and initial phoneme duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Kawamoto, A. H., & Zemblige, J. H. (1992). Pronunciation of homographs. *Journal of Memory and Language, 31,* 349–374.

Plaut, D. C., & Kello, C. T. (in press). The interplay of speech comprehension and production in phonological development: A forward modeling approach. In B. MacWhinney (Ed.), *The emergence of language*. Mahweh, NJ: Erlbaum.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Seidenberg, M. S. (1985). The time course of information activation and utilization in visual word recognition. In D. Besner, T. G. Waller, & E. M. MacKinnon (Eds.), *Reading research: Advances in theory and practice* (pp. 199–252). New York: Academic Press.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

Seidenberg, M. S., & Waters, G. S. (1989). Word recognition and naming: A mega study [Abstract 30]. *Bulletin of the Psychonomic Society, 27,* 489.

Spieler, D. H., & Balota, D. A. (this issue). Bringing computational models of word naming down to the item level. *Psychological Science.*

Taraban, R., & McClelland, J. L. (1987). Conspiracy effects in word recognition. *Journal of Memory and Language, 26,* 608–631.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General, 124,* 107–136.