Technical Commentary

THE UTILITY OF ITEM-LEVEL ANALYSES IN MODEL EVALUATION: A Reply to Seidenberg and Plaut

David A. Balota[1] and Daniel H. Spieler[2]

[1]*Washington University and* [2]*State University of New York at Binghamton*

Abstract—*Seidenberg and Plaut (this issue) argue that the implications of our analyses (Spieler & Balota, 1997) for the two extant connectionist models of word naming are limited by two factors. First, variables outside the scope of these models influence naming performance, so it is not surprising that the models do not account for much of the variance at the item level. Second, there is error variance associated with large item-level data sets that obviously should not be captured by these models. We point out that there are a number of variables that have been incorporated within the targeted connectionist models that should provide these models an advantage over the simple predictor variables that we selected as a baseline to evaluate the efficacy of the models (e.g., log frequency, length in letters, and number of orthographic neighbors). We also point out that there is considerable consistency across four large-scale studies of item means. Finally, we provide evidence that even under conditions of a standard word-naming study (with a small set of items), simple word frequency, orthographic neighborhoods, and length accounted for more variance than the extant connectionist models. We conclude that item-level analyses provide an important source of evidence in the evaluation of current models and the development of future models of visual word recognition.*

In a previous article (Spieler & Balota, 1997), we reported that the two major connectionist models of speeded-naming performance (Seidenberg & McClelland, 1989 [hereafter, SM89]; Plaut, McClelland, Seidenberg, & Patterson, 1996 [hereafter, PMSP]) captured a relatively small amount of the item-level variance that was produced by 31 participants naming each of the 2,820 words that the models were trained on.[1] We believe this observation is useful in evaluating the currently implemented models because, as the authors of these models assert, one of the attractive aspects of connectionist models is the ability to capture the continuous nature of variables such as frequency, neighborhood structures, and spelling-to-sound correspondence. We also found that traditional variables such as log frequency, orthographic neighborhoods, and orthographic length accounted for more variance than the models alone (see Besner & Bourassa, 1995, for similar analyses on 300 words).

In their reply, Seidenberg and Plaut (this issue; hereafter, S&P) argue that there are two important limitations to our observations. First, S&P suggest that there are a number of factors, such as processes involved in articulation and individual differences across subjects, that produce variability that is clearly outside the scope of the models, and so it is not surprising that the models do not do well at the item level. Second, S&P suggest that "there is a large amount of error associated with the item means that we would not expect nor want the models to capture" (p. 234). In this brief reply, we address both of these issues, and end on a note indicating that we remain committed to the utility of item means in both model evaluation and model development.

THE INFLUENCE OF VARIABLES OUTSIDE THE SCOPE OF THE MODELS

We surely agree with S&P that there are many variables outside the scope of the models that may influence performance, and we demonstrated that coding phonetic onsets accounted for a considerable amount of the variance above and beyond what both the models and the standard predictor variables (log frequency, orthographic neighbors, and length) could account for. However, the emphasis in our report was not simply on how much absolute variance the models captured (although this is also informative), but on how much variance the models captured compared with the three standard predictor variables. We were surprised that the models captured relatively little variance, and in the case of the PMSP model, not as much as the standard predictor variables. This finding was surprising because the models were developed to capture many factors that we did not include as predictor variables. For example, both models were developed, in part, to capture subtle aspects of spelling-to-sound consistency (e.g., *pint* vs. *hint* and *have* vs. *gave*), such as the frequency and density of friends and enemies and their relationship to the frequency of a given target word (e.g., Jared, McRae, & Seidenberg, 1990). Moreover, we grossly coded orthographic neighborhood size as simply the number of words that can be generated by changing one letter in the target (e.g., Coltheart's *N*). One of the advantages of the models is that they should be able to capture the more graded influence of orthographic similarity across words. Thus, because many variables in addition to log frequency, orthographic neighborhood size, and length in letters are actually within the scope of the models, we find it informative that these three factors accounted for as much variance as the models, if not more.

We should also note here that S&P argue that one of our standard predictors (orthographic length) is beyond the scope of their models (e.g., "The residual effects of length are a reminder that there are aspects of word recognition and pronunciation that are beyond the scope of the implemented models . . .," p. 235). Of course, one might ask if all of the effects of length are outside the scope of the models. Consider the correlation between length and the settling times from the PMSP model and between length and the error scores from the SM89 model after frequency and neighborhood size have been par-

Address correspondence either to David A. Balota, Department of Psychology, Washington University, St. Louis, MO 63130, e-mail: dbalota@artsci.wustl.edu, or to Daniel Spieler, Department of Psychology, SUNY at Binghamton, Binghamton, NY 13902-6000, e-mail: spieler@binghamton.edu. Readers may obtain the item means and the predictor variables at the following website: http://www.artsci.wustl.edu/~dbalota/naming.html.

[1] There are alternative computational models (e.g., Coltheart, Curtis, Atkins, & Haller, 1993; Grainger & Jacobs, 1996) that one might test to determine the amount of variance captured. We have focused on the models of Seidenberg and McClelland and Plaut et al. because they naturally acquire spelling-to-sound correspondence effects in a frequency-dependent fashion instead of building in such effects within the models. We welcome requests for our item means so that they may be used in testing alternative models.

tialed out. In the PMSP model, one finds a highly reliable correlation of .083, $p < .001$, but in the SM89 model, there is only a small correlation of .016, $p = .385$. PMSP discussed the predictive power of length in their article (see p. 85). Thus, although all of the effect of length may not be within the scope of the models, at least some of the effect of length is within the scope of the most recent PMSP model.

In sum, we agree that variables outside the scope of the models decrease the amount of variance that the models can account for, and we demonstrated the power of one such factor (phonetic onsets) in our previous article. However, we find it intriguing that variables that are clearly within the scope of the models (e.g., spelling-to-sound consistency, degree of orthographic similarity of the neighbors, and relative frequency of the neighbors to the target) do not provide the models a significant advantage over the standard predictor variables. This is the most surprising aspect of our original observation. Finally, it is unclear why orthographic length should be outside the scope of at least the PMSP model. However, even if length were totally outside the scope of both models, the major thrust of our arguments would not change substantially.

HOW MUCH VARIANCE IS THERE TO ACCOUNT FOR?

S&P point out that a "striking aspect" of our data is "how much variance is unexplained by any known factor" (p. 235). If there is relatively little systematic variance to be explained, then it is not surprising that the connectionist models do a poor job of accounting for variance at the item level. We have four responses to this issue.

First, one might ask how much variance one should expect to capture across the 2,820 items. Given the limited variability in reaction times for the modal items, and all of the potential variables that could influence onset latency (e.g., variations in sensitivity of voice keys across different speakers, and across trials within speakers), we were encouraged that we could account for 21.7% of the variance with the three standard predictor variables, and nearly 42% of the variance when we included coding for onsets. Of course, this is ultimately a matter of the glass being half empty or half full, but given the constraints of such a data set, we believe that the glass is half full.

Second, one might ask if there is something peculiar about the data set we used. Possibly other data sets would produce different patterns. There are now four megastudies of naming performance (Seidenberg & Waters, 1989; Spieler & Balota, 1997; Spieler & Balota, 1998; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995) to address this issue. As shown in the first four columns of Table 1, although there is some variability, there is considerable consistency across these data sets regarding the relative predictive power of the standard variables and the models.

Third, S&P point out that our participants produced relatively fast naming latencies, and this may have minimized the effects of variables such as spelling-to-sound consistency because fast participants produce smaller consistency effects than slow participants (see Seidenberg, 1985). Although we believe that some of the differences in speed may be due to equipment differences (such as display characteristics and sensitivity of voice keys), this is clearly an important issue. S&P display the means from a selected set of items (taken from Taraban & McClelland, 1987) and show that the predicted Frequency × Regularity interaction did not occur for these items in our data set. Interestingly, as shown in S&P's Figure 1, when these same items are pulled from Seidenberg and Waters's (1989) megastudy, which produced considerably slower response latencies than our study, one still finds a failure to produce the pattern found by Taraban and McClelland (i.e., low-frequency regular words produce faster response latencies than high-frequency regular words). In addition, we (Spieler & Balota, 1998) have recently replicated the experiment described in our previous article (Spieler & Balota, 1997) with a group of 29 adults age 60 and older and the same set of 2,820 words. As shown in the third column of Table 1, although the older adults in our 1998 study exhibited slower naming latencies than the young adults in our 1997 study (663 ms vs. 468 ms, respectively), the older adults produced the same pattern of accounted-for variance in the predictor variables and the models. Finally, we conducted a median split (based on a participant's overall response latency) on the (young) participants from our 1997 study and looked at the correlations for fast participants (mean = 434 ms) and slow participants (mean = 501 ms). The estimates of variance accounted for by both the SM89 and the PMSP models were higher for the fast participants (9.4% and 2.7%) than the slow participants (7.2% and 2.4%). This result is clearly inconsistent with the notion that the models will do better for slower subjects. Thus, the relatively fast naming latencies in our 1997 data set do not appear to be a critical variable that diminished the models' ability to capture item-level variance.

Table 1. *Variance accounted for by predictors in four word-naming corpora*

Predictor	SW89	SB97	SB98	TMBR95	TM87
Log frequency	3.3	7.3	12.2	4.6	7.6
Log frequency, Coltheart's *N*, and orthographic length	14.5	21.8	21.5	8.2	8.2
SM89 error scores	6.5	10.1	10.8	5.1	3.1
PMSP settling times	3.0	3.3	2.9	3.3	1.0

Note. SW89 refers to Seidenberg and Waters (1989), SB97 refers to Spieler and Balota (1997), SB98 refers to Spieler and Balota (1998), TMBR95 refers to Treiman, Mullennix, Bijeljac-Babic, and Richmond-Welty (1995), TM87 refers to our replication of Taraban and McClelland's (1987) frequency-by-regularity study, SM89 refers to Seidenberg and McClelland (1989), and PMSP refers to Plaut, McClelland, Seidenberg, and Patterson (1996).

Of course, it is possible that having participants pronounce 2,820 words produced a type of behavior that does not reflect "normal" lexical-processing performance. Thus, it is possible that one might find a different set of correlations if one predicted item means from a more traditional experiment in which participants pronounced 100 to 200 words. In order to test this possibility directly, we conducted a simple speeded-naming study (with 20 participants) using the set of Taraban and McClelland items that S&P identify as being critical tests of the models' performance. As shown in Figure 1, we replicated Taraban and McClelland's original interaction for this restricted set of items. One might expect that the connectionist models would predict item-level variance in this data set better than the standard predictor variables. However, as shown in column 5 of Table 1, the standard predictors do better than the models even for this limited set of items. Thus, the models' poor ability to capture item-level variance is not due to peculiarities of requiring participants to name a large set of items. Even with a relatively small set of items, one finds the same pattern of predictive power.

Of course, one might ask whether it is more appropriate to model selected items from the large megasets or results from experiments with relatively few items. It is possible that the data obtained from a relatively small set of items have idiosyncratic characteristics that are less prevalent in a larger sample of items. For example, it is at least possible that with a restricted set of items, there are subtle cross-item priming influences of particular spelling-to-sound correspondences, and this may contribute to the observed Frequency × Regularity interaction.

Finally, we were not surprised by S&P's observation that the models do a better job in correlating with our item means than do any of the individual participants in our study. Clearly, one would expect considerable variability across participants because of variability in exposure to stimulus words and speech patterns, to mention two examples. We did not address how much variance the models could account for at the individual participant level, and this is precisely why we looked at the means across participants to obtain an estimate of the modal response latency, as in standard cognitive experiments. Our major point is that given the same data set (item means across 31 participants), the three standard measures did at least as well as the models. It is of course possible that 31 runs of the models would be a better way to test the models against the participants' data. However, as acknowledged by S&P in their footnote 2, multiple runs did not improve the predictive power of the models.

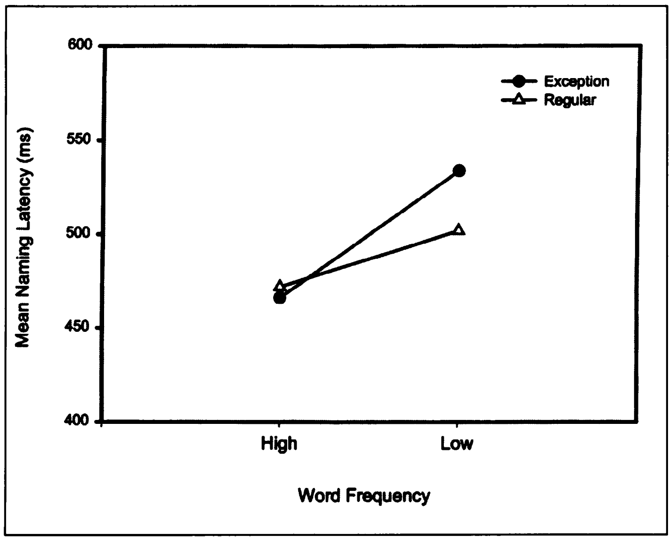

Fig. 1. Mean response latencies as a function of frequency and regularity with Taraban and McClelland's (1987, Experiment 1a) stimuli.

CONCLUSION

We remain confident that item-level analyses provide a valuable complement to traditional factorial studies in the evaluation of extant models of lexical processing. As we noted in our earlier article, there is the potential danger of factor-level analyses producing too much emphasis on variables that account for relatively little consistent variance. In this same light, we do not believe that accounted-for variance should be the only metric for model evaluation and development. Factorial studies of theoretically motivated variables are quite important for making progress in this area. Finally, we remain optimistic about the connectionist enterprise that SM89 and PMSP have espoused. The computational specificity of the models and the ability of the models to extract the effects of a number of important factors in the literature simply via frequency-based exposure to items are important advantages of these models over first-wave metaphorical models. We sincerely hope that our item-level analyses can provide an additional source of helpful guidance in future model development.

Acknowledgments—This work represents a totally collaborative effort and was supported by National Institute on Aging Grant PO1 AG03991 and Training Grant AG00030. We thank David Plaut for his helpful comments on an earlier version of this manuscript.

REFERENCES

Besner, D., & Bourassa, D.C. (1995, June). *Localist and parallel 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, Nova Scotia.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100,* 589–608.

Grainger, J., & Jacobs, A.M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103,* 518–565.

Jared, D., McRae, K., & Seidenberg, M.S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language, 29,* 687–715.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S., & Patterson, K. (1996). Understanding normal and impaired 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]. *Bulletin of the Psychonomic Society, 27,* 489.

Spieler, D.H., & Balota, D.A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science, 8,* 411–416.

Spieler, D.H., & Balota, D.A. (1998). [Item-level variance in the speeded-naming performance of healthy older adults across 2,820 words]. Unpublished raw data.

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.

(Received 9/21/97; Accepted 12/9/97)