Error, Error Everywhere: A Look at Megastudies of Word Reading

Daragh E. Sibley (dsibley@wisc.wdu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Christopher T. Kello (ckello@ucmerced.edu)
Cognitive Science Program, P.O. Box 2039
Merced, CA 95344 USA

Mark S. Seidenberg (seidenberg@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

Two primary methods have been used in studies of word reading: small-scale factorial studies and larger scale "megastudies" involving thousands of words. We conducted comparisons between the two, using the frequency X regularity interaction in word naming as test case. Whereas the effect replicates across small-scale studies, the same results were not obtained using item means from 3 megastudies. Correlations between the megastudies are also relatively low. The considerable error variance in the megastudies limits their use in creating mini ("virtual") experiments. The megastudies yield small but more consistent results using regression analyses that examine specific factors.

Keywords: Word reading; megastudies; methodology.

Introduction

How people read words is one of the most extensively studied topics in cognitive science. This is because of the complexity and importance of the skill, and because it provides a domain for exploring general theoretical frameworks (e.g., Parallel Distributed Processing, Bayesian, dual-route) and computational modeling methods. Extensive data have been acquired using many complementary methods (e.g., behavioral studies of skilled or impaired readers; neuroimaging). Considerable progress has been made, and the work has begun to suggest how reading should be taught (e.g., Rayner et al., 2002).

Although many methods have been used to study reading, most of the primary data are drawn from simple tasks like reading words aloud (naming) and making lexical decisions. These tasks are used to examine how characteristics of words affect processing, including experiential factors like frequency and age of acquisition; semantic properties such as ambiguity and abstractness; and structural factors like length and spelling-sound consistency. For years researchers relied on experiments that factorially manipulated properties of words while attempting to hold other factors constant. These usually involved relatively small numbers of items per condition, because of demands of experimental designs. More recently, researchers have conducted "megastudies" in which latencies are gathered for large numbers of words. In the first study of this type, Seidenberg and Waters (SW; 1989) obtained naming latency and error data for 2,900 words from 30 subjects. Kessler, Treiman, & Mullennix (KTM; 2002) gathered similar data for 2,326 words, and Balota et al. (English Lexicon Project, ELP; 2007) took this approach much further, gathering naming data for 40,481 words from 444 individuals. These large corpora sample the lexicon broadly and permit additional types of data analyses, particularly regression models. Balota et al. have provided an extensive analysis of their corpus, which is an important tool that is freely available on the Internet. The two other corpora are also downloadable and have been used by other researchers.

We were interested in two questions. First, can these corpora be used to conduct "virtual experiments"? Researchers can create factorial-type experiments by sampling words from the corpora. Thus one could generate huge varieties of virtual experiments, while avoiding the demands of collecting new behavioral data. This would be an enormously productive research strategy. The question, however, is whether the large scale studies yield data that is comparable to that obtained in the smaller-scale factorial experiments. Data obtained from subjects who have read hundreds or thousands of words could differ from data obtained in studies of 100 words. We examined this question by looking at whether 5 well-known studies of a common finding, the frequency by regularity interaction, replicate using data drawn from megastudies.

A second, related question concerns how the results of the megastudies compare to each other. Differing methodologies (e.g. recording equipment, subject samples, instructions, and the number and type of stimuli per session) could introduce considerable experiment-specific variance. This is an important consideration, as accounting for item-wise variance in naming latencies has become a criterion for evaluating computational models of reading (Perry, Ziegler, & Zorzi, 2007).

To foreshadow the results, we show that the virtual experiment methodology is problematic: it tends to underestimate effects, creating Type II errors. Item-wise correlations between the megastudies are surprisingly low, indicating considerable error variance. Thus, some effects identified using factorial studies would not have been detected using virtual experiments. The differences between the corpora have important implications for their use in evaluating computational models of reading. Regression analyses yield more consistent results across the megastudies, but some effects are small and hard to detect. We conclude that factorial and megastudy methodologies have different strengths and weaknesses, which suggests using them in a complementary manner.

Virtual Experiments

The frequency by regularity interaction is a standard finding in the reading literature. Seidenberg et al. (1984) found that English words with irregular pronunciations (e.g., HAVE, PINT) produced longer latencies than regular words (e.g., MUST, PINE) only when they were relatively low in frequency. Intuitively, common words are read equally easily, other factors aside; for less common words, irregulars incur a penalty in latency, even for skilled readers. In early studies “regularity” was defined with respect to whether a word’s pronunciation is rule-governed (as in a dual-route model; Coltheart et al., 2001). Later, it was reconceptualized in terms of the degree of consistency in the mapping between spelling and sound (as simulated in connectionist models; Seidenberg & McClelland, 1989). These effects (under either name) have replicated multiple times in different labs using a variety of stimuli. Plaut et al. (1996) provided a formal analysis of the relationship between frequency and consistency in connectionist models.

Our first question was this: will the pattern obtained in the factorial studies replicate if we create virtual studies using latencies for the same stimuli but collected as part of the megastudies? For this analysis we used the following studies: Seidenberg (1985, sets A and B collapsed); Taraban and McClelland (1987); Jared (1997); Paap and Noel, 1991); and Seidenberg et al. (1984). Other studies were not included because of space limitations.

Methods and Results

We recreated each study using the means for the original stimuli, but taken from the megastudy data sets. Items in the original experiment were occasionally missing from a megastudy. For present purposes we simply excluded these items from the analyses, resulting in slightly different numbers of stimuli per condition across experiments and megastudies. However, the number of excluded items was small. We conducted the same item analyses as in the original studies; the full ANOVAs are available from the authors.

Figure 1 shows the results of the Seidenberg (1985) study and the three virtual experiments. The significant frequency X regularity interaction in the original study was marginal using the KTM corpus ($p < .07$) and nonsignficant using ELP and SW. The SW latencies are noticeably faster than in the other data sets, although in the same range as the original study. All of the megastudies show qualitatively similar patterns as the original study, taking into account differences in overall naming speed. ELP produced main effects of frequency and regularity but no interaction; these subjects were also slowest on the higher frequency words, which suggest they treated them more like lower frequency items, perhaps because they were less familiar with them on average. In summary, the basic pattern replicates across the megastudies but the statistical effects are not reliable.

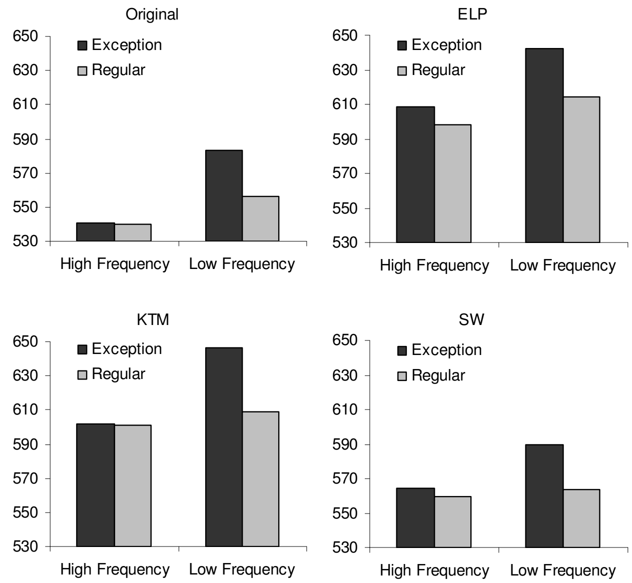

Figure 1: Results for Seidenberg (1985) Experiment

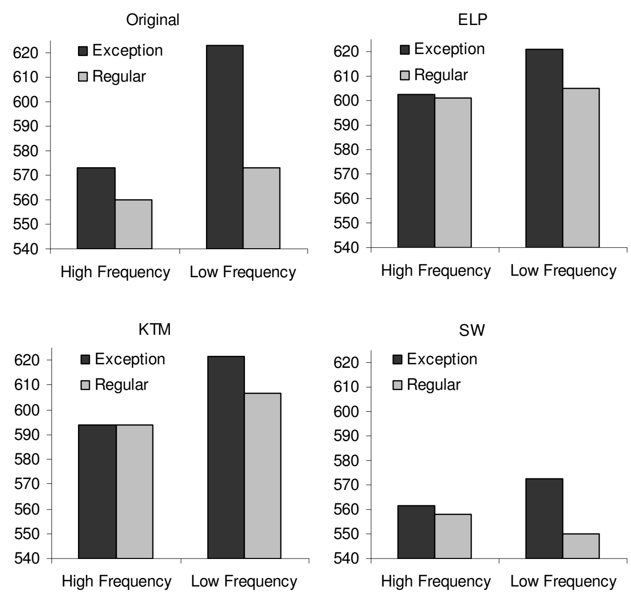

Figure 2: Results for Taraban & McClelland (1987)

Replications of the Taraban and McClelland (1987) experiment yielded similar results (Figure 2). In the original study, there was a significant frequency X regularity interaction (the apparent difference between the two high frequency conditions was nonsignificant). The replications produced similar patterns, but the effects are again not statistically significant (the only significant effect was frequency in the KTM analysis).

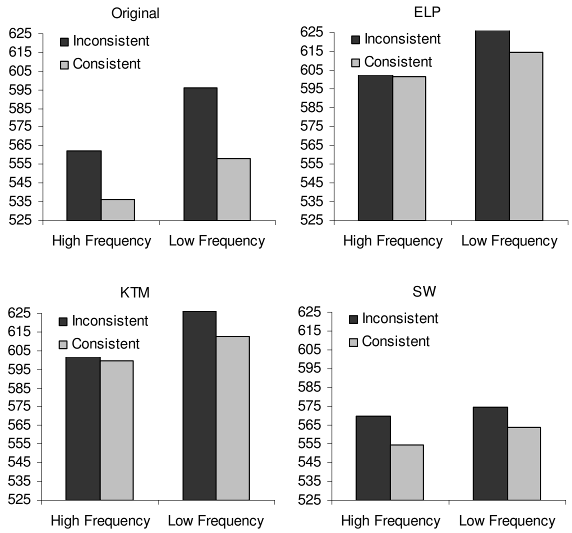

Figure 3: Jared (1997) Results

The Jared (1997) study (Figure 3) compared "consistent" and "inconsistent" words, where the inconsistents included items that had been categorized as either exceptions (such as BREAK) or regular inconsistents (such as PAID, which as the "enemy" SAID) in previous research. The "consistent" items are essentially the same as "regular" words in other studies. This study is widely cited because it produced a consistency effect for "high" frequency words as well as low (however, the frequencies of these HF words were somewhat lower than in previous studies, and the consistency effect was larger for the lower frequency words). In the virtual experiments none of the statistical effects (frequency, consistency, and the interaction) were significant and none reproduced the original latency pattern.

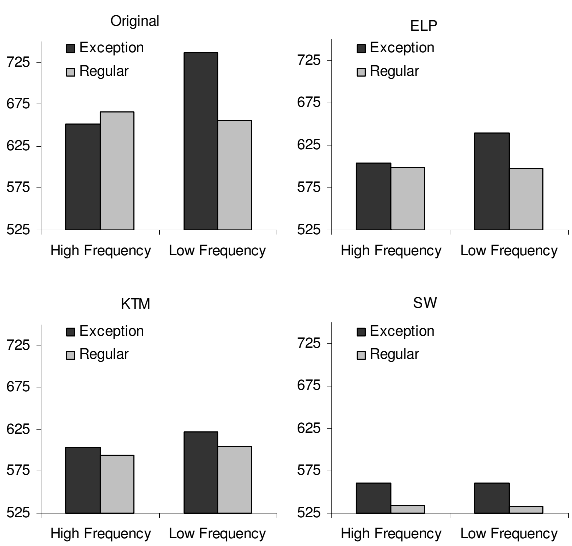

Figure 4: Results for Paap and Noel (1991)

The Paap and Noel (1991) study used a different methodology: subjects performed a secondary task while naming words aloud. Although the study is cited as yielding a frequency X regularity interaction (e.g., Coltheart et al., 2001), the relevant statistical test was not reported in the article. There was a regularity effect for LF words and a reversed effect for HF words (reg > exc). The megastudy replications are variable. ELP produces frequency and regularity effects and a marginal ($p < .07$) interaction. For the other studies only the regularity effect in the SW replication is significant.

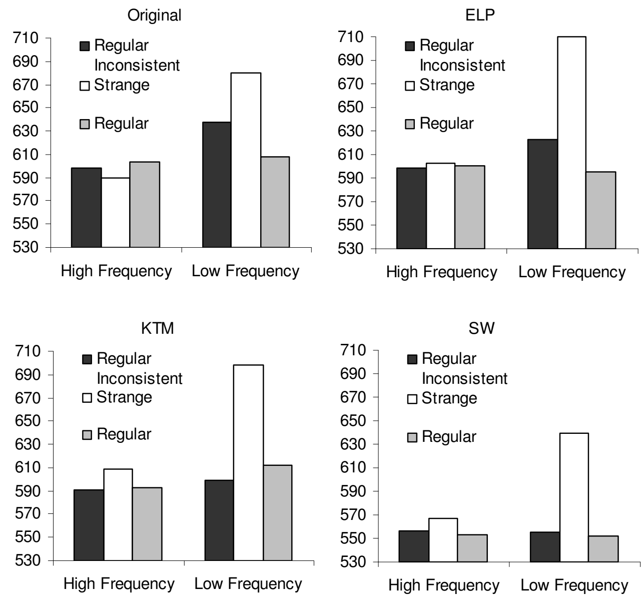

Figure 5: Seidenberg et al. (1984) Results

Finally, Figure 5 shows the results for an early study by Seidenberg et al. (1984). The experiment examined regular words, inconsistent words (such as GAVE) and "strange" words (such as AISLE), which are highly irregular. For the HF words, there were no significant differences between conditions; for the LF words, the order of latencies was strange >> regular inconsistent > regular. This study replicates best. ELP and SW produced the same pattern and statistical results; KTM produces the same pattern except that low frequency inconsistents did not differ from low frequency regulars.

To summarize, we examined whether 5 widely-cited studies of the effects of frequency and regularity/ consistency would replicate if the same experiments were run using data from three megastudies. The results are somewhat disappointing. In some cases, effects do not replicate. In other cases, the virtual experiments show the same pattern as the original, but effects are not borne out by the statistical analyses. Differences in subject speeds across studies also need to be considered. Subjects in ELP and KTM were typically slower than in the original studies, whereas the SW subjects were faster, producing smaller effects.

The freq X regularity/consistency interaction has replicated in multiple small-scale experiments conducted in different labs. The main effects of these variables are also well documented. The fact that the effects do not replicate in the megastudies is therefore a cause for concern. The megastudies yield substantially different results at the level of individual items and condition means.

Item Analyses

Our virtual experiments may not have replicated the original factorial studies for a number of reasons. One possibility is that the megastudy data are not as reliable at the item level as data collected in factorial experiments. This would decrease statistical power and so increase the rate of Type II errors. One clue to the reliability of megastudies is the amount of shared item variance across studies.

Table 1 shows the percentage of item variance that each megastudy accounts for in each of the other megastudies. This is based on naming latencies for the 2,303 words included in all three datasets. Notably, none of the megastudies accounts for as much as half the variance in another study. The remaining variance is attributable to factors other than characteristics of the words themselves.

Table 1: Percentage Of Shared Item Variance (computed using r^2) Across Megastudies

	ELP	KTM
KTM	43.59	
SW	36.11	34.90

One issue may be the use of different stimuli in each study, which can produce list context effects. The ELP includes mono- and multisyllabic words, whereas SW and KTM are only monosyllabic. The ELP and KTM studies were conducted at multiple institutions, in labs that employ different apparatus. In fact, the KTM dataset was gathered to examine variation produced by different voice keys. Finally, there is always subject-wise error: subjects differ in reading ability, attention to the task, and so on.

These and other sources of variance seem to prevent the creation of reliable virtual experiments. As the frequency X regularity analyses showed, different results will be obtained using different corpora and the effects are not robust. The smaller-scale studies produced consistent results, suggesting they provide more reliable estimates.

The relatively low correlations between megastudies affect their utility in evaluating models of word reading (Balota & Spieler, 1998; Seidenberg & Plaut, 1998). Recent computational models have been assessed with respect to item-wise correlations between measures of model performance and mean naming latencies. These correlational analyses involve thousands (Perry, Zeigler, & Zorzi, 2007) or tens of thousands (Sibley & Kello, submitted) of words. Much has been made of the fact that one model accounts for more item-wise variance in naming latencies than others; increasing the amount of variance accounted for is taken as a primary modeling goal (see Coltheart et al., 2001, and Perry et al., 2007). This approach assumes that megastudies are reliable at the item level. However, Table 1 indicates the presence of considerable experiment-specific variance. The amount of item variance that each megastudy predicts in the other megastudies suggests an upper limit for the variance that a computational models should predict. A model that fit a particular dataset more closely would probably be modeling error.

The megastudies have other uses, however. Balota et al. (2004) and Yap & Balota (submitted) used regression to explore how variables such as frequency and regularity affect latencies across large portions of their corpus. These regression analyses offer advantages compared to smaller-scale factorial experiments. The larger data sets allow a broader range of statistical analyses, including ones that examine multiple properties of words simultaneously, or the effect of a factor (such as frequency) while statistically controlling other factors (such as length). Factorial studies use relatively few words per condition because of the need to equate stimuli across conditions with respect to other factors. Regression analyses are not subject to statistical problems associated with treating continuous variables as categorical (Cohen, 1983). This is particularly relevant to factors such as consistency, which is thought to be a graded phenomenon (Seidenberg & McClelland, 1989).

The regression analyses yield more consistent results than the virtual experiments. We ran regression models on the 2,252 words which were included in all three megastudies for which estimates of frequency and consistency were available. We first examined frequency because of its prominence in studies of word reading. The logarithm of American History Dictionary frequency estimates had a statistically significant relationship ($p < .001$) with latency in all of the megastudies. Effect size estimates, using r^2, are 0.091, 0.074, and 0.027 in the ELP, KTM, and SW datasets, respectively. While these effect sizes vary they are all in the same direction and so support similar conclusions.

We also performed regression analyses examining the effect of a word's length in letters on naming latencies. Length was chosen because it is a simple objective measure, with substantial theoretical importance (see Perry et al., 2007 for discussion). Orthographic length accounts for a statistically significant amount of variance in all three megastudies (p's < .001). Estimates of effect sizes are also fairly consistent: 0.159, 0.153, and 0.124 on the ELP, KTM, and SW studies, respectively.

Regression analyses were conducted on four measures of spelling-sound consistency from Yap & Balota (submitted). Their measures examine consistency in the mapping from orthography to phonology (feedforward consistency) and vice versa (feedback consistency) for both onset and rime. Three of the four consistency measures had statistically significant relationships with latencies in all three megastudies. Only feedforward rime consistency had no statistical relationship to any of the megastudies. The size of the relationship between each consistency measure and each megastudy is displayed in Table 3. This table shows that analyses based on any of the megastudies would have yielded the same conclusion: that there is a small but significant effect of spelling-sound consistency in naming aloud. However, the effects are very small. This may be because the variability in these consistency measures is not large across the large corpora because most words in English are consistent. It is also odd that feedforward consistency had no effect, because most experiments on consistency effects manipulated this property.

Table 3: Effect size measure of consistency

	ELP	KTM	SW
Feedforward onset	.012	.022	.025
Feedforward rime	.000	.000	.000
Feedback onset	.027	.033	.032
Feedback rime	.020	.011	.019

Finally we conducted regression analyses examining the interaction of frequency and consistency. Interactions were assessed as suggested by Cohen et al. (2003; see also Yap, 2007): the two relevant main effect variables were first entered into a stepwise regression and then their interaction was entered. However, using this technique only feedback onset consistency had a statistically signifigant effect and only in the ELP and KTM megastudies.

The regression analyses of the megastudies fail to pick up the frequency X regularity/consistency interaction seen in many smaller-scale studies. This discrepancy is a reminder as to why smaller experiments with well-controlled contrasts between conditions are valuable. The English language has relatively few low frequency inconsistent words. As a result, frequencies X consistency interactions are difficult to detect without using experimental designs that oversample low frequency inconsistent words. Although the effects are small in the lexicon as a whole, they are nonetheless theoretically important and thus important to identify using sensitive designs (see Seidenberg & Plaut, 2006).

Conclusions

The present work examined whether data collected in megastudies of word reading can be used to draw reliable conclusions about skilled reading. We first explored whether virtual experiments could be created by determining whether existing studies would replicate using item data drawn from the three data sets. This technique could be used to generate many studies, expediting the research process. We found that 5 prominent studies of word naming failed to produce conclusive results using this method. Given the robustness of the effects in question (e.g., the frequency X regularity interaction), these failures to replicate suggest that the virtual experiment methodology has limited utility.

Regression analyses produced more consistent results across the three megastudies. Frequency and length in letters produced reliable effects in all three data sets. The estimates of the sizes of the effects varied somewhat, as expected given the presence of measurement error. The regression analyses were less successful in picking up effects of consistency and its interaction with frequency. There were significant effects of consistency but they were tiny, and the interaction with frequency was not detectable.

Our results highlight important differences between the two methodologies. The small-scale factorial experiments are useful for identifying factors that affect performance, such as spelling-sound consistency. They intentionally involve creating conditions that are most likely to reveal whether the factor in question has an effect. The studies typically involve relatively few words per condition because of the need to equate stimuli with respect to other properties. Such studies often succeed in identifying robust phenomena, such as the frequency X regularity interaction, using different stimuli tested in different labs with different subjects. Such phenomena often have considerable theoretical interest, of course.

Whereas the small-scale experiments are useful for identifying the *existence* of an effect, they provide little information about its *size*, precisely because they sample biased portions of the lexicon and the number of stimuli is small. If the question of interest is the size of the effect, megastudies are the way to go. However, some additional issues should be noted.

First, sometimes the size of an effect is less important than the fact that it exists. Consistency effects are small across the lexicon as a whole for two reasons: (a) because English is relatively consistent, and (b) because the effect is modulated by frequency. Since most of the inconsistent monosyllabic words are high in frequency, the overall effect is small. However, consistency is theoretically important because it is one of the few phenomena for which competing models of word reading (dual-route and connectionist) make different predictions. Coltheart et al.'s (2001) DRC model treats consistency effects as artifactual, resulting from confounded factors. If such effects are real, however, they provide strong evidence against the DRC approach (see Seidenberg & Plaut, 2006, for discussion and evidence). The size of the effect, however, is of limited interest. It could be small and require examining specific parts of the lexicon. That is one purpose of small, well-designed experiments. Such effects may be difficult to detect in regression analyses based on large data sets.

If the size of an effect is of interest, then analyses of large corpora can be conducted. Effects of factors such as frequency and length can be detected, as in our analyses and others' (e.g., Balota et al., 2004). This is particularly relevant for continuous factors. Estimates of the sizes of the effects vary across corpora because of measurement error, however. This error seems to be substantial because the correlations between megastudies are only moderate in size: no megastudy accounted for even half the variance in another megastudy. This conclusion is consistent with the results of the virtual experiments, which tended to replicate patterns, but did not reach statistical significance.

What can be concluded from these analyses? First, the virtual experiment methodology, though it would have been enormously useful, seems inadvisable. It combines the weakest elements of the two methodologies: the relatively small number of stimuli in the factorial experiments, and the relatively high error variance in the megastudies. This combination is a recipe for Type II errors. Second, there is no methodological Silver Bullet. Each of the methods has strengths and weaknesses. The methods are also relevant to different kinds of questions. The smaller scale factorial experiments can be used to identify factors that are theoretically important but small when considered with respect to the entire lexicon. The megastudies can be used to examine the relative sizes of effects and correlations among factors, modulo the problem of error variance across data sets. Moreover, they can be used with a much broader range of data analysis tools, not merely the simple regression analyses reported here. Thus the methods seem to have complementary strengths and weaknesses, and have complementary functions.

Acknowledgments

D. Sibley was supported by the training grant T32 HD49899 from the NICHD.

References

Balota, A. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis. B., Neely, J. H., Nelson, D., Simpson, G. B., et al. (2007). The English lexicon project. *Behavior Research Methods*, 39(3) 445-459.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*, 283-316.

Balota, D.A., & Spieler, D.H. (1998). The utility of item-level analyses in model evaluation: A reply to Seidenberg and Plaut. *Psychological Science, 9*, 238-240.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249-253.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.

Coltheart, M., Rastle, K., Perry, C., Langdon, R. & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204-256.

Forster, K. I. & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*, 627-635.

Jared, D. (1997). Spelling-sound consistency affects the naming of high-frequency words. *Journal of Memory and Language*, 36, 505–529.

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. Journal *of Memory & Language, 46*, 723-750.

Kessler, B., Treiman, R., & Mullennix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language, 47*, 145-171.

Paap, K. R. & Noel, R. W. (1991). Dual-route models of print to sound: Still a good horse race. *Psychological Research*, 53, 13-24.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ Model of reading aloud. *Psychological Review, 114,* 273-315.

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.

Rayner, K., Foorman, B.R., Perfetti, E., Pesetsky, D., & Seidenberg, Mark S. (2001). How psychological science informs the teaching of reading. Psychological Science In The Public Interest Monograph, 2, 31-74. American Psychological Society.

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

Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior, 23,* 383-404.

Seidenberg, M. S. (1985). The time course of phonological activation in two writing systems. *Cognition*, 19, 1–30.

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

Seidenberg, M., and Plaut, D. C. (1998). Evaluating word reading models at the item level: Matching the grain of theory and data. *Psychological Science*. *9*, 234-237.

Sibley, D. E., & Kello, C. T. (Submitted). A connectionist model of monosyllabic and bisyllabic naming. Submitted to the *European Journal of Cognitive Psychology*.

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

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

Yap, M. (2007). Visual word recognition: Explorations of megastudies, multisyllabic words, and individual differences. PhD thesis, Washington University.

Yap, M. & Balota., D. A. (Submitted). Visual word recognition of multisyllabic words.