5 Megastudies

What do millions (or so) of trials tell us about lexical processing?

David A. Balota, Melvin J. Yap,
Keith A. Hutchison, and Michael J. Cortese

Many disciplines have an agreed-upon knowledge base for study. For cellular neuroscientists, it is the neuron, for geneticists, it is the genome, for some areas of chemistry, molecular interactions are the primary target. The success in these fields is in part due to the accumulation of a well-established set of principles. For example, in each of these domains there is a target knowledge base (i.e., the genome, the periodic table, etc.), which then allows researchers to investigate changes across different contexts and how the system interacts with other systems.

Within cognitive science, one might argue that words are a fundamental building block in psychology. Words have been central to developments in computational modeling (McClelland & Rumelhart, 1981), cognitive neuroscience (e.g., Petersen, Fox, Posner, Mintun, & Raichle, 1988, 1989), memory (Craik & Lockhart, 1972), psycholinguistics (Pinker, 1999), among many other areas. Words are wonderful stimuli because they have a relatively limited set of constituents (e.g., letters/phonemes) that can be productively rearranged to capture virtually all the meaning that humans convey to each other. In this light, one might argue that words, like cells for biologists, are a major building block of cognitive science.

If words are so fundamental to our discipline, then surely we must have accumulated an enormous wealth of information about how humans process words. Indeed, this is largely true. For example, psychologists and psycholinguists have identified many variables that appear to influence speeded lexical processing, including word frequency, familiarity, age of acquisition, imageability, number of meanings, letter length, phoneme length, syllable length, number of morphemes, syntactic class, orthographic neighborhood, phonological neighborhood, frequency of orthographic and phonological neighborhoods, spelling-to-sound consistency, among many others. Given the enormous effort that has been devoted to studying words, one would naturally assume that there is a well-specified set of constraints that one could use to accurately predict processing performance for any set of words. Specifically, there ought to be a standard set of assumptions about lexical processing that researchers have agreed upon.

In this chapter, we review a relatively recent approach to studying lexical processing, which involves developing large databases that are made available for researchers to study across three distinct domains; isolated visual word recognition, semantic priming, and recognition memory. One of the goals of this research endeavor is to help define the common set of principles that researchers can rely upon in better understanding how lexical processing influences critical aspects of cognition. This megastudy approach contrasts with the more traditional approach of factorial experiments targeting specific variables within small-scale studies. We will ultimately argue that progress in this field is going to depend on a judicious combination of targeted factorial studies and large scale databases.

Factorial studies of lexical processing

The vast majority of studies of words have involved standard factorial studies in which investigators cross-targeted variables. For example, one might be interested in the interaction between word frequency and length (i.e., number of letters). Hence, the researcher will select a set of items (typically 10–20) that fit the four (or more) cells of the experimental design by crossing word frequency and length.

As noted, the standard factorial approach has yielded a wealth of knowledge. However, there are also some limitations. First, a critical assumption of this approach is that one can equate stimuli on all other relevant variables to fit the critical cells within such designs. Given the plethora of variables available, this is clearly a daunting task (see Cutler, 1981, for a discussion of this point). Second, one needs to worry about list context effects. Specifically, it is possible that by loading up on a given variable, one may actually be modulating the effect of this variable. For example, if one is interested in spelling-to-sound regularity effects, it is possible that other similarly spelled words could influence the obtained effects (consider a list which contains both HINT and PINT; see Seidenberg, Waters, Sanders, & Langer, 1984). Indeed, there is clear evidence of overall list context effects across a number of variables in the literature (see Lupker, Brown, & Colombo, 1997; Monsell, Patterson, Graham, Hughes, & Milroy, 1992; Zevin & Balota, 2000). Third, most variables are not categorical, but are continuous in nature. Moreover, it is unlikely that variables are linearly related to the behavior of interest. By arbitrarily setting a categorical boundary for a variable that is nonlinearly scaled (e.g., word frequency), one may either magnify or diminish the influence of the variable. Along these same lines, one loses statistical power by turning a continuous variable into a categorical variable (see Cohen, 1983; Humphreys, 1978; Maxwell & Delaney, 1993). Finally, one needs to worry about experimenter biases in selection of items. Forster (2000) has demonstrated that experimenters have implicit knowledge about how lexical variables drive performance in a given task. Hence, it is possible that such knowledge may inadvertently influence item selection.

The megastudy approach

A complementary approach to factorial designs is to let the language define the stimuli, as opposed to selecting stimuli based on a limited set of criteria. This indeed is the megastudy approach reviewed here. If there is an agreed-upon useful database, then researchers may use this dataset to explore the influence and interrelationships amongst targeted variables and also test contemporary theoretical models in a more continuous manner instead of the categorical manner which has dominated model development.

Of course, there are also potential limitations to the megastudy approach. One concern is that researchers may exploit the dataset. For example, one could resample multiple times from the large dataset and capitalize on chance to find 'significant' effects of theoretical interest. We believe that the normal peer-review process is sufficiently sensitive to such sampling possibilities, and have not seen megastudy databases misused in this manner. A more important problem is that the large databases may not be stable enough to detect more subtle effects. That is, there may be sufficient error variance in the databases to decrease sensitivity to variables that appear to be well-established in the factorial literature. Interestingly, this possibility has been raised recently by Sibley, Kello, and Seidenberg (2009). Because of the potential importance of this concern, we will briefly address this issue.

Sibley et al. (2009) argued that one should be cautious in relying on megastudy databases because these databases may not be sensitive to more subtle manipulations in the literature that have important theoretical implications. To pursue this issue, they tested the adequacy of megastudies for finding an important interaction between spelling-to-sound consistency/regularity and word frequency. Specifically, high-frequency words are typically influenced less by spelling-to-sound consistency/regularity than are low-frequency words. They correctly argued that these variables have been critical in the development of models of word naming. Therefore, Sibley et al. selected the items from published studies investigating these variables, and attempted to determine if the same pattern would be observed when the item means were obtained from the megastudies. Sibley et al. examined the stability of consistency effects across four different datasets, and indeed there was variability across the datasets with respect to producing the pattern. Here, we simply investigate the stability of the effects using the English Lexicon Project (ELP), because this has been the most well explored database to date and includes a wide range of both monosyllabic and multisyllabic words. At its completion, the ELP was by far the largest megastudy (approximately 20 times larger than other English datasets), although as described below, there are recent databases approaching its size. Finally, this dataset is readily available for researchers to access via the website (http://elexicon.wustl.edu), along with a search engine that affords access to a rich set of lexical characteristics.

In their paper, Sibley et al. (2009) focused on mean *raw* item naming response time (RT) data. However, in the ELP, different participants contribute to the

mean RT of any item, and it is therefore more appropriate to look at *z*-scored RTs instead of raw RTs, as suggested by Balota, Yap, Cortese, Hutchison, Kessler, Loftus, Neely, et al. (2007). In this way, no subject disproportionately influences the item means. Of course, it is also useful to look at the accuracy data. To explore this, we recently used the ELP to conduct analyses similar to Sibley et al.'s, and the collective results are remarkably consistent with published studies. For example, Seidenberg (1985) observed an interaction between spelling-to-sound consistency and word frequency. In our analysis of the Seidenberg items taken from the ELP dataset, the mean *z*-scores clearly were indeed in the same direction, albeit non-significant, while the accuracy data was significant and clearly replicated the Seidenberg (1985) pattern. Turning to Seidenberg et al. (1984), the ELP dataset again produced the same reliable interaction in all three dependent measures (i.e., raw RTs, *z*-scored RTs, accuracy). Turning to Papp and Noel (1991), who did not report the results from statistical tests, the ELP produced an interaction in the same direction for *z*-scores ($p = .07$), and the interaction was again reliable in accuracy in the predicted direction. Jared (2002) observed main effects of consistency and word frequency with her stimuli, with no interaction. This pattern was reliably replicated in the accuracy data of the ELP dataset with her items, and also in the pattern of mean z-scores and raw RTs. Finally, Taraban and McClelland (1987: 613) did *not* report a reliable frequency by regularity interaction for their stimuli, but in separate tests did report a reliable effect of regularity for low-frequency words, but not for high-frequency words. Indeed, in the ELP, there is a regularity effect for Taraban and McClelland low-frequency words in both accuracy and *z*-scores, but not for high-frequency words. In sum, we view the data from the ELP as being remarkably consistent with the critical studies that have manipulated both spelling-to-sound regularity and word frequency. Instead of only questioning the ELP database, it would be useful for the field to test the reliability of standard effects in the lexical processing literature across different institutions for a baseline. In fact, given the likelihood of idiosyncratic effects of list context and voice-key measurement issues, we were surprised (and pleased) by the level of stability. Clearly, based on the above analyses, it appears that the ELP does quite well in producing the standard effects regarding spelling-to-sound regularity/consistency and word frequency.

One might also ask if other standard interactions are observed in the ELP database. As indicated in Balota et al., (2004) and Yap and Balota (2009), many standard interactions reported in the literature (i.e., length by frequency, orthographic N by frequency, orthographic N by length) are well-replicated in the ELP. Most importantly, if the ELP had an extraordinary amount of error variance, one might expect little variance in the total dataset to be accounted for by standard lexical variables. However, this is clearly not a problem in the ELP. For example, Balota et al. accounted for 49% and 42% of the variance in speeded pronunciation and lexical decision latencies respectively for monosyllabic words. Moreover, Yap and Balota (2009) accounted for over 60% of the variance for all monomorphemic multisyllabic words with standard predictors. Because it is likely that there are still unknown variables and possible better ways of conceptualizing

current variables (i.e., nonlinear functions), the current estimates may underestimate the amount of variance accounted for (see Rey, Courrieu, Schmidt-Weigand, & Jacobs, 2009, for further discussion).

Of course, the stability of these datasets is an important issue for testing current computational models. If indeed the datasets are not stable/reliable, then the utility of these datasets would be minimized. Indeed, Spieler and Balota (1997) were initially surprised by the relatively small amount of variance captured by standard computational models. For example, for the monosyllabic dataset, 10.1% was captured by the Seidenberg and McClelland model, and 3.3% was captured by the Plaut, McClelland, Seidenberg, and Patterson (1996) model. In contrast, word frequency, orthographic neighborhood size and length accounted for 21.7% of the variance. Hence, it is not the case that one can simply dismiss the poor fits due to error-prone datasets.

Thus, in developing models of word recognition, researchers have become more interested in using accounted for variance as one useful (but not only) metric of evaluating model performance. In addition, Perry, Zorzi, and Ziegler (2010) have labeled an additional metric called the Yap and Balota criteria, which basically involves the *specific* proportion of variance in model performance accounted for by different variables (also see Sibley and Kello, this volume, Chapter 2). Of course, there are multiple ways of evaluating model adequacy. The point we are emphasizing here is simply that a possible lack of stability of the megastudy datasets is not a reason to dismiss this useful constraint on model development (see Adelman, Marquis, & Sabatos-DeVito, 2011, for further discussion of explained and unexplained variance in word recognition performance).

We shall now turn to a selective review of what we have learned from the megastudy approach at a more empirical level. We will first discuss studies of isolated word recognition, which are by far the most well investigated. We will then turn to more recent developments in the domains of semantic priming and episodic recognition memory performance.

Isolated word recognition performance

In addition to providing a testbed for evaluating computational models of visual word recognition, there are three additional contributions from megastudies to better understand lexical processing. First, these datasets allow researchers to rigorously evaluate the strength of relatively novel variables that theoretically should modulate word recognition. Second, the databases allow for one to compare the relative predictive power of competing metrics. Third, the datasets allow for a finer-grained assessment of the functional relationships (e.g., linear vs nonlinear) between lexical variables and word recognition performance.

Evaluating the influence of novel variables

Megastudies are very useful for benchmarking new variables by evaluating whether they account for additional variance above and beyond traditional variables. In an

early example of such work, Treiman, Mullennix, Bijeljac-Babic, and Richmond-Welty, (1995) explored readers' sensitivity to the consistency of spelling-sound mappings at different grain sizes (Ziegler & Goswami, 2005). A word is considered consistent if its pronunciation matches that of most similarly spelled words. For example, PINT is inconsistent because the pronunciation of its rime (vowel and following consonants, i.e., -INT) conflicts with that of similarly spelled words (e.g., HINT, MINT, TINT). At the time the Treiman et al. paper was published, the dominant view was that spelling–sound relations were most appropriately described at the level of graphemes and phonemes, and Treiman et al. were interested in whether consistency defined for higher-order units (e.g., rimes) was also able to predict speeded pronunciation performance. On the basis of regression analyses of two independent megastudies (consisting of 1327 and 1153 words respectively), they demonstrated that the consistency of higher-order rime units indeed reliably accounted for pronunciation variance, after the consistency of individual graphemes and other variables were controlled for. To strengthen the general conclusions from the megastudies, Treiman et al. also conducted additional factorial studies where rime consistency was manipulated. Importantly, the results from these studies converged nicely with the findings from the megastudies, suggesting that large-scale and factorial studies provided complementary perspectives on phenomena of interest.

The Treiman et al. (1995) study focused on monosyllabic consonant-vowel-consonant (CVC) words. Chateau and Jared (2003) carried out a megastudy (including 1000 words) where they compared the consistency of various orthographic segments in six-letter disyllabic words. Specifically, they obtained measures of spelling-sound consistency for simple (i.e., C_1, V_1, C_2) and higher-order (i.e., C_1V_1, V_1C_2) orthographic segments in the first and second syllables. In addition, they computed consistency for the BOB (body-of-the-BOSS; Taft, 1992) which includes the first vowel and as many following consonants to form a legal word ending (e.g., the BOB for VERTEX is ERT). They found that the consistency of the BOB and the second-syllable vowel predicted pronunciation performance, confirming that readers were sensitive to the consistency of multiple grain sizes when pronouncing words aloud.

Other theoretically motivated variables whose validity have been evaluated using megastudy data include imageability (Cortese & Fugett, 2004), age of acquisition (Cortese & Khanna, 2008), semantic richness (Yap, Tan, Pexman, & Hargreaves, 2011a), a new measure of orthographic similarity called Levenshtein Orthographic Distance (Yarkoni, Balota, & Yap, 2008), a new measure of phonological similarity called the Levenshtein Phonological Distance measure (see http://elexicon.wustl.edu), contextual diversity (i.e., the number of contexts a word appears in; Adelman, Brown, & Quesada, 2006), phonographic neighborhood size (i.e., the number of neighbors that are both orthographic and phonological; Adelman & Brown, 2007), and Sensory Experience Rating (Juhasz, Yap, Dicke, Taylor, & Gullick, 2011), a new variable motivated by the grounded cognition framework which indexes the degree to which a word evokes sensory/perceptual experiences. The general strategy is to assess the extent to which a

novel predictor accounts for unique variance in megastudies, after other correlated variables have been controlled for. While a full description of these studies is outside the scope of this chapter, the studies listed above have shed light on the role of semantic variables on word recognition (Cortese & Fugett, 2004; Cortese & Khanna, 2008; Juhasz, Yap, Dicke, Taylor, and Gullick, 2011; Yap et al., 2011a), the influence of a new orthographic distinctiveness metric that can be used for long words and outperforms the traditional measure of orthographic neighborhood size (Yarkoni et al., 2008), and the superiority of contextual diversity (Adelman et al., 2006) and phonographic neighborhood size (Adelman & Brown, 2007) over raw word frequency and orthographic neighborhood size respectively. Finally, it is noteworthy that the megastudy approach has also been used to provide evidence *against* the reliability of a new variable (see Kang, Yap, Tse, & Kurby, 2011, for an example).

Comparing competing metrics

Megastudies are also ideal for adjudicating between competing measures of the same construct. For example, word frequency is one of the most studied variables in cognitive science. Although many frequency counts are available, most researchers unfortunately continue to rely on the Kučera and Francis (1967; KF67) norms, which are dated and based on a relatively small corpus of written texts. Brysbaert and New (2009), using lexical decision data from recently published large-scale data, compared a number of frequency counts, including KF67, HAL (Hyperspace Analog to Language; Burgess & Livesay, 1998), CELEX (Center for Lexical Information; Baayen, Piepenbrock, & van Rijn, 1993), TASA (Touchstone Applied Science Associates; Zeno, Ivens, Millard, & Duvvuri, 1995), and BNC (British National Corpus; Leech, Rayson, & Wilson, 2001). KF67, CELEX, TASA, and BNC are based on written texts, while HAL is based on internet newsgroup postings. The proportion of variance each frequency measure accounted for in lexical decision performance from the ELP was used as a criterion of its quality. These were all evaluated against an intriguing new frequency measure (SUBTL) based on a 50-million word corpus comprising film and television subtitles. A subtitle-based corpus possesses the advantages of being more reflective of day-to-day spontaneous language exposure and is also relatively easy to accumulate. Indeed, the analyses conclusively demonstrated that KF67 frequency was clearly the worst measure (replicating studies by Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Zevin & Seidenberg, 2004), whereas subtitle-based frequency accounted for more variance than the other leading frequency measures. Interestingly, Yap, Balota, Brysbaert, and Shaoul (2011a) have also shown that a measure of rank frequency (simply the rank order of word frequency values instead of their actual frequency, see Forster, this volume, Chapter 3) predicts very similar amounts of variance. In the examples in this section, word frequency was the construct of interest but in principle, a similar strategy can be used to compare different instantiations of other constructs. Indeed, this is precisely what Yarkoni et al. (2008) reported in

their comparison of the new Levenshtein Distance measure and the standard orthographic N measure (see Davis, this volume, Chapter 9, for a discussion of orthographic neighborhood structure).

Exploring functional relationships amongst variables in word recognition

Large-scale data have been used productively to explore the functional relationships between lexical variables and word recognition performance. For example, word recognition researchers have tended to focus on linear relationships between variables and response times, but it is clear that non-linear contributions also need to be taken into account (see Baayen, Feldman, & Schreuder, 2006).

Consider the relationship between word length in letters and response times. It is commonly assumed that the relationship between length and word recognition performance (as reflected by tasks such as lexical decision, speeded pronunciation, perceptual identification, and eye tracking) is linear, and that word recognition latencies increase monotonically as a function of length. However, this view is complicated by inconsistent results across tasks and studies. Specifically, some studies find inhibitory effects (length and response times positively correlated), others find facilitatory effects (negative correlation), while yet others yield null effects (see New, Ferrand, Pallier, & Brysbaert, 2006, for a review). New et al. explored this inconsistency by conducting regression analyses on a dataset of lexical decision latencies for over 33,000 words (ranging from 3 to 13 letters) from the English Lexicon Project (Balota et al., 2007). They observed an intriguing U-shaped relationship between length and lexical decision latencies, whereby length was facilitatory for 3–5-letter words, null for 5–8-letter words, and inhibitory for 8–13-letter words. Hence, to the extent that different investigators are using stimuli of different lengths, this U-shaped relationship provides a partial explanation for the varied results across different experiments. More recently, Yarkoni et al. (2008) have suggested that this nonlinear function may be accommodated by differences in the orthographic neighborhood characteristics as reflected by the novel Levenstein Distance metric, discussed in the previous section.

The functional form of the relationship between word frequency and word recognition latencies has also been receiving considerable attention in the literature (see Forster, this volume, Chapter 3). Traditionally, researchers have assumed that a linear relationship exists between the logarithm of frequency and recognition times. However, recent models have begun to make explicit predictions about the form of the frequency effect (see Adelman & Brown, 2008, for a review). For example, Norris' (2006) Bayesian Reader model predicts a logarithmic relationship between word frequency and lexical decision latencies; this model conceptualizes readers as optimal Bayesian decision-makers who use Bayesian inference to combine perceptual information with knowledge of prior probability during word recognition. In contrast, Murray and Forster's (2004) serial search model of lexical access predicts that RTs will be 'directly related to

the rank position of a word in a frequency-ordered list, not to its actual frequency or to any transform of it' (p. 723). Finally, instance-based models, which are predicated on the assumption that each encounter with a word leaves a memory trace (e.g., Goldinger, 1998; Logan, 1988), predict that frequency and latencies are represented as a power function. Using large-scale datasets, Adelman and Brown (2008) evaluated the rank frequency function against other functional forms (e.g., logarithmic and power functions) and concluded that the empirical data appeared to be most consistent with some versions of the instance-based models (but see Murray & Forster, 2008). Clearly, without these large-scale databases one would not be able to test the functional form of this theoretically important relationship.

Identifying the unique predictive power of targeted variables

The ultimate goal of factorial studies is to afford a better understanding of the influence of theoretically motivated variables. This goal has been somewhat elusive, for the various reasons discussed in the Introduction. As noted, it is difficult to control for the many variables that have been shown to influence word recognition (Cutler, 1981), since many of these variables are correlated. Megastudies minimize these problems by having the language, rather than the experimenter, select the stimulus set, and using regression analyses to control for correlated variables.

The studies by Treiman et al. (1995) and Chateau and Jared (2003) described earlier exploit this approach for exploring spelling-to-sound consistency, but Balota et al. (2004) were the first to explore the effects of a comprehensive array of variables on word recognition. Specifically, they examined the unique predictive power of surface variables (phonological features in the onsets), lexical variables (e.g., measures of consistency, frequency, familiarity, neighborhood size, and length), and semantic variables (e.g., imageability and semantic connectivity) on word recognition performance for virtually all monomorphemic monosyllabic words. They also compared lexical decision to pronunciation data to study task-dependent effects, and young adult to older adult performance to study the effects of aging. Space limits preclude a full description of the study, but Balota et al. were able to demonstrate that the influence of many variables were modulated by the nature of the task, hence shedding light on a number of empirical controversies. For example, surface variables, length, neighborhood size, and consistency accounted for more variance in pronunciation, compared to lexical decision, because of the pronunciation task's emphasis on generating phonology. In contrast, word frequency and semantics better predicted lexical decision, because of lexical decision's reliance on familiarity-based information for discriminating between familiar words and unfamiliar nonwords (e.g., Balota & Chumbley, 1984). There was also an interesting age-related dissociation where older adults were more influenced by objective frequency while younger adults were more influenced by subjective frequency (i.e., subjective ratings of a word's frequency; Balota, Pilotti, & Cortese, 2001), suggesting that standard

frequency estimates based on written texts may be better tuned to the older adult lexicon.

To a large extent, the visual word-recognition literature has been overwhelmingly dominated by the study of monosyllabic words, because these are relatively simple stimuli to work with. However, monosyllabic words only constitute a small minority of a person's lexicon, and it is unclear if behavioral effects reported for monosyllabic words generalize to longer multisyllabic words. The Chateau and Jared (2003) study discussed earlier is noteworthy for being the first large-scale exploration of multisyllabic words. However, they were predominantly interested in how consistency influences the pronunciation of six-letter disyllabic words.

Using data from the ELP, Yap and Balota (2009) extended the work by Balota et al. (2004) and Chateau and Jared (2003) by using hierarchical regression analyses to identify the effects of surface, lexical, and semantic variables for 6115 monomorphemic multisyllabic words. In addition to considering the role of traditional variables (e.g., frequency, length, orthographic neighborhood size), they also explored variables specific to multisyllabic words (e.g., stress pattern, number of syllables). Importantly, processing of multisyllabic words does not appear to radically differ from the processing of monosyllabic words. However, there were also a number of surprising differences. First, onset characteristics, which account for considerable variance in monosyllabic pronunciation, are far less influential in multisyllabic pronunciation. This may suggest differences in the emphasis on onsets during production in the monosyllabic words compared to the more complex multisyllabic words. Second, number of syllables was positively correlated with both lexical decision and pronunciation latencies, even after controlling for a host of variables. This suggests that multiple codes mediate lexical access and output processes, and the syllable is one of those codes. Third, the analyses included novel measures of orthographic and phonological distinctiveness (Levenshtein measures; Yarkoni et al., 2008) to complement traditional measures (e.g., orthographic neighborhood size) that are not optimized for long words. The interesting finding here is that words which have relatively close visually and phonologically confusable neighbors produced faster response latencies in both naming and lexical decision performance, which is inconsistent with a simple competitive lexical identification process.

Stable individual differences revealed in the ELP

Interestingly, there has been relatively little work in the visual word recognition literature on the reliability of measures of lexical processing within individuals (see Andrews, Volume 2, Chapter 3). This issue is important for a number of reasons. First, if one is ultimately interested in extending visual word recognition models based on lexical processing studies to individuals who have breakdowns in reading performance, e.g., individuals with developmental dyslexia, then one needs to be concerned about the stability of the lexical processing tasks within individuals. Second, in evaluating the adequacy of computational models, it is possible that the mean performance across individuals at the item level should not

simply be fit to one static model, because a single model will miss the important diversity across individuals. Third, there may indeed be important tradeoffs in the effects of variables (such as spelling-to-sound correspondence vs lexical-semantic processing), which provide important information on how mechanisms associated with specific variables may tradeoff across individuals.

Of course, in order to investigate the stability of lexical processing one needs to have sufficient number of observations of a large number of participants to obtain stable estimates at different points in time. The ELP affords an excellent database to examine stability since it contains naming or lexical decision performance for a large number of participants responding to a large set of different words (and nonwords in LDT) across two sessions, separated by a 24-hour to a 1-week interval. Yap, Balota, Sibley, and Ratcliff (2012) recently undertook this endeavor, and found that the participants in the ELP database provided considerable consistency in performance across these sessions in mean performance, reaction time distributional parameters (such as estimates from the ex-Gaussian function, see Balota & Yap, 2011), and even estimates from the diffusion model (see Ratcliff, Gomez, & McKoon, 2004), and sensitivity to individual lexical variables such as word frequency. Moreover, this database indicated that subjects who had higher vocabulary in general produced faster response latencies, more accurate word recognition performance, and *attenuated* sensitivity to lexical variables. Finally, there was no evidence of tradeoffs in lexical and non-lexical processing across individuals. Clearly, megastudies such as the ELP provide useful information regarding basic aspects of individual differences in lexical processing and are only just beginning to be explored.

To summarize, the work described in this section has provided interesting new constraints on current models and future theory development. While megastudies clearly cannot (and indeed should not) replace well-designed factorial studies for establishing what the benchmark effects should be, they provide a powerful complementary, convergent approach for investigating visual word recognition.

Extending the megastudy approach beyond English

The megastudy approach to isolated word recognition has recently been developing to understand lexical processing in other languages. To our knowledge, there are now two published recent megastudies in non-English languages (note that Keuleers, Lacey, Rastle, & Brysbaert, 2012, have a paper on the British lexicon): The French Lexicon Project (FLP; Ferrand, New, Brysbaert, Keuleers, Bonin, Méot, et al., 2010) and a study of 14,000 Dutch mono- and disyllabic words and nonwords (Keuleers, Diependaele, & Brysbaert, 2010). As an adjunct to the English Lexicon Project, these databases offer a number of benefits. For example, consider the FLP. In addition to stimulating psycholinguistic research in French, researchers can also develop a better understanding of the similarities and differences between English and French, which might yield insights into research aimed at teasing apart language-specific from language-general processes. For example, although English and French are both alphabetic languages,

French is far more morphologically productive, has more transparent mappings from spelling to sound, and has unambiguous syllable boundaries (Ferrand et al., 2010). In the FLP, lexical decision latencies for 38,840 French words and nonwords were collected; due to financial and logistical constraints, speeded pronunciation data have not yet been collected. Although the FLP has only been recently completed, it has already yielded a number of noteworthy findings. Similar to English, a frequency measure based on subtitles predicted lexical decision variance better than book-based frequency estimates. Interestingly, the intriguing quadratic length effect seen in the ELP data (see earlier discussion; New et al., 2006) was also replicated in the FLP, indicating that the non-linear effects of length generalize across languages and are also not specific to the methodological idiosyncrasies of the ELP.

Yap, Balota, Brysbaert, and Shaoul (2010b) have also recently developed a megastudy for Malay. Malay, a language spoken by about 250 million people in Indonesia, Malaysia, Brunei and Singapore, contrasts well with English, due to its very shallow alphabetic orthography (i.e., spelling–sound mappings are predictable and transparent), simple syllabic structures, and transparent affixation. Speeded pronunciation and lexical decision latencies were collected for 9592 Malay words, and regression analyses revealed some interesting processing differences between Malay, a shallow orthography, and English, a deeper orthography. For example, word *length* predicted Malay word recognition performance far better than word frequency. In contrast, frequency is the best predictor in English word recognition. This is consistent with the idea that transparent orthographies heavily implicate a frequency-insensitive sublexical mechanism that assembles pronunciations using a limited set of spelling-sound rules (see Frost, Katz, & Bentin, 1987). Although frequency effects were greatly attenuated in Malay, they were nonetheless reliable, demonstrating that lexical processing plays a role even in very shallow orthographies.

Megastudies have also been used to make other types of cross-linguistic comparisons. Using a progressive demasking task, Lemhöfer, Dijkstra, Schriefers, Baayen, Grainger, and Zwitserlood (2008) compared the word recognition performance of French, German, and Dutch bilinguals for the same set of 1025 monosyllabic English words. English was the second language for these bilinguals. Regression analyses were used to examine the data, and a large number of within-language (e.g., length, word frequency, morphological characteristics, semantic characteristics) and between-language (e.g., cognate status, number of orthographic neighbors in the *first* language) variables were included as predictors. Lemhöfer et al. noted that there was substantial overlap in response time distributions between the three bilingual groups, suggesting that word recognition performance in English generalizes to different bilingual groups with distinct mother tongues. More interestingly, word recognition performance of all three groups was primarily driven by within-language characteristics, i.e., the characteristics of the target language, English. The characteristics of the first language played a relatively limited role in influencing English word recognition. Finally, comparisons of the bilingual groups against a control native English-speaking

group yielded subtle but interesting differences. For example, both written and spoken frequency independently influenced word recognition performance in nonnative speakers, while only spoken frequency had an effect for native speakers. Effects of word frequency were also stronger for nonnative, compared to native, speakers.

In sum, the development of megastudies across different languages already has shed some interesting observations on language-specific vs language-general principles. Although the first steps have already been initiated (indeed there are interesting ongoing megastudies of other languages such as Slovenia; Repovs, personal communication), a future goal of this work would be to establish links across languages in these large databases to provide insights into the mapping of orthography onto meaning, and eventually the mapping of phonology onto meaning, in speech perception. Such a cross-language repository of lexical processing would greatly facilitate our understanding of fundamental characteristics of language. At one level, it would be useful to have an international consortium established to test individuals on identical experimental platforms to insure comparability across the languages, and participants. On the other hand, it would also be useful to have more participants from a wide variety of backgrounds. Indeed, a promising study by Dufau, Dunabeitia, Moret-Tatay, McGonigal, Peeters, Alaorio, et al. (2011) has recently initiated a large international lexical decision study that relies on a common smart-phone platform that participants all over the world can access freely. Data collection has been remarkably fast using this approach, accumulating as many observations in months that the ELP took years to accomplish. Possibly, this approach will lay the foundation of a multilingual psycholinguistic resource, containing performance and lexical characteristics for multiple languages.

Megastudies of semantic priming

Although much has been gleaned about the processes underlying isolated word recognition from both factorial and megastudy approaches, words are typically not recognized in isolation, and there is an extensive literature concerning the influence of semantic/associative context on word recognition (see McNamara, 2005; Jones & Estes, accompanying volume). In the semantic priming paradigm, participants are presented with a target word (e.g., TABLE) for a speeded response (typically pronunciation or lexical decision) that was immediately preceded by either a related (e.g., CHAIR) or an unrelated (e.g., WATCH) prime word. The semantic priming effect refers to the consistent finding that people respond faster to target words preceded by related, relative to unrelated, primes. If one merely wished to demonstrate the existence of semantic priming, then the factorial limitations described earlier would not impede progress because semantic priming researchers typically counterbalance primes and targets across subjects by repairing the same prime-target pairs to create unrelated pairs. Thus, any facilitation in responding to targets could not be due to item selection differences between related and unrelated conditions.

Simple demonstrations of priming, however, are no longer the primary issue of interest. Today, researchers use the semantic priming paradigm as a tool to better understand the organization and retrieval of semantic knowledge. In doing so, researchers select sets of items that differ on a dimension deemed relevant for semantic priming. For example, researchers may test how priming differs as a function of target characteristics such as word frequency (Becker, 1979), regularity (Cortese, Simpson, & Woolsey, 1997), or imageability (Cortese et al., 1997). Alternatively, researchers may examine priming as a function of prime-target relatedness using measures such as forward associative strength (FAS), backward association strength (BAS; Hutchison, 2002; Shelton & Martin, 1992; Thomson-Schill, Kurtz, & Gabrieli, 1998), semantic feature overlap (McRae & Boisvert, 1998; Moss, Ostrin, Tyler, & Marslen-Wilson, 1995), type of semantic relation (Hodgson, 1991), global co-occurrence (Jones, Kintsch, & Mewhort, 2006; Lund, Burgess, & Atchley, 1995), or relational similarity (Estes & Jones, 2006).

Such factorial designs, while important, can distort the relative importance of the variable of interest in accounting for semantic priming. Primes and targets from different item sets often are not matched on potentially important variables. For instance, studies examining priming for categorically related (e.g., HORSE–DONKEY) vs associatively related (e.g., THIRSTY–WATER) pairs often confound type of relation with target frequency such that associatively related targets are often higher in frequency (Bueno & Frenk-Mastre, 2008; Ferrand & New, 2003; Williams, 1996). Since low-frequency words typically show larger priming effects (Becker, 1979), this can artificially inflate the importance of categorical, relative to associative, relations (see Hutchison, 2003).

In addition to matching problems, list context effects also plague factorial semantic priming studies. McKoon and Ratcliff (1995) showed that priming of a particular type of semantic relation (e.g., synonyms or antonyms) is modulated by the proportion of similar types of relations within a list, even when the overall proportion of related items in the list (i.e., the relatedness proportion) is held constant (also see Becker, 1980). Therefore, including many such items within a list likely inflates priming for that particular type of relation (e.g., category members, script-relations, antonyms, etc.). Supporting this argument, Hutchison (2003) observed that priming from perceptually-similar items (e.g., COIN–PIZZA) only occurs when such items constitute a majority of the list. In addition to relation types, the salience of specific item characteristics (e.g., word frequency, regularity, imageability, etc.) is also increased when participants are presented with extreme values on variables in factorial studies.

Finally, as noted earlier, the methodological problems inherent in categorizing continuous variables also apply to semantic priming. Selecting items high or low on a particular dimension can reduce the power to detect true relationships between variables and can even produce spurious effects that do not truly exist when the entire sample is considered (Cohen, Cohen, West, & Aiken, 2003). Most importantly, the use of extreme scores fails to capture the importance of the variable across its full range. If a variable is really an important factor in priming, it should capture the magnitude of priming, not just its presence or absence (McRae, De Sa, & Seidenberg, 1997).

In addition to comparing priming effects across different types of relations and items, researchers have also compared priming effects across different groups of participants including young vs older adults (Balota & Duchek, 1988; Laver & Burke, 1993), and those high vs low perceptual ability (Plaut & Booth, 2000), reading ability (Betjemann & Keenan, 2008), vocabulary (Devitto & Burgess, 2004; Yap & Balota, 2009) and working memory capacity (Hutchison, 2007; Kiefer et al., 2007).[1] As with items, subjects from different populations likely differ in many ways other than the variable of interest and it is impossible to match on everything. One particularly critical difference is often baseline RT. If RTs are not first standardized within participants, priming effects from the slower group will be artificially inflated, often creating a significant group × priming interaction (Faust, Balota, Spieler, & Ferraro, 1999). This methodological flaw can then leave theorists attempting to explain such hyper-priming among their clinical population. In addition, selecting extreme groups on a dimension (e.g., high vs low reading ability) can over- or underestimate the importance of that dimension in priming among the general population.

Hutchison, Balota, Cortese, and Watson (2008)

In an early attempt to highlight, and partially circumvent, such item selection problems, Hutchison et al. (2008) examined priming for 300 strong forward associate pairs (e.g., CAT–DOG) among 108 younger and 95 older adults. Priming effects were measured across both lexical decision and pronunciation tasks using both short (200 ms) and long (1200 ms) stimulus onset asynchrony (SOA) conditions. Because the items were initially selected based upon FAS, this variable was somewhat restricted in range (99% between 0.50 and 0.94). However, there was considerable variability in the other prime-target relation variables examined (global co-occurrence, backward associative strength) and in both prime and target lexical characteristics (frequency, length, orthographic neighborhood). In addition to these variables, baseline RTs for target words were obtained through the use of a neutral prime condition (i.e., the prime BLANK) and RTs for the prime words that were available from the ELP website (another example of the use of such megastudy databases).

Reaction times were first standardized across participants to control for individual differences in baseline RT and variability. Then z-scores for items (averaged across the z-scores calculated within each participant) were obtained in related and unrelated conditions. Multiple regression analyses were then used to predict standardized priming for each item based upon characteristics of the primes and targets (length, log printed word frequency, orthographic neighborhood, baseline RT) as well as the prime-target relatedness variables FAS, BAS, and Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). Across tasks, priming effects were well predicted by the prime characteristics, target characteristics, and prime-target relatedness measures.

There were a number of findings relevant to the present discussion. First, collapsing across tasks, priming at the 200-ms SOA was greater following related

primes that were short, high in frequency, and had few orthographic neighbors. Thus, under such time constraints, priming likely depends upon one's ability to quickly identify the prime word. Second, priming effects were greater for targets that had long baseline RTs, especially within the LDT. These two findings are problematic for any previous or future priming study that uses different prime and/or target words across item sets. In some cases, both primes and targets differ across item sets (e.g., categorical vs associatively related items) whereas in other cases a researcher will contrast priming for the same target (e.g., ANGER) preceded by one of two different related primes (e.g., a synonym RAGE vs an antonym HAPPY). In either case, differential priming effects may be determined entirely by the lexical characteristics of the different items themselves, rather than to any type of prime-target relation per se.

In addition, when collapsed across SOA, FAS predicted RT and error priming in both tasks whereas BAS predicted priming only in the LDT. This pattern is consistent with Neely's (1991) three-process model of semantic priming in which a backwards semantic-matching mechanism contributes to priming in the LDT, but not in pronunciation. This model appropriately predicts that backward relations (from the target to the prime) should increase priming for LDT only.

Finally, LSA similarity did not predict priming in any of the four task × SOA conditions. This finding is problematic for global co-occurrence as a major factor in producing semantic priming. Even though LSA was able to predict that priming would occur in this study (i.e., related items had higher LSA values than unrelated items), it could not predict differences in the degree of semantic priming among related items. In summary, this preliminary regression study of semantic priming has important methodological and theoretical implications for the study of semantic priming and semantic memory.

The Semantic Priming Project (SPP)

The SPP (Hutchison, Balota, Cortese, Neely, Niemeyer, & Bengson, 2011) is an attempt to greatly extend the methodology of Hutchison et al. (2008) to a broader range of items and subjects. Like its predecessor, the ELP, the SPP is a National Science Foundation funded collaborative effort among four universities (Montana State University; University of Albany, SUNY; University of Nebraska, Omaha; and Washington University in St Louis) to investigate a wide range of both item and individual differences in semantic priming. The resulting database (see http://spp.montana.edu) will hopefully aid researchers throughout the world to advance theories and computational models of the processes that allow humans to use context during word recognition.

SPP priming task

A total of 768 native-English-speaking healthy young adults with normal or corrected-to-normal vision were recruited for the semantic priming task: 256 in speeded pronunciation and 512 in lexical decision. Each participant responded to

1661 target words preceded by either a related or unrelated prime. Related pairs were selected from the Nelson, McEvoy, and Schreiber (2004) association norms with the constraint that no item occurred more than twice in the study (once as a prime and once as a target, presented on different days). For each target, a first associate prime (for which the target is the first associate given) and a randomly selected other associate prime (i.e., the target is not the first associate given) were chosen. Unrelated trials were created by randomly re-pairing items within the first and other sets of related pairs. Experimental trials were separated into two sessions with two blocks of trials within each session (a 200-ms SOA block and a 1200-ms SOA block, counterbalanced).

SPP item measures

For item-specific characteristics, the SPP includes the measures (length, frequency, orthographic neighborhood, ELP RT and error rate) used by Hutchison et al. (2008). In addition to these measures, the SPP includes measures of concreteness, imageability, bigram frequency, phonological onset, part-of-speech, and polysemy. For prime-target relational characteristics, the SPP will also include associative measures such as FAS, BAS, associative rank order, semantic measures such semantic feature overlap, connectivity, and type of semantic relation (e.g., synonym, antonym, category coordinate, etc.), and global co-occurence measures such as BEAGLE (Jones et al., 2006) and HAL (Burgess, 1998). The inclusion of such a broad range of variables across the large sample of items in the priming task should greatly increase our understanding of the extent to which item characteristics and types of relatedness contribute to semantic priming.

SPP individual difference measures

As was done for the ELP, we have obtained information about each participant's gender, age, education level, ethnic background, knowledge of non-English languages (e.g., fluency in a second or third language), amount of reading per week on a seven-point scale, circadian rhythm, and self-rated health information. In addition to these measures, the SPP includes measures of reading comprehension, vocabulary, and attentional control (operation span task, Stroop task, and antisaccade task, taken from Hutchison, 2007). As noted previously, performance on each of these measures has been linked to semantic priming performance for various items or under various conditions.

Targeted audience for the website

We anticipate that this database will be an invaluable tool for researchers developing theories of semantic priming and models of semantic memory. Of primary importance is identifying variables crucial for predicting priming across the database. For instance, is semantic priming more accurately predicted by primary word association, number of overlapping features, or similarity in global

co-occurrence? The answer to this question is central to understanding the basic structure of semantic memory. Overall predictability can be tested as well as possible interactions between predictor variables. For instance, perhaps normative association strength (or associative rank order) will produce larger influences on priming when feature overlap and/or global co-occurrence is low, or vice-versa. Perhaps these effects are further modulated by SOA, attentional control, vocabulary, or some combination of these.

This project should also serve as a tool for researchers interested in generating hypotheses for future factorial experiments of semantic priming and actually conducting virtual experiments by accessing the database. In addition, researchers from other areas (e.g., memory, perception, neuroimaging, neuropsychology) will be able to use this database to select items that produce large, medium, or small priming effects and are equated along a number of relevant dimensions. Finally, researchers interested in examining populations such as children, aphasics, schizophrenics, Alzheimer's patients, or healthy older adults could use patterns of priming in this database as a control to test predicted deviations for their population under certain conditions or with certain types of stimuli.

A megastudy of recognition memory

The megastudy approach is obviously not limited to investigations of psycholinguistic variables in lexical decision and pronunciation performance. This approach can be extended across many domains of cognition. For example, item characteristics (e.g., word frequency, concreteness/imageability, orthographic neighborhood size, spelling-to-sound regularity) have also been examined in factorial studies of recognition memory. However, item analyses have been surprisingly rare in the memory literature. Therefore, it is difficult to know if an effect is consistent across items and generalizes to the population of items (Clark, 1973, see special issue of *Journal of Memory and Language* Volume 59, Issue 4, 2008 for detailed discussion of these issues). This is problematic because models of recognition memory do indeed make predictions about particular classes of items, and clearly could be tested at the item level.

Cortese, Khanna, and Hacker (2010) have recently reported the first megastudy of episodic word recognition.[2] The Cortese et al. study provides recognition memory estimates (e.g., hits, false alarms, etc.) for 3000 monosyllabic words. This set of words was selected because estimates for key predictor variables such as imageability and age of acquisition (AoA) were readily available for the majority of these words. In two studies, participants completed 30 study and test lists consisting of 50 and 100 monosyllabic words, respectively, across two 2-hour sessions. The main difference between studies was that in Study 1, participants determined the study duration for each word whereas in Study 2, each word was presented for 2000 ms during study. Across participants, each word was responded to as an old or new item about equally often. The dependent measures were hit rate, false alarm rate, hit minus false alarm rate, d', and C (Snodgrass & Corwin, 1988). Each of these dependent variables was initially analyzed via multiple regression in

which eight predictor variables (see below) were entered simultaneously. Of the 3000 words used in the studies, there were 2578 for which predictor variable values were available. The results across the two studies were very similar (supporting the stability of the data) so the data reported here have been collapsed across studies.

The results of the Cortese et al. (2010) study are useful for the following reasons: First, these data can be used to assess theories of recognition memory. For example, most theories (e.g., Glanzer, Adams, Iverson, & Kim, 1993) predict that items which produce a high hit rate should also produce a low false alarm rate and vice versa (i.e., the mirror effect which yields a negative correlation between hits and false alarms across items). In addition, item noise models (e.g., McClelland & Chappell, 1998) predict that memory will be hampered for items that are similar to many other items. Hypothetically, this similarity could occur at any level (e.g., orthography, phonology, semantics). For highly similar items, there will be more feature matches between the test items and memory representations, increasing the false alarm rates for these items. In addition, one might also hypothesize that individual features will be weakly stored in highly similar words, and this would produce a lower hit rate as well. We can test these possibilities by investigating the influence of orthographic and phonological measures as reflected by the recently developed Levenshtein distance metrics. Semantic similarity may be captured by Age of Acquisition effects. Finally, by regressing the dependent recognition memory measures onto a set of targeted predictor variables, one has the advantage of capturing unique variance of each of the predictor variables, with other variables controlled. Previous research has identified a number of item characteristics that influence recognition memory, but the *relative* influence of each factor remains largely unknown.

The results from this study yielded a number of intriguing observations. Across items, the mean hit rate was 0.72 (SD = 0.10) and the mean false alarm rate was 0.20 (SD = 0.09). The set of predictor variables accounted for 45.9% of the variance in hit rates, 14.9% of the variance in false alarm rates, and 29.2% of the variance in hits minus false alarms. Interestingly, contrary to the prediction that item hit rates should be negatively related to their false alarm rates, hit rates were positively correlated with false alarm rates ($r = 0.145$, $p < .0001$). Hence, when one looks at the item level, as opposed to the factor level, there is not much support for the mirror effect. This is particularly compelling because the same items served as old and new items on the recognition test, and so any idiosyncratic item information that drives hits should also increase false alarms.

The results also indicated that traditional measures of recognition memory including hits minus false alarms and d' were positively correlated with imageability, AoA, and negatively related with word frequency, phonological and orthographic similarity, and word length. Interestingly, imageability and length were the two strongest predictors. Consistent with item noise models, item similarity effects were observed for both phonological density and orthographic density measures. These findings suggest that items sharing similarities with many other items are less distinct (i.e., associated with more noise) and more

difficult to recognize. In addition, there were effects of age of acquisition that may tap into semantic similarity. It is again important to note that the effects of AoA were above and beyond the correlated influence of the related variables.

In sum, the results from the mega recognition memory study have yielded a number of intriguing findings that further our understanding of episodic recognition. Based on these results, it appears that the mirror effect in episodic recognition does not naturally extend to the item level. Moreover, there is some support for item noise models suggesting that there are strong similarity effects along orthographic, phonological, and semantic measures. Finally, imageability and word length provide unique predictive power in recognition performance above and beyond correlated variables of word frequency and familiarity. Clearly the megastudy approach to episodic recognition nicely exemplifies the utility of this approach to models of episodic recognition.

Conclusions

The present chapter reviews evidence of the utility of the megastudy approach in providing further leverage in understanding cognitive performance across a number of distinct domains. As we have emphasized throughout this chapter, we are not suggesting that this is the only way to study such domains, but believe that it is indeed important to use converging evidence across both factorial and megastudy approaches. Hopefully, the megastudies will nurture the development of cumulative datasets that serve to lay the foundation of accepted findings and principles that appear to be common place in other scientific disciplines.

- Megastudies of visual word recognition involve the collection behavioral responses across a large set of participants and items. Such databases provide a complementary, converging approach to the standard factorial studies that dominate the literature.
- Recently, there has been a concern raised that data from megastudies may not be sensitive enough to pick up some subtle effects in the word recognition literature. We directly addressed this issue and showed that the concern is not substantiated in the largest megastudy database available, the English Lexicon Project.
- The megastudy literature was reviewed to demonstrate the utility of this approach in a) the evaluation of extent computational models, b) the development of new models (e.g., models of multisyllabic word processing), c) the evaluation of hitherto unexplored variables (e.g., Levenstein Distance measures), d) comparing the predictive power of multiple measures of the same variable (e.g., the functional form of various word frequency norms), e) comparing the predictive power of different measures (e.g., word frequency vs letter length), f) measuring individual differences in sensitivity to lexical properties.

- Although the megastudy approach was originally developed in the visual word recognition domain in English, this approach has now been extended to other languages (e.g., Dutch, French, German, and Malay), and to other behavioral domains such as Semantic Priming and Episodic Word Recognition. These newer studies were also reviewed.

Acknowledgments

We thank James Adelman and two anonymous reviewers for their comments on this chapter. This work was supported by NSF BCS 0001801 and NSF BCS 0517942.

Notes

1 Both Hutchison (2007) and Kiefer et al. (2005) actually included subjects within the full range of working memory capacity in their studies for their correlational analyses, but included the extreme-groups analyses mainly for illustrative purposes.
2 It should be noted however, that in an analyses of items drawn from 13 experiments, Rubin and Friendly (1986), conducted regression analyses to predict free recall performance for 925 nouns. They found that imageability, emotionality, and the likelihood of being generated as an associate via free association (i.e., availability) were the best predictors of free recall.

References

Adelman, J. S. & Brown, G. D. A. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review*, *14*, 455–459.

Adelman, J. S. & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, *115*, 214–227.

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word naming and lexical decision times. *Psychological Science*, *17*, 814–823.

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2012). *The unexplained nature of reading*. Manuscript submitted for publication.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory & Language*, *55*, 290–313.

Balota, D. A. & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory & Language*, *24*, 89–106.

Balota, D. A. & Duchek, J. M. (1988). Age-related differences in lexical access, spreading activation, and simple pronunciation. *Psychology & Aging*, *3*, 84–93.

Balota, D. A. & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. *Current Directions in Psychological Science*, *20*, 160–166.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, *29*, 639–647.

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

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K.A., Kessler, B., Loftus, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project: A user's guide. *Behavior Research Methods*, *39*, 445–459.

Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, *5*, 252–259.

Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & Cognition*, *8*, 493–512.

Betjemann, R. S. & Keenan, J. M. (2008). Phonological and semantic priming in children with reading disability. *Child Development*, *79*, 1086–1102.

Brysbaert, M. & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990.

Bueno, S. & Frenk-Mastre, C. (2008). The activation of semantic memory: Effects of prime exposure, prime-target relationship, and task demands. *Memory & Cognition*, *36*, 882–898.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods: Instruments & Computers*, *30*, 188–198.

Burgess, C. & Livesay, K. (1998). The effect of corpus size in predicting RT in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods, Instruments, & Computers*, *30*, 272–277.

Chateau, D. & Jared, D. (2003). Spelling-sound consistency effects in disyllabic word naming. *Journal of Memory & Language*, *48*, 255–280.

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior*, *12*, 335–339.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–53.

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

Cortese, M. J. & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments & Computers*, *36*, 384–387.

Cortese, M. J. & Khanna, M. M. (2008). Age of acquisition ratings for 3000 monosyllabic words. *Behavior Research Methods*, *40*, 791–794.

Cortese, M. J., Simpson, G. B., & Woolsey, S. (1997). Effects of association and imageability on phonological mapping. *Psychonomic Bulletin & Review*, *4*, 226–231.

Cortese, M. J., Khanna, M. M., & Hacker, S. (2010) Recognition memory for 2,578 monosyllabic words. *Memory*, *18*, 595–609.

Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior*, *11*, 671–684.

Cutler, A. (1981). Making up materials is a confounded nuisance: or Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, *10*, 65–70.

Dufau, S., Dunabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alaorio, F., Balota, D. A., Brysbaert, M., Carreiras, M., Ferrand, L., Ktoir, M., Perea, M., Rastle, K., Sasburg, O., Yap, M. J., Ziegler, J. C., & Grainger, J. (2011). Smart phone,

smart science: How the use of Smartphones can revolutionize research in cognitive science. *PLoS ONE* 6(9): e24974.

Devitto, Z. & Burgess, C. (2004). Theoretical and methodological implications of language experience and vocabulary skill: Priming of strongly and weakly associated words. *Brain and Cognition*, *55*, 295–99.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*, 777–99.

Ferrand, L. & New, B. (2003). Semantic and associative priming in the mental lexicon. In P. Bonin (Ed.), *Mental lexicon: Some words to talk about words* (pp. 25–43). Hauppage, NY: Nova Science Publishers.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*, 488–496.

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, *28*, 1109–1115.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Strategies*, *13*, 104–115.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*, 546–567.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–259.

Humphreys, L. G. (1978). Research on individual differences requires correlational analysis, not ANOVA. *Intelligence*, *2*, 1–5.

Hutchison, K. A. (2002). The effect of asymmetrical association on positive and negative semantic priming. *Memory & Cognition*, *30*, 1263–1276.

Hutchison, K. A. (2003). Is semantic priming due to association strength or featural overlap? A *micro*-analytic review. *Psychonomic Bulletin & Review*, *10*, 785–813.

Hutchison, K. A. (2007). Attentional control and the relatedness proportion effect in semantic priming. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *33*, 645–662.

Hutchison, K. A., Balota, D. A., Cortese, M., & Watson, J. M. (2008). Predicting semantic priming at the item-level. *Quarterly Journal of Experimental Psychology*, *61*, 1036–1036.

Hutchison, K. A., Balota, D. A., Cortese, M. J., Neely, J. H., Niemeyer, D. P., & Bengson, J. J. (2011). The Semantic Priming Project: A web database of descriptive and behavioral measures for 1,661 nonwords and 1,661 English words presented in related and unrelated contexts. http://spp.montana.edu, Montana State University.

Hodgson, J. M. (1991). Informational constraints on pre-lexical priming. *Language and Cognitive Processes*, *6*, 169–205.

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

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semanticspace accounts of priming. *Journal of Memory & Language*, *55*, 534–552.

Juhasz, B. J., Yap, M. J., Dicke, J., Taylor, S. C., & Gullick, M. M. (2011). Tangible words are recognized faster: The grounding of meaning in sensory and perceptual systems. *The Quarterly Journal of Experimental Psychology*, *64*, 1683–91.

Kang, S. H. K., Yap, M. J., Tse, C-S., & Kurby, C. A. (2011). Semantic size does not matter: 'Bigger' words are not recognised faster. *The Quarterly Journal of Experimental Psychology*, *64*, 1041–1047.

Kučera, H. & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies. A lexical decision study on 14000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Language Sciences, Psychology*, *1*, 174.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The Britsh lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*, 287–304.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Laver, G. D. & Burke, D. M. (1993). Why do semantic priming effects increase in old age? A meta-analysis. *Psychology and Aging*, *8*, 34–43.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.

Lemhöefer, K., Dijkstra, A., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *34*, 12–31.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In J. D. Moore & J. F. Lehman (Ed.), *Proceedings of the 17th annual meeting of the Cognitive Science Society*, (pp. 660–665). Pittsburgh, PA: Lawrence Erlbaum Associates.

Lupker, S. J., Brown, P., & Colombo, L. (1997). Strategic control in a naming task: Changing routes or changing deadlines? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *23*, 570–590.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*, 375–407.

McClelland, J. L. & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*, 724–760.

McKoon, G. & Ratcliff, R. (1995). Conceptual combinations and relational contexts in free association and in priming in lexical decision and naming. *Psychonomic Bulletin & Review*, *2*, 527–533.

McNamara, T. P. (2005). *Semantic Priming: Perspectives from memory and word recognition*. New York, NY: Psychology Press.

McRae, K. & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 558–572.

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*, 99–130.

Maxwell, S. E. & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*, 181–190.

Monsell, S., Patterson, K., Graham, A., Hughes, C. H., & Milroy, R. (1992). Lexical and sublexical translations of spelling to sound: Strategic anticipation of lexical status. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 452–467.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 863–883.

Murray, W. S. & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review*, *111*, 721–756.

Murray, W. S. & Forster, K. I. (2008). The rank hypothesis and lexical decision: A reply to Adelman and Brown (2008). *Psychological Review*, *115*, 240–251.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In Besner, D. & Humphreys, G. W. (Eds) *Basic processes in reading: Visual word recognition*. (pp. 264–336). Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. (2004). The University of South Florida word association, rhyme and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*, 402–407.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Re-examining word length effects in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*, 45–52.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*, 327–357.

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. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, *61*, 106–151.

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., & Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature*, *331*, 585–589.

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., & Raichle, M. E. (1989). Positron emission tomographic studies of the processing of single words. *Journal of Cognitive Neuroscience*, *1*, 153–170.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: HarperCollins.

Plaut, D. C. & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, *107*, 786–823.

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.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159–182.

Rey, A., Courrieu, P., Schmidt-Weigand, F., & Jacobs, A. M. (2009). Item performance in visual word recognition. *Psychonomic Bulletin & Review*, *16*, 600–608.

Rubin, D. C. & Friendly, M. (1986). Predicting which words get recalled: Measures of free recall, availability, goodness, emotionality, and pronunciability for 925 nouns. *Memory & Cognition*, *14*, 79–94.

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

Seidenberg, M. S., Waters, G. S., Sanders, M., & Langer, P. (1984). Pre and post-lexical loci of contextual effects on word recognition. *Memory & Cognition*, *12*, 315–328.

Shelton, J. R. & Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 1191–1210.

Sibley, D. E., Kello, C. T., & Seidenberg, M. S. (2009). *Error, error everywhere: A look at megastudies of word reading.* Proceedings of the Annual Meeting of the Cognitive Science Society. Amsterdam, The Netherlands.

Snodgrass, J. G. & Corwin, J. (1988). Pragmatics of measuring recognition memory. Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50.

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

Taft, M. (1992). The body of the BOSS: Subsyllabic units in the lexical processing of polysyllabic words. *Journal of Experimental Psychology: Human Perception & Performance*, *18*, 1004–1014.

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

Thompson-Schill, S. L., Kurtz, K. J., & Gabrieli, J. D. E. (1998). Effects of semantic and associative relatedness on automatic priming. *Journal of Memory and Language*, *38*, 440–458.

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.

Williams, J. N. (1996). Is automatic priming semantic? *European Journal of Cognitive Psychology*, *22*, 139–151.

Yap, M. J. & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory & Language*, *60*, 502–529.

Yap, M. J., Balota, D. A., Brysbaert, M., & Shaoul, C. (2010a). *Are three frequency measures better than one? Creating a composite measure of word frequency*. Unpublished manuscript.

Yap, M. J., Rickard Liow, S. J., Jalil, S. B., & Faizal, S. S. B. (2010b). The Malay lexicon project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, *42*, 992–1003.

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, *18*, 742–750.

Yap, M. J., Tse, C-S., & Balota, D. A. (2009). Individual differences in the joint effects of semantic priming and word frequency revealed by RT distributional analyses: The role of lexical integrity. *Journal of Memory and Language*, *61*, 303–325.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 53–79.

Yarkoni, T., Balota, D. A., & Yap, M. J. (2008). Beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971–979.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science.

Zevin, J. D. & Balota, D. A. (2000). Priming and attentional control of lexical and sublexical pathways during naming. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*, 121–135.

Zevin, J. D. & Seidenberg, M. S. (2004). Age-of-acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory & Cognition*, *32*, 31–38.

Ziegler, J. & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*, 3–29.