The Quarterly Journal of Experimental Psychology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100704

Do the effects of subjective frequency and age of acquisition survive better word frequency norms?

Marc Brysbaert[a]; Michael J. Cortese[b]
[a] Ghent University, Ghent, Belgium [b] University of Nebraska-Omaha, Omaha, NB, USA

First published on: 09 August 2010

To cite this Article Brysbaert, Marc and Cortese, Michael J.(2010) 'Do the effects of subjective frequency and age of acquisition survive better word frequency norms?', The Quarterly Journal of Experimental Psychology,, First published on: 09 August 2010 (iFirst)
To link to this Article: DOI: 10.1080/17470218.2010.503374
URL: http://dx.doi.org/10.1080/17470218.2010.503374

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
iFirst, 1–15

Psychology Press
Taylor & Francis Group

Do the effects of subjective frequency and age of acquisition survive better word frequency norms?

Marc Brysbaert
Ghent University, Ghent, Belgium

Michael J. Cortese
University of Nebraska–Omaha, Omaha, NB, USA

Megastudies with processing efficiency measures for thousands of words allow researchers to assess the quality of the word features they are using. In this article, we analyse reading aloud and lexical decision reaction times and accuracy rates for 2,336 words to assess the influence of subjective frequency and age of acquisition on performance. Specifically, we compare newly presented word frequency measures with the existing frequency norms of Kučera and Francis (1967), HAL (Burgess & Livesay, 1998), Brysbaert and New (2009), and Zeno, Ivens, Millard, and Duvvuri (1995). We show that the use of the Kučera and Francis word frequency measure accounts for much less variance than the other word frequencies, which leaves more variance to be "explained" by familiarity ratings and age-of-acquisition ratings. We argue that subjective frequency ratings are no longer needed if researchers have good objective word frequency counts. The effect of age of acquisition remains significant and has an effect size that is of practical relevance, although it is substantially smaller than that of the first phoneme in naming and the objective word frequency in lexical decision. Thus, our results suggest that models of word processing need to utilize these recently developed frequency estimates during training or setting baseline activation levels in the lexicon.

Keywords: Word frequency; Familiarity; Age of acquisition; Word recognition.

The extent to which frequency and age of acquisition (AoA) relate to naming and lexical decision performance has become a central issue for researchers interested in word recognition. It is important to understand this issue because a genuine AoA effect has considerable impact on theoretical approaches to word recognition. For example, in parallel distributed processing (PDP) models (see, e.g., Zevin & Seidenberg, 2002) connection weights are adjusted (i.e., connections between inputs and outputs are strengthened) each time the model encounters a word during its learning phase. Therefore, frequency of occurrence affects the speed of processing in PDP models. In contrast, AoA influences performance only when the mapping between input and output is arbitrary. In English, while the relationship between spelling and meaning is arbitrary, the

Correspondence should be addressed to Marc Brysbaert, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Gent, Belgium. E-mail: marc.brysbaert@ugent.be

The authors thank Melvin Yap and James Adelman for kindly giving access to their frequency ranks. They also thank the reviewers, Harald Baayen and Wayne Murray, for many helpful comments.

http://www.psypress.com/qjep
DOI:10.1080/17470218.2010.503374

relationship between orthography and phonology is much less so. Therefore, Zevin and Seidenberg (2002) argued that an effect of AoA would not be observed in word naming. However, one might expect an effect in lexical decision if the decision is partly based on semantic information.

Zevin and Seidenberg's (2002) analysis was a response to a provocative article by Morrison and Ellis (1995; also see Carroll & White, 1973), who claimed that all word frequency effects actually were age-of-acquisition (AoA) effects in disguise. According to Morrison and Ellis, words were processed more efficiently not because they had been encountered more often but because they had been acquired first by the individual. More specifically, Morrison and Ellis (1995) showed that if they matched a list of high-frequency words and a list of low-frequency words on AoA, there was no frequency effect left any more. In contrast, when they matched a list of early-acquired words and a list of late-acquired words on frequency, there still was a significant AoA effect.

Zevin and Seidenberg (2002) criticized Morrison and Ellis (1995; and related work) by pointing out how much the results depended on the quality of the frequency measure. They demonstrated that although Morrison and Ellis's early- and the late-acquired words were matched on the Kučera and Francis (1967; hereafter, KF) norms, they were not matched according to two other frequency estimates: Celex (Baayen, Piepenbrock, & van Rijn, 1993) and the Educator's Word Frequency Guide (Zeno, Ivens, Millard, & Duvvuri, 1995, hereafter Zeno). The words were not matched on subjective familiarity either, making Zevin and Seidenberg conclude that the AoA effect reported by Morrison and Ellis easily could be a frequency effect in disguise, due to the bad quality of the KF frequency measure.

Cortese and Khanna (2007) took up the challenge set by Zevin and Seidenberg (2002) and collected ratings of subjective frequency and AoA for the majority of English monosyllabic words included in the megastudy of Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004). They found that the effect of AoA remained significant in word naming and lexical decision, even after the Zeno et al. (1995) frequency and subjective frequency (Balota, Pilotti, & Cortese, 2001) were partialled out and after other possible confounds such as word length, first phoneme, and grapheme–phoneme consistency were taken into account.

Another criticism against Zevin and Seidenberg was raised by Ghyselinck, Lewis, and Brysbaert (2004), who argued that subjective familiarity ratings were an ill-defined measure, which were likely to be influenced by the time the word was acquired (i.e., participants indicate that they are more familiar with early-acquired words than with late-acquired words). Other authors also noted that the instructions given to participants for making familiarity estimates were quite vague and may have allowed participants to use more than just frequency information (e.g., Alderson, 2007; Thompson & Desrochers, 2009). For instance, Balota et al. (2001) demonstrated that meaningfulness, a semantic variable, was related to familiarity estimates collected under the instructions used by Gernsbacher (1984), but not under subjective frequency instructions that specifically asked the participants how often they came across the words in their daily life (see the Method section). Similarly, Baayen, Feldman, and Schreuder (2006) reported higher familiarity ratings for words that occur more often in spoken than in written language, higher ratings for verbs than for nouns, higher ratings for words that do not resemble many other words (e.g., "yacht", which has a quite distinctive orthography), and different ratings depending on the derivational and inflectional entropy (i.e., the number and the frequency of morphologically and syntactically related words). Given that all these variables also affected lexical decision times, it is not inconceivable that the familiarity ratings of words are partly determined by the ease with which they are recognized (i.e., raters assume that easy recognizable words are familiar words, whereas hard to process words "must be" unfamiliar).

Finally, the usefulness of familiarity ratings is likely to depend on the quality of the objective

frequency counts. Indeed, Gernsbacher's (1984) evidence for the importance of familiarity ratings was based on the limited frequency counts available at the time (Kučera & Francis, 1967; Thorndike & Lorge, 1944). For instance, Gernsbacher mentioned the examples of *boxer*, *icing*, *joker*, *loire*, *gnome*, and *assay*, which all had a frequency of 1 per million in Thorndike and Lorge and KF, but which differed widely in the familiarity estimates given by students. However, these words also differ reliably in more recent frequency counts based on larger databases from a wider range of language registers. For example, the frequencies of these words in film subtitles (Brysbaert & New, 2009; see also below) are: boxer (196), icing (71), joker (233), loire (0), gnome (32), and assay (8; frequencies based on the number of counts in a corpus of 51 million words).

Clearly, the central issue in the whole discussion is how to operationalize the total number of times people have come across a particular word. For a long time, the quality of objective frequency measures was based on face validity with questions such as "how representative is the corpus?" and "how large is the corpus?" Recently, however, research has taken a new turn by correlating frequency measures with an objective criterion: the processing efficiencies for large numbers of words. This line of research was started by Burgess and Livesay (1998) who showed that word frequencies based on a corpus of Internet user groups were significantly better predictors of lexical decision times than the KF frequencies. The research of Burgess and Livesay was limited to samples of a few hundred words. A more ambitious project was published by Balota et al. (2004). They correlated different frequency counts with naming times and lexical decision times for some 2,800 monosyllabic English words and observed substantial quality differences between the measures with particularly poor performance for the KF measure.

Arguably, the most extensive analysis was published by Brysbaert and New (2009). Using the naming times and lexical decision times for nearly 40,000 words from the English Lexicon Project (Balota et al., 2007), Brysbaert and New showed that the quality of a frequency measure depended on the size of the corpus and the language register on which the corpus is based. In particular, they found weak results for corpora based on fewer than 16 million words, and they showed that three language registers seemed to provide interesting converging information. The first register came from written frequencies, as estimated by Burgess and Livesay's (1998) Internet-based HAL frequencies. The second interesting source came from child books, as measured by Zeno et al. (1995). Finally, Brysbaert and New also found very good performance for word frequencies based on a corpus of film subtitles (hereafter SUBTLEX), arguably because these frequencies better capture spoken word use in social relationships. The quality differences were not minor. Whereas KF explained only 19.6% of the variance in lexical decision accuracy scores, the combination of HAL, Zeno, and SUBTLEX accounted for 33.7% of the variance. For the reaction times, percentages of variance accounted for were, respectively, 57.7% and 64.1%. The difference was even larger for the short (monosyllabic and disyllabic) words mostly used in psychological research.

A similar approach was taken by Yap and Balota (2009). They combined two sources: childhood frequencies and frequencies of Internet use. For the former, they used the Zeno et al. (1995) frequencies as well. For the latter, two measures were used: the HAL frequencies and the Google frequencies. The latter were computed by Brants and Franz (2006) on the basis of over 500 billion words stored by the Internet search engine Google. In addition to using frequencies from different sources, Yap and Balota introduced a new transformation. Since the early days of word frequency research, psychologists have used the logarithmic transformation to capture the fact that the relationship between word frequency and processing efficiency is not linear but concave, with bigger processing gains for increases in low frequencies than for equivalent increases in high frequencies. Murray and Forster (2004), however, argued that frequency ranks were a

better alternative, a finding confirmed by Yap (2007) after an intensive analysis of the available megastudies. Therefore, Yap and Balota (2009) proposed the combined rank (CORA) measure as the best possible predictor of frequency of occurrence. Again, this measure explained substantially more variance than the other frequency measures.

The recent developments above allow us for the first time to properly test the relationship between objective frequency, subjective frequency, and AoA. In particular, the following questions were addressed:

1. Do better objective frequency measures result in more variance explained in word processing efficiency measures, or does the gain due to objective word frequency happen at the expense of subjective frequency and AoA?
2. If the impact of subjective frequency depends on the quality of the objective frequency measure, is there still evidence for a genuine contribution of subjective frequency once objective frequency is optimized?
3. If the impact of AoA depends on the quality of the objective frequency measure, is there still evidence for a genuine AoA effect once objective frequency is optimized?

The first question examines to what extent subjective ratings add specific information or correct poor objective measurements. Do subjective ratings add something extra to objective counts or are they only useful as long as the objective measurements are suboptimal? For instance, Gernsbacher (1984) argued that familiarity ratings were particularly needed for low-frequency words, because it was difficult to get reliable objective frequency estimates for these words. In this view, subjective ratings are "corrective", and their impact will decrease as more reliable objective measures become available. In contrast, AoA researchers (e.g., Brysbaert & Ghyselinck, 2006; Ellis & Lambon-Ralph, 2000) are adamant that the AoA effect cannot be reduced to a cumulative frequency effect, dependent on the total number of times participants have encountered the words in their lives. In their view, no matter how good a frequency measure is able to capture the cumulative frequency, there will always remain a core of the AoA effect due to the *order* in which the words were learned. This is the background against which Questions 2 and 3 were examined. When addressing these questions, it is important to realize that not only statistical significance determines the outcome, but also the amount of variance explained. Given that statistical significance is obtained rapidly if the analyses are based on thousands of observations, and given that all the variables we are testing are correlated with other word features, the questions to be addressed is whether subjective frequency and AoA still explain a meaningful percentage of variance once the effect of objective frequency is partialled out.

To test the hypotheses, we pitted subjective frequency and AoA against seven different frequency measures. Because we did not want our results to be influenced by semantic variables, we used Balota et al. (2001) subjective frequency ratings rather than Gernsbacher's (1984) familiarity ratings (see the Method section for more details about the different measures). The frequency measures were:

1. KF.
2. Zeno.
3. HAL.
4. SUBTLEX.
5. COMP (i.e., composite, or mean of Zeno, HAL, and SUBTLEX).
6. CORA.
7. MonoRank (i.e., the rank ordering of the words assessed in the current study, derived from the composite measure).

To examine the hypotheses we took advantage of a rich set of naming and lexical decision data collected by Balota and colleagues using the megastudy approach (Balota et al., 2004; Spieler & Balota, 1997; see also Cortese & Khanna, 2007). In these megastudies naming and lexical decision times were collected for thousands of monosyllabic words, allowing researchers to properly examine the factors influencing word processing.

Method

Data

The current study utilized the dataset compiled by Balota and colleagues (for details see Balota et al., 2004; Cortese & Khanna, 2007; Spieler & Balota, 1997). The primary measures were naming time, naming accuracy, lexical decision time, and lexical decision accuracy. The primary data for the current analyses consisted of 2,336 item means obtained for each of these measures. A total of 31 Washington University college undergraduates participated in the naming study, and 30 Washington University undergraduates participated in the lexical decision study.

Stimuli

There were 2,870 monosyllabic words used in the naming study and 2,906 monosyllabic words and an equal number of length-matched nonwords used in the lexical decision study. Stimuli ranged between two and seven letters in length. Nonwords were constructed by changing one to three letters of each word.

For 2,336 items we had information about all the variables included in the regression analyses. Apart from the different objective frequency measures, the two most important variables were subjective frequency and AoA. Subjective frequency estimates were collected by Balota et al. (2001). In their study, participants estimated the frequency for which words were encountered or produced in various modalities (reading, hearing, written, spoken). Participants also provided a general "encountered" estimate that presumably included all of the various modalities. This general "encountered" rating was that used by Balota et al. (2004), by Cortese and Khanna (2007), and in the present analyses. Words appeared in booklets of 109 to 160 words, and ratings were indicated by filling in ovals on scan sheets with a pencil. Each word was rated by 30 or 32 participants. Ratings were made to indicate the frequency of encounters with a word based on the following scale: 1 = never, 2 = once a year, 3 = once a month, 4 = once a week, 5 = once every two days, 6 = once a day, 7 = several times a day.

AoA estimates were collected by Cortese and Khanna (2008). In their study, 32 participants provided an AoA estimate for each of 3,000 words. Participants were instructed to estimate the age at which each word was acquired according to Gilhooly and Logie's (1980) scale where 1 = age 0–2 years, 2 = age 2–4 years, 3 = age 4–6 years, 4 = age 6–8 years, 5 = age 8–10 years, 6 = age 10–12 years, and 7 = age 13 years or older. This scale appeared below each word, which appeared in the centre of a computer screen, and participants made their ratings via the keyboard.

Because it is well known that the word frequency effect levels off for values above 100 per million (Baayen et al., 2006; Balota et al., 2004), unless otherwise indicated we used polynomials of the second degree to estimate the impact of word frequencies (i.e., both log frequency and $\log^2$ frequency were used as predictors).[1] A similar nonlinearity was not present for the frequency ranks, because the number of words with frequencies higher than 100 per million are quite limited and, hence, do not differ much in rank.

Procedure

For each participant, the study consisted of two sessions that were separated by at least 24 hours and no more than a week. In the lexical decision study, there were 600 trials in Blocks 1–9, and there were 412 trials in Block 10. For each session of the naming study, there were 150 trials in Blocks 1–9 and 85 trials in Block 10. In both studies, each session began with 20 practice trials, and there were 2 practice trials after every break.

Participants were instructed to respond quickly and accurately. The following events occurred on each trial: (a) a fixation mark appeared at the centre of a monitor for 400 ms; (b) a blank screen occurred for 200 ms, and (c) the stimulus appeared in the centre of the monitor until a

[1] The results remained the same when restricted cubic splines were used instead of second-degree polynomials (there were only differences in the third digit of the percentages of variance accounted for).

response occurred. In the naming task, the response initiated a voice key and erased the stimulus from the screen. Each participant coded his or her naming responses via the mouse as correct or incorrect, and an intertrial interval of 1,200 ms followed. A naming response was coded as incorrect if either the word was mispronounced or there was a voicekey error (i.e., the voicekey failed to register or it was prematurely initiated, e.g., by a coughing sound). In the lexical decision task, a word decision was indicated by pressing the "/" key (labelled "YES") on the keyboard, and a nonword decision was indicated by pressing the "z" key (labelled "NO"). Correct responses initiated a 1,200-ms intertrial interval. Incorrect lexical decision responses initiated a message that appeared slightly below the centre of the screen for 1,500 ms. The message informed the participant that his or her response was incorrect. On these trials, the participant pressed the space key to initiate the 1,200-ms delay.

Regression analyses

Following Balota et al. (2004) and Cortese and Khanna (2007), we performed hierarchical regression analyses on the 2,336 item means separately for naming reaction time, naming proportion correct, lexical decision reaction time, and lexical decision proportion correct. Reaction time means were log transformed.[2] We alternated seven objective frequency measures in the predictor set for each dependent variable: KF, Zeno, SUBTLEX, COMP, CORA, and MonoRank (i.e., the rank ordering of the words assessed in the current study, derived from the composite measure). For each nonranked objective frequency measure, we added 1 to the raw frequency value, then calculated the frequency per million, and finally calculated the log of the frequency of the per million value—that is, $\log[(\text{freq} + 1)/N]$; freq = frequency, N = corpus size in millions. Thus, words that did not have a frequency value in the corpus were associated with a raw frequency value of 1.

Because there were seven frequency measures and four dependent variables, we had a total of 28 separate analyses. For the 2,336 words employed in the studies, the mean reaction time in naming was 468.3 ms ($SD = 20.7$, range = 153.0), the mean proportion correct in naming was .96 ($SD = .05$, range = .52), the mean reaction time in lexical decision was 615.4 ms ($SD = 62.0$, range = 353.7), and the mean proportion correct in lexical decision was .92 ($SD = .08$, range = .47). We note that the mean accuracy rate for naming was quite high, and so the results of the analyses involving proportion correct in naming should be interpreted with caution.

Results

Table 1 presents the percentages of variance accounted for by (a) the objective frequency measures, (b) when subjective frequency and AoA were included separately with objective frequency in the regression equation, and (c) when all three factors were entered simultaneously in the regression equation. The increasing quality of the objective frequency measures is obvious. Whereas KF accounts for 6.1% of the variance in naming times and 32.3% in lexical decision times, the composite measure explains 9.6% and 44.0%, and CORA explains 9.0% and 45.4%, respectively. Zeno, HAL, and SUBTLEX have values in-between, with better performance in general for SUBTLEX than for Zeno or HAL. These findings are in line with the results reported by Brysbaert and New (2009) and by Yap and Balota (2009) and confirm the superior quality of the new frequency measures.[3]

[2] The same results are obtained when the inverse values of the reaction time (RT) means are used or when the raw RT means are used.

[3] To make sure that we gave frequency ranks the same priority as raw frequencies, we additionally checked the frequency ranks based on Kučera and Francis (1967), Zeno et al., Celex, and the British National Corpus, collected by Adelman and Brown (2008). A further advantage of these frequency ranks is that they have been corrected for words unlikely to be known to university students. The correlations between the new frequency ranks and the dependent variables were always lower than those between CORA and the dependent variables.

Table 1. *Adjusted R^2s across objective frequency measures for each dependent measure*

		Objective frequency measure						
		KF	*Zeno*	*SUBTLEX*	*HAL*	*COMP*	*CORA*	*MonoRank*
Naming reaction times	Objective frequency[a]	.061**	.076**	.098**	.088**	.096**	.090**	.094**
	Obj + subj freq[b]	.076**	.084**	.098**	.092**	.096**	.094**	.095**
	Obj + AoA[b]	.099**	.099**	.110**	.111**	.110**	.107**	.109**
	Obj + subj freq + AoA[c]	.099**	.099**	.110**	.111**	.110**	.106**	.109**
Naming accuracy	Objective frequency[a]	.009**	.017**	.015**	.018**	.019**	.021**	.021**
	Obj + subj freq[b]	.012**	.018**	.015**	.018**	.018**	.021**	.013**
	Obj + AoA[b]	.022**	.024**	.023**	.025**	.025**	.024**	.020**
	Obj + subj freq + AoA[c]	.022**	.024**	.022**	.026**	.025**	.026**	.021**
Lexical decision reaction time	Objective frequency[a]	.323**	.394**	.435**	.372**	.440**	.454**	.415**
	Obj + subj freq[b]	.439**	.472**	.474**	.460**	.485**	.480**	.434**
	Obj + AoA[b]	.498**	.490**	.507**	.507**	.518**	.518**	.494**
	Obj + subj freq + AoA[c]	.515**	.515**	.519**	.522**	.530**	.522**	.495**
Lexical decision accuracy	Objective frequency[a]	.181**	.253**	.274**	.247**	.290**	.299**	.212**
	Obj + subj freq[b]	.275**	.317**	.311**	.311**	.331**	.303**	.226**
	Obj + AoA[b]	.309**	.320**	.329**	.336**	.345**	.328**	.275**
	Obj + subj freq + AoA[c]	.326**	.344**	.342**	.349**	.359**	.328**	.276**

Note: For all frequency (freq) estimates except CORA, log freq and $\log^2$ freq were included in the regression analysis to capture the nonlinearity in the curve. KF = Kučera and Francis (1967). Zeno = Zeno, Ivens, Millard, and Duvvuri (1995). SUBTLEX = Brysbaert and New (2009), HAL = Lund and Burgess, 1996. COMP = composite frequency (i.e., average of Zeno, SUBTLEX, and HAL). CORA = composite rank frequency from Yap and Balota (2009). MonoRank = monosyllablic ranked frequency. AoA = age of acquisition.

[a]Only objective (obj) frequency included in the regression equation. [b]Subjective (subj) frequency and AoA (age of acquisition) included separately with objective frequency in the regression equation. [c]All three factors are entered simultaneously in the regression equation.

**$p < .01$.

A second important observation in Table 1 is that subjective frequency and AoA make up for most of the differences in variance accounted for by the objective frequency measures. When all three variables are entered, differences in R^2 are relatively small. This agrees with Zevin and Seidenberg's (2002) argument that suboptimal measures of word frequency open the way for other variables to fill the gap. Whereas subjective frequency and AoA "account" for 19.2% of the lexical decision times when the objective frequency is assessed with KF, this percentage of variance shrinks to 9.0% with the composite measure, and to 6.8% when the objective frequency is assessed with CORA. It is not the case that better objective frequency measures lead to more variance explained overall.

We generated two scatterplots to display the relationship between objective frequency and subjective frequency when a suboptimal frequency measure (i.e., KF, see Figure 1) is employed versus a more optimal measure (i.e., the Zeno, SUBTLEX, and HAL composite measure, see Figure 2). These figures reveal quite clearly that there is much less predictability between KF and subjective frequency at the low end of the frequency scale than at the high end of the frequency scale whereas the predictability of the composite measure is much more constant throughout the entire range of frequency values. Scatterplots

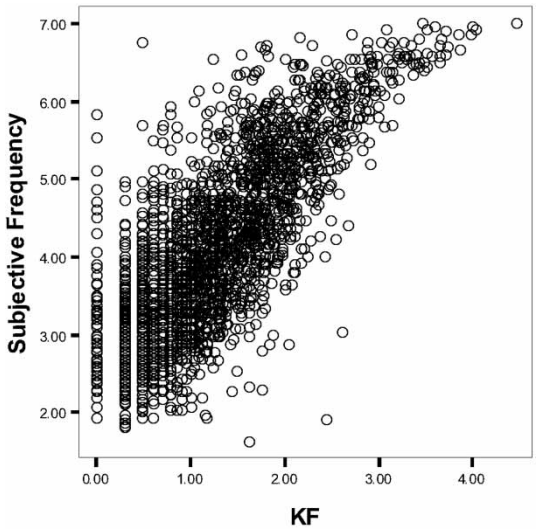

Figure 1. *A scatterplot of the relationship between KF (Kučera & Francis, 1967), frequency and subjective frequency. KF frequencies are expressed in log10 per million words, meaning that a value of 0 corresponds to 1 per million, 2 corresponds to 100 per million, and 4 corresponds to 10,000 per million.*

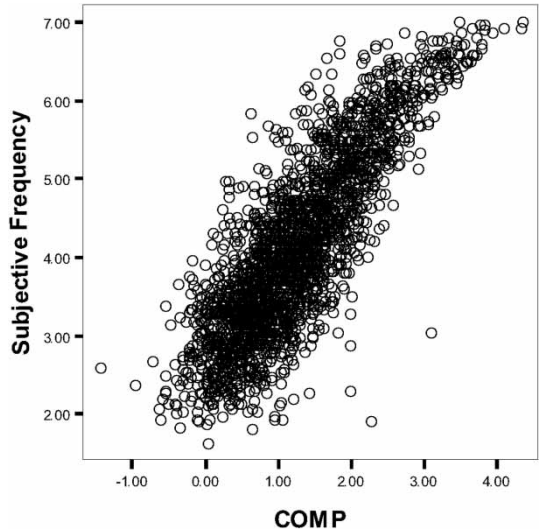

Figure 2. *The relationship between the composite frequency measure (COMP; i.e., the mean of Zeno, SUBTLEX, and HAL) and subjective frequency. Zeno = Zeno, Ivens, Millard, and Duvvuri (1995), SUBTLEX = Brysbaert and New (2009), HAL = Lund and Burgess, 1996. The COMP frequencies are expressed in log10 per million words, meaning that a value of −1 corresponds to 0.1 per million, 0 to 1 per million, and 2 to 100 per million.*

showing the relationships between KF and AoA and between COMP and AoA were also generated (see Figures 3 and 4).

To get a better idea of the specific effects of the different frequency measures, we repeated the analysis reported in Cortese and Khanna (2007). That is, phonological onset variables (i.e., the characteristics of the first phoneme) were entered in Step 1, sublexical and lexical variables were entered in Step 2, and semantic variables were entered in Step 3. For the present purposes, the important things to note are that objective frequency and subjective frequency were entered in Step 2, and AoA and imageability were entered in Step 3 (see Table 2 for the intercorrelation matrix of the predictor variables entered in Steps 2 and 3). In Step 1 (see Cortese & Khanna, 2007, for standardized regression coefficients), the amount of variance accounted for by the phonological onset variables collectively was 34.9% for naming reaction times, 4.3% for naming proportion correct, 1.1% for lexical decision reaction times, and 1.0% for lexical decision proportion correct.

Table 3 displays the standardized regression coefficients obtained in Steps 2 and 3 and their corresponding adjusted R^2s for each analysis. To improve the interpretation of the coefficients, we excluded the $\log^2$ frequency variable from this analysis, so that the same variables are included for each frequency measure (including $\log^2$ frequency does not change any of the conclusions drawn; it just distributes the frequency effect of KF, Zeno, and SUBTLEX over two variables).

The first important result to note in Table 3 is the nearly perfect negative correlation between the weights of the objective frequency measures and the weights of the subjective frequencies: The higher the quality of the objective frequency, the lower the weight of the subjective frequency. Specifically, the weights for subjective frequency were clearly highest for KF and clearly lowest for the composite measure and for CORA and MonoRank. This pattern was consistent across all dependent measures. There was also a negative

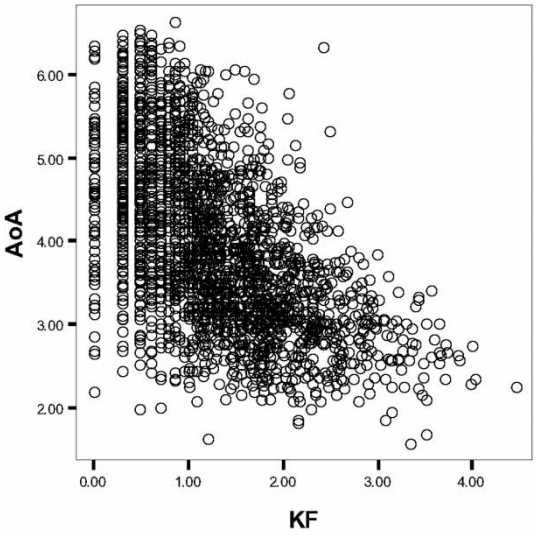

Figure 3. *The relationship between KF (Kučera & Francis, 1967) and AoA (age of acquisition).*

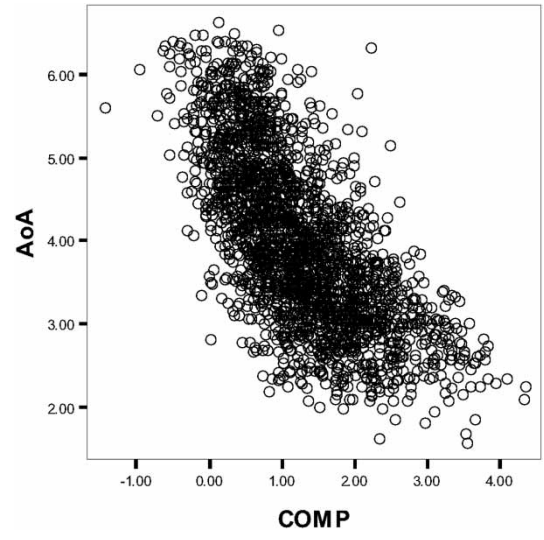

Figure 4. *The relationship between the composite frequency measure (COMP; i.e., the mean of Zeno, SUBTLEX, and HAL) and AoA (age of acquisition). Zeno = Zeno, Ivens, Millard, and Duvvuri (1995), SUBTLEX = Brysbaert and New (2009), HAL = Lund and Burgess, 1996.*

correlation between the quality of the objective frequency measure and the standardized weight of AoA, but this effect was less strong.

Of further interest is the observation that the weights of the other variables (e.g., word length, orthography–phonology consistency, imageability) vary little with the quality of the objective frequency index, making it unlikely that these variables are confounded by word frequency.

Discussion

This article is the first to provide an in-depth analysis of the consequences of quality differences in objective frequency measures for the impacts of word familiarity and AoA. It is already known for some time that the KF norms are deficient (although many researchers still seem to be unaware of it, given that the KF measure continues to be used in about 200 articles published in top journals each year).

Interestingly, the use of better frequency measures does not lead to much higher percentages of variance in word processing efficiency accounted for (Table 1). Rather, what happens is that the void is filled, first and foremost by the subjective frequency ratings and second by the AoA ratings. This gives these variables more weight than they deserve and at the same time shows how much they are needed if researchers are working with outdated frequency counts based on a small corpus (KF is based on a corpus of 1 million words only). There is little point in matching stimuli on an inferior variable such as KF, although arguably the measure can still be used to make a crude distinction between high-frequency and low-frequency words (given the reasonable correlations between KF and the other frequency measures; Table 2).

The fact that the weight of the subjective frequency measure depends so much on the quality of the objective frequency measure used undermines Gernsbacher's (1984) argument that this is an important variable to control for. Its impact depends more on the weakness of the objective frequency count than on its inherent worth. In principle, one can expect a small genuine contribution from this variable when the familiarity raters and the participants of the word-processing experiment share the same features (e.g., both are

Table 2. *Correlation matrix for the predictor variables entered in Steps 2 and 3 regression analyses*

		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
1.	Length		−.654**	−.158**	−.133**	−.163**	−.220**	−.212**	−.173**	−.189**	−.005	−.024	.021	.007	−.067**	.261**
2.	Neighbourhood size			.122**	.113**	.139**	.182**	.176**	.176**	.182**	.114**	.011	.090**	.124**	.062**	.207**
3.	Subjective frequency				.773**	.785**	.843**	.837**	.737**	.815**	−.072**	−.104**	−.006	−.041*	−.020	−.722**
4.	KF					.906**	.857**	.911**	.766**	.873**	−.085**	−.123**	−.013	−.077**	−.134**	−.561**
5.	Zeno						.871**	.943**	.858**	.918**	−.059**	−.119**	−.002	−.061**	.014	−.692**
6.	SUBTLEX							.954**	.843**	.920**	−.082**	−.116**	−.025	−.047*	−.029	−.703**
7.	COMP								.883**	.971**	−.073**	−.119**	.016	−.057**	.036	−.683**
8.	CORA									.925**	−.028	−.075**	.027	−.044*	.157**	−.692**
9.	MonoRank										.044*	−.093**	.001	−.038	.048*	−.695**
10.	FF onset consistency											.023	.217**	.059**	.066**	−.004
11.	FF rime consistency												.071**	.223**	.037	.080**
12.	FB onset consistency													.057**	.041*	−.033
13.	FB rime consistency														.079**	−.020
14.	Imageability															−.382**
15.	AoA															

Note: KF = Kučera and Francis (1967). Zeno = Zeno, Ivens, Millard, and Duvvuri (1995). SUBTLEX = Brysbaert and New (2009), HAL = Lund and Burgess, 1996. COMP = composite frequency (i.e., average of Zeno, SUBTLEX, and HAL). CORA = composite rank frequency from Yap and Balota (2009). MonoRank = monosyllablic ranked frequency. FF = feedforward. FB = feedback. AoA = age of acquisition.

**$p < .01$.

Table 3. *The standardized regression coefficients obtained in Steps 2 and 3 and their corresponding adjusted R^2s as a function of objective frequency measure for each dependent measure*

			Objective frequency measure						
			KF	*Zeno*	*SUBTLEX*	*HAL*	*COMP*	*CORA*	*MonoRank*
Naming reaction times									
Step 1	Adjusted R^2		.349**	.349**	.349**	.349**	.349**	.349**	.349**
Step 2		Length	.154**	.152**	.144**	.147**	.145**	.154**	.152**
		Orthographic neighbourhood size	−.110**	−.107**	−.105**	−.106**	−.105**	−.098**	−.099
		Feedforward onset consistency	−.025	−.023	−.025	−.023	−.024	−.022	−.022
		Feedforward rime consistency	−.074**	−.076**	−.076**	−.074**	−.076**	−.071**	−.073**
		Feedback onset consistency	−.075**	−.073**	−.078**	−.077**	−.076**	−.073**	−.076**
		Feedback rime consistency	−.075**	−.075**	−.073**	−.074**	−.076**	−.076**	−.073**
		Subjective frequency	−.173**	−.129**	−.106**	−.142**	−.103**	−.113**	−.097**
		Objective frequency	−.078**	−.134**	−.154**	−.117**	−.158**	−.163**	−.168**
	Adjusted R^2		.491**	.495**	.495**	.493**	.496**	.500**	.498**
Step 3		Imageability	.001	.004	−.002	.002	−.001	.023	.010
		AoA	.122**	.097**	.098**	.118**	.100**	.094**	.100**
	Adjusted R^2		.497**	.498**	.499**	.499**	.500**	.503**	.501**
	Adjusted R^2 with $\log^2$ freq included		.499**	.500**	.500**	.500**	.500**		
Naming accuracy									
Step 1	Adjusted R^2		.043**	.043**	.043**	.043**	.043**	.043**	.043**
Step 2		Length	−.085*	−.084**	−.082**	−.081**	−.081**	−.086**	−.085**
		Orthographic neighbourhood size	.008	.005	.007	.004	.005	−.003	.002
		Feedforward onset consistency	.167**	.167**	.167**	.167**	.167**	.166**	.166**
		Feedforward rime consistency	.155**	.157**	.156**	.156**	.157**	.154**	.155**
		Feedback onset consistency	.099**	.099**	.100**	.101**	.100**	.098**	.100**
		Feedback rime consistency	.077**	.078**	.076**	.078**	.077**	.080**	.077**
		Subjective frequency	.081**	.044	.072*	.047	.041	.007	.034
		Objective frequency	.029	.075*	.037	.072*	.076*	.131**	.085*
	Adjusted R^2		.134**	.135**	.134**	.135**	.135**	.141**	.136**
Step 3		Imageability	.041+	.040+	.037	.043+	.042+	.028	.037
		AoA	−.070*	−.057	−.071+	−.065+	−.060+	−.046	−.061+
	Adjusted R^2		.139**	.139**	.138**	.140**	.139**	.143**	.139**
	Adjusted R^2 with $\log^2$ freq included		.139**	.141**	.138**	.141**	.140**		
Lexical decision reaction times									

(*Continued overleaf*)

Table 3. *Continued.*

			Objective frequency measure						
			KF	*Zeno*	*SUBTLEX*	*HAL*	*COMP*	*CORA*	*MonoRank*
Step 1	Adjusted R^2		.011**	.011**	.011**	.011**	.011**	.011**	.011**
Step 2		Length	.002	−.002	−.019	−.012	−.016	.004	.000
		Orthographic neighbourhood size	.005	.013	.016	.013	.016	.045*	.038+
		Feedforward onset consistency	−.067**	−.064**	−.069**	−.064**	−.066**	−.061**	−.061**
		Feedforward rime consistency	−.052**	−.057**	−.056**	−.053**	−.057**	−.045**	−.050**
		Feedback onset consistency	.001	.004	−.006	−.004	−.003	.007	−.003
		Feedback rime consistency	−.015	−.017	−.012	−.015	−.015	−.022	−.014
		Subjective frequency	−.507**	−.371**	−.337**	−.440**	−.342**	−.250**	−.248**
		Objective frequency	−.144**	−.318**	−.340**	−.229**	−.336**	−.503**	−.458**
	Adjusted R^2		.387**	.417**	.411**	.398**	.412**	.493**	.448**
Step 3		Imageability	−.183**	−.173**	−.186**	−.178**	−.184**	−.128**	−.163**
		AoA	.305**	.257**	.255**	.301**	.263**	.227**	.248**
	Adjusted R^2		.502**	.505**	.508**	.509**	.511**	.546**	.528**
	Adjusted R^2 with $\log^2$ freq included		.532**	.531**	.535**	.540**	.546**		
Lexical decision accuracy									
Step 1	Adjusted R^2		.010**	.010**	.010**	.010**	.010**	.010**	.010**
Step 2		Length	.083**	.082**	.091**	.088**	.090**	.075**	.080**
		Orthographic neighbourhood size	−.014	−.023	−.023	−.023	−.025	−.061*	−.041
		Feedforward onset consistency	.060**	.061**	.064**	.061**	.062**	.058**	.059**
		Feedforward rime consistency	.071**	.077**	.076**	.074**	.076**	.068**	.073**
		Feedback onset consistency	.018	.016	.022	.021	.020	.012	.020
		Feedback rime consistency	.011	.016	.013	.015	.015	.024	.015
		Subjective frequency	.469**	.320**	.322**	.360**	.310**	.116**	.223**
		Objective frequency	.007	.201**	.186**	.149**	.202**	.494**	.314**
	Adjusted R^2		.224**	.239**	.234**	.232**	.236**	.333**	.257**
Step 3		Imageability	.242**	.242**	.249**	.251**	.252**	.208**	.238**
		AoA	−.215**	−.184**	−.184**	−.200**	−.182**	−.118**	−.172**
	Adjusted R^2		.339**	.342**	.342**	.348**	.345**	.393**	.352**
	Adjusted R^2 with $\log^2$ freq included		.369**	.384**	.382**	.394**	.397**		

Note: KF = Kučera and Francis (1967), Zeno = Zeno, Ivens, Millard, and Duvvuri (1995), SUBTLEX = Brysbaert and New (2009), HAL = Lund and Burgess, 1996; COMP = composite frequency (i.e., average of Zeno, Subtitle, and HAL). CORA = composite rank frequency from Yap and Balota (2009), MonoRank = monosyllablic ranked frequency. AoA = age of acquisition.

*$p < .05$. **$p < .01$. +$.05 < p < .10$.

native English-speaking university students from the USA), because these features may not be present in the objective frequency counts. However, as soon as the raters differ from the participants of the word experiment, we conjecture that the subjective frequency rating will not add anything beyond what can be captured by a good frequency count. Also, while the subjective frequency estimates of Balota et al. (2001) seem to be less affected by semantic variables such as meaningfulness, it is unclear whether other types of nonfrequency information are being used to make the estimates. The subjective nature of such estimates always allows for that possibility.

Regarding word frequency, our findings, as well as those reported by Brysbaert and New (2009), have important implications for models of word recognition. For example, models utilizing a lexicon (e.g., Perry, Ziegler, & Zorzi, 2007) may assign a baseline level of activation to each word based on its frequency. Obviously, the models will gain in predictive power to the extent that they utilize optimal frequency estimates. In PDP models (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996), the exposure a model has with a word is determined by its frequency. Therefore, utilizing optimal frequency measures has important implications for PDP models as well.

It is also important to note that the work of Murray and Forster (2004), as well as more recent work by Keuleers, Diependaele, and Brysbaert (2010), indicates that frequency effects are much stronger at the lower end of the frequency spectrum than at the higher end. In fact, the frequency effect for words between 0.1 and 1 per million is larger than that between 10 and 100 per million. Thus models of word recognition not only need to utilize optimal estimates, they need to be able to capture the nonlinear relationship that frequency has with processing speed. Furthermore, the effect of frequency for very-low-frequency words is remarkable given the number of exposures people have with these words. Assuming a steady input of 200 words per minute for 16 hours a day (likely to be an overestimate), a 20-year old has received $200 \times 60 \times 16 \times 365.25 \times 20 = 1.4$ billion words. This means that a typical 20-year-old reader on average has encountered a word with a frequency of 0.1 per million no more than 140 times in his or her life.

Turning to AoA, Zevin and Seidenberg (2002) have argued that AoA ratings are also tainted by the quality of the objective word frequency measure (or the lack of such quality). Our analyses confirm this claim, although the impact is much less than for familiarity: The standardized weight of AoA does not drop as much between KF and CORA or SUBTLEX as the standardized weight of subjective frequency. Our analyses also show that even with the best frequency measures currently available AoA ratings still explain a reliable percentage of variance. This is in agreement with the models of word processing that take into account the fact that the first learned information enjoys a processing advantage over later learned information, even when the total number of encounters are equated or made advantageous to the later acquired information (Ellis & Lambon Ralph, 2000; Steyvers & Tenenbaum, 2005). In this respect it may also be worthwhile to remember that the present analyses were based on word-processing times of very highly performing university students. It will be interesting to examine to what extent the impact of word frequency differs as a function of schooling and reading practice.

Given the significant effect of AoA, even when frequency is well controlled, it is sensible to control stimuli for this variable if the measure is available. At the same time, we must keep in mind that AoA accounts for at most 7% extra variance (in lexical decision times). This can be compared to the 35% as a result of the onset phoneme in naming times and the 46% due to objective frequency in lexical decision times. This clearly shows the pecking order of variables to control for if compromises must be made. It also illustrates that for investigators starting to examine languages with few data available it is more worthwhile to invest in good objective frequency counts rather than to invest in the development of AoA ratings. If the frequency counts are not good, a considerable part of the effect ascribed to AoA is likely to be a frequency effect in disguise.

In sum, recently developed frequency norms provide optimal estimates that should be seriously considered when interpreting word frequency effects, training models of word recognition, and/or controlling for frequency. The use of these estimates makes the use of subjective frequency estimates much less important than in the past. Finally, the practical impact of AoA in word recognition, while still important for theoretical approaches, may also have been overemphasized in the past.

Original manuscript received 26 June 2009
Accepted revision received 1 June 2010
First published online day month year

REFERENCES

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

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28,* 383–409.

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

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). The CELEX Lexical Database [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

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

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods*, *39,* 445–459.

Brants, T., & Franz, A. (2006). Web IT 5-Gram Version 1. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency-related, partly frequency-independent. *Visual Cognition, 13,* 992–1011

Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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.

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

Carroll, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture naming latency. *Quarterly Journal of Experimental Psychology*, *25*, 85–95.

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology*, *60,* 1072–1082.

Cortese, M. J., & Khanna, M. M. (2008). Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods*, *40,* 791–794.

Ellis, A. W., & Ralph, M. A. L. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26,* 1103–1123.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113,* 256–281.

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literatuire and a new multi-task investigation. *Acta Psychologica, 115*, 43–67.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12,* 395–427.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). *Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords*. Manuscript submitted for publication.

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

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*. *28*, 203–208

Morrison, C. M., & Ellis, A. W. (1995). Roles of word-frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21,* 116–133.

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

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.

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

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29,* 41–78.

Thompson, G. L., & Desrochers, A. (2009). Corroborating biased indicators: Global and local agreement among objective and subjective estimates of printed word frequency. *Behavior Research Methods*, *41,* 452–471.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.

Yap, M. J. (2007). *Visual word recognition: Explorations of megastudies, multisyllabic words, and individual differences*. Unpublished doctoral dissertation, Washington University in St. Louis, St. Louis, MO.

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

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. K., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47,* 1–29.