Published in final edited form as:
J Exp Psychol Hum Percept Perform. 2013 June ; 39(3): 802–823. doi:10.1037/a0030859.

Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers

Victor Kuperman and
Department of Linguistics and Languages, McMaster University, Togo Salmon Hall 626, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4M2

Julie A. Van Dyke
Haskins Laboratories, New Haven, CT, USA

Abstract

The importance of vocabulary in reading comprehension emphasizes the need to accurately assess an individual's familiarity with words. The present article highlights problems with using occurrence counts in corpora as an index of word familiarity, especially when studying individuals varying in reading experience. We demonstrate via computational simulations and norming studies that corpus-based word frequencies systematically overestimate strengths of word representations, especially in the low-frequency range and in smaller-size vocabularies. Experience-driven differences in word familiarity prove to be faithfully captured by the subjective frequency ratings collected from responders at different experience levels. When matched on those levels, this lexical measure explains more variance than corpus-based frequencies in eye-movement and lexical decision latencies to English words, attested in populations with varied reading experience and skill. Furthermore, the use of subjective frequencies removes the widely reported (corpus) frequency-by-skill interaction, showing that more skilled readers are equally faster in processing *any* word than the less skilled readers, not disproportionally faster in processing lower-frequency words. This finding challenges the view that the more skilled an individual is in generic mechanisms of word processing the less reliant he/she will be on the actual lexical characteristics of that word.

As with any acquired skill, *experience* plays a crucial role in the development of reading competence. Greater exposure to printed materials trains the many subskills that are implicated in reading, including skills in phonological decoding, rapid access to orthographic information, and morpho-syntactic competence (e.g., Stanovich & West, 1989; Stanovich & Cunningham, 1992). Perhaps the most prominent consequence of accrued reading experience, however, is increased vocabulary size. Beyond the obvious observation that not knowing the meanings of words makes comprehension impossible, a number of studies have demonstrated that increased vocabulary size correlates negatively with cognitive effort in word recognition, as evidenced by reduced response latencies in lexical decision and word naming tasks (Butler & Hains, 1979; Yap, Balota, Sibley, & Ratcliff, 2012). These advantages are demonstrated not only for recognizing familiar words, but also recognizing *unfamiliar* words (Chateau & Jared, 2000; Llewelen, Goldinger, Pisoni, & Greene, 1993; Stanovich & West, 1989, among many others). Moreover, individuals with a higher vocabulary size also appear to require fewer exposures to learn new words (Perfetti, Wlotko, & Hart, 2005).

The primary role of vocabulary in reading comprehension underscores the necessity for accurate methods for assessing an individual's familiarity with particular words. The goal of

the present article is to provide a critical evaluation of current methods of assessing word experience, and to describe an alternative conceptualization of these methods. In particular, we highlight problems with the use of *word frequency* as an index of word familiarity. Word frequency has long been identified as a reliable and pervasive index of the cognitive effort associated with word recognition. Words that occur more frequently in language tend to be recognized faster than less frequent words, as reflected, for instance, in shorter eye-fixation latencies and shorter response times to the word in lexical decision and similar chronometric paradigms (Balota & Chumbley, 1984; Monsell, Doyle & Haggard, 1989; Rayner & Duffy, 1986; see also references in Rayner, 1998). Thus, it is unsurprising that word frequency appears as one of the core parameters in a number of computational models of lexical processing, including models of eye-movement control in reading (e.g., Reichle, Pollatsek, & Rayner, 2006; Engbert, Nuthmann, Richter, & Kliegl, 2005), as well as models of the lexical decision task (e.g., Ratcliff, Gomez, & McKoon, 2004; Balota, Yap, Cortese, & Watson, 2008).

Word frequency also plays a prominent role in current theorizing about the relationship between skill and experience in reading. Research into individual differences in reading reports robust interactions of word frequency by reading skill. In over a dozen studies with different combinations of experimental paradigms, manipulations and participant populations, a recurring pattern was observed: more proficient readers showed a weaker effect of word frequency (estimated as the count of a word's occurrences in a corpus) on behavioral outcomes of the task. The interaction is found in lexical decision and naming tasks that require recognition of isolated words: see, for example, Chateau and Jared (2000), Frederiksen (1978), Pugh et al. (2008), Shaywitz et al. (2003). For lexical decision in particular, Diependaele, Lemhöfer, and Brysbaert (submitted) demonstrated that the magnitude of the word frequency effect varies with an individual's vocabulary size, with weaker effects for those with larger vocabularies. Likewise, weaker word frequency effects in more proficient readers also emerged in several eye-tracking studies of sentence reading, the task that we employ here. Thus, Ashby, Rayner, and Clifton (2005) compared eye-movement latencies between cohorts of average and skilled readers (where skill was measured via the Nelson-Denny test of reading comprehension) as a function of manipulated word frequency. Both cohorts proved to be slower when reading low-frequency and low-predictability words, but the average readers were slowed down more (a 40 ms contrast between low- and high-frequency words) than skilled readers (an 11 ms contrast). Also, the contrast between under-performing readers and controls was much smaller for high-frequency words than it was for low-frequency words (6 ms vs 35 ms). Qualitatively similar interactions in eye-movement latencies and numbers of fixations were observed in the study of dyslexic vs. normal readers of German (Hawelka, Gagl, & Wimmer, 2010) and in the non-college-bound population of English-speaking young adults in Kuperman & Van Dyke (2011). Interactions of lexical item characteristics by participant characteristics have also emerged with respect to the effect of aging on reading, and in studies of reading acquisition in children. For instance, older readers showed overall slowing, as well as stronger effects of word frequency on eye-movement, lexical decision and naming latencies as compared with younger readers (Allen, Madden & Slane, 1995; Balota & Ferraro, 1996; Laubrock, Kliegl & Engbert, 2006; Rayner, Reichle, Stroud, Williams & Pollatsek, 2006; Spieler & Balota, 2000; see however the across-tasks differences in Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004). Rayner et al. suggested that the more prominent role of word frequency in recognition speed in older readers results from their greater experience with recognizing printed words gained over the lifespan: this account capitalizes on the observation that older readers tend to perform better in vocabulary tests that younger controls (O'Dowd, 1984).

The most common interpretation for these interactions makes skilled word recognition akin to object recognition: overlearned objects can be recognized automatically as a single unitized object, whereas less familiar items require more deliberate, piecemeal processing. A quote from a recent paper (Yap et al., 2012) elaborates the implications of this view:

> Consider the general hypothesis that as readers acquire more experience with words, they become increasingly reliant on automatic lexical processing mechanisms (LaBerge & Samuels, 1974; Stanovich, 1980). In this case it is possible that as automatic mechanisms develop, [recognition of] words may be less influenced by lexical characteristics (Butler & Hains, 1979), and even context (Stanovich & West, 1983; Yap et al., 2009).

Thus, the actual orthographic form of a particular word is thought to be implicated to a lesser extent if the reader possesses superior reading skill, gained via extensive experience with reading any and all words (and critically – this does not relate directly to experience with the particular word of interest). This conclusion has recently been challenged as potentially spurious by Faust, Balota, Spieler, and Ferraro (1999) and Yap et al. (2012), following up on Butler and Hains (1979). Their observation was that the magnitude of lexical effects on word recognition latencies in a participant is positively correlated with the magnitude of that participant's recognition latencies: longer reaction times to the word will elicit larger effects of word frequency (see also Diependaele et al., submitted). Matching the overall response latency across individuals or groups by applying the z-score transformation to their response latencies revealed the familiar shape of the frequency-by-skill interaction in word naming (readers with a higher score in the vocabulary size task showed *weaker* effects of word frequency on RTs), but a reverse shape of that interaction in lexical decision (readers with a larger vocabulary size showed *stronger* effects of word frequency on RTs). Thus, the frequency by skill interaction in chronometric tasks is argued to be an artifact of the base-rate effect, rather than a genuine experience-driven difference in processing difficulty. To anticipate our own results, we observed that the base-rate effect, removed either through a z-transformation or non-parametric transformations, cannot explain away all instances of the frequency by skill interaction. Nevertheless, we suggest an alternative reason that these interactions may be spurious: the use of inadequate estimates of word frequency itself.

Estimates of word frequency are typically based on counts of word occurrences derived from corpora of written or spoken language. To review the eye-tracking studies listed above, Ashby et al. (2005) and Rayner et al. (2006) used estimates of word frequency based on the 1-million word Francis and Kucera (1982) corpus; Hawelka et al. (2010) and the data analyzed in Laubrock et al. (2006) utilized the 6-million German component of CELEX lexical database (Baayen, Piepenbrock, & van Rijn, 1993); and Kuperman and Van Dyke (2011) used Burgess and Livesay's (1998) 130-million token HAL corpus. The advantages of using corpus-based counts for estimating word frequencies of occurrence in a language are well-known: such counts mirror language use over a large population rather than any single idiolect, and are perfectly replicable by all users of a corpus. Additionally, through selecting a corpus of an adequate size and coverage, the corpus data may be able to offer either a balanced representation of language genres and historical periods, or a detailed introspection into a specific genre, dialect or period; see Burgess and Livesay (1998) and Brysbaert and New (2009), among others, for a discussion of corpus selection for the purpose of obtaining stable word frequency estimates. Yet an inquiry into word frequency and its effect on human behavior faces several challenges. For instance, the lower range of frequency shows a greater variability both within and between corpora or subjects. This variability is negatively correlated with frequency and is more susceptible to poor measurment accuracy. Finally, speed and accuracy of word recognition behavior are asymptotically bound (physically determined at the low end and constrained to 100%

accuracy at the high end), and these limits have implications for the functional relation of performance indices with word frequency[1]. This paper specifically addresses one of these challenges, the within-subjects variability in frequency estimates and their impact on word recognition. These are precisely the advantages of corpus-based frequency counts which may become weaknesses in the exploration of individual differences, where accurate estimation of participant-specific experience is essential. It is precisely this consideration which motivates the use of a frequency measure that would be sensitive to the quality of lexical representation in an individual, also taking into consideration that person's reading experience or verbal skill (Landauer et al., 2011). Objective, replicable frequency counts from corpora that are not annotated for the expected or actual experience level of the language producers or comprehenders obviously fail to meet this desideratum (for a seminal exception see the school grade-based Zeno et al.'s (1995) corpus).

Alternative methods for assessing word frequency

In the quest for measuring individual familiarity with a word, an alternative to using corpus frequency is the use of introspective *subjective* measures of lexical knowledge (Thomspon & Desrochers, 2009). These latter measures appear beneficial for individual-differences research as they aim to capture characteristics of words *as learned by particular individuals*, as opposed to characteristics of words as estimated over an extensive language output of a large linguistic community, i.e. a corpus. Most commonly, two aspects of individual experience with words are addressed: (a) the degree of a word's entrenchment in one's mental lexicon, operationalized as a rating of individual familiarity with the word (*subjective familiarity*) or as a rating of how much exposure an individual has had with a word (*subjective frequency*), and (b) the primacy of the word in one's mental lexicon, operationalized as a subjective judgment of the age when the word was acquired (e.g., Balota, Pilotti, & Cortese, 2001; Balota et al., 2004; Brysbaert & Cortese, 2011; Brysbaert & Ghyselinck, 2006; Chafin, Morris, & Seely, 2001; Colombo, Pasini, & Balota, 2006; Connine, Mullenix, Shernoff, & Yellen, 1990; Gaygen & Luce, 1998; Gernsbacher, 1984; Gilhooly & Logie, 1980; Gordon, 1985; Juhasz, 2005; Juhasz & Rayner, 2003; 2006; Whalen & Zsiga, 1994; Williams & Morris, 2004). To compare the functional relationship between different subjective and objective frequency measures and their predictivity of linguistic behavior, a megastudy of Balota et al. (2001) collected subjective frequency norms for 2,938 words from a cohort of over 2,000 participants balanced in terms of age and formal education. The participants were asked to provide judgments of how often they encounter presented words. Balota et al. (2004) further incorporated the average by-item subjective frequency norms into a set of predictors of lexical decision and naming latencies collected for 2,428 monosyllabic words. These studies demonstrated that subjective frequency norms were preferable indices of individual experience with a word as compared with subjective familiarity ratings. In addition, Balota et al. (2001) showed that average subjective frequency norms strongly correlated with objective frequencies obtained from a number of corpora (e.g., $r = 0.78$ with the Zeno et al. (1995) objective frequencies), while the correlation was much stronger for words with higher objective frequency than those with lower frequencies, as also reported in Gordon (1985) for English words, and Thompson and Desrochers (2009) for French words.

The question of whether an individual's familarity with a word and their performance in word recognition tasks are best explained by any one or any combination of the following measures – objective corpus-based frequency, subjective frequency or AoA norms – is a topic of an ongoing debate: see, for instance, Brysbaert and Cortese (2011, pages 546–548) for a detailed review of conflicting evidence. We begin here with a few examples that favor

[1]We are indebted to Kevin Diependaele for this discussion.

subjective frequency as an explanatory factor. Balota et al. (2001) observed that subjective norms explained unique variance in both lexical decision and naming latencies over and above objective frequency. Furthermore, eye-tracking and chronometric studies that factorially manipulated subjective and objective (corpus-based) estimates of individual exposure to a word converged in that both manipulations elicit a reliable difference in behavioral responses (higher objective or subjective frequency words were read or responded to faster): The effect of subjective frequency or familiarity was particularly pronounced in words with low corpus frequency (cf. Gernsbacher, 1984; Chafin et al., 2001; Williams & Morris, 2004). Perhaps the strongest piece of evidence so far in favor of subjective frequency norms surfaced in Experiment 1 of Williams and Morris (2004): for words matched on moderate subjective frequency (mean 6.6 on the 7-point scale), eye-movement latencies did not differ between words with low vs. high corpus frequency. These studies corroborate the notion that subjective estimates play an important role in word processing, one not fully captured by objective frequency estimates.

On the other hand, arguments were made that subjective frequency ratings explain no unique variance over and above objective frequency and AoA norms in behavioral latencies, especially if a corpus is chosen with highly accurate frequency counts (Brysbaert & Cortese, 2011). Other concerns have been raised regarding the use of subjective lexical measures (Balota et al., 2001; Baayen, Feldman, & Schreuder, 2006; Thompson and Desrochers, 2009): namely, subjective judgments are arguably *contaminated* by irrelevant information, which can spuriously strengthen the correlation between the subjective frequency rating and the behavioral response obtained in word recognition. Thus, prior evidence with respect to the role of subjective frequency in word recognition is mixed. We address these issues below by pitting subjective frequency norms against lexical decision and eye-tracking data collected from populations with highly variable reading experience. We will argue that subjective frequency (or similar introspective measures) is crucial, and more advantageous than objective frequency measures, for an adequate characterization of individual variability in word recognition.

Our present inquiry into the tools and theories of characterizing individual lexical representations has the following structure. In the first part of the paper we examine the hypothesis that objective frequencies may be particularly poor estimates for readers with smaller vocabularies, an attribute frequently associated with poor reading skill and/or limited reading experience. We address this question via corpus-based simulations (Study 1), and by exploring the functional relationship between objective and subjective frequencies and testing whether it differs systematically and substantially according to experience (Study 2), or by corpus size and corpus genre coverage (Study 3). The second part of the paper (Study 4) directly contrasts subjective and objective frequency ratings as predictors of two types of behavioral measures, eye-tracking during natural reading and lexical decision. Our chief goal is to determine which of the two measures explains more variance in the record of eye-movements from readers with a range of skill-levels and reading experience. In particular, we focus on how the use of a particular frequency measure changes the nature of the frequency-by-skill interaction in word recognition latencies. To anticipate our results, we show that subjective frequency estimates cause this widely reported interaction to disappear. We conclude by exploring the implications of this result for theories of word recognition.

Objective word frequency counts: A Critical Review

Our first point of interest is to evaluate how well corpus-based frequency counts estimate relative frequencies of words in individual vocabularies. Given the strong correlation between vocabulary size and reading experience (Stanovich & Cunningham, 1992), this

examination will shed light on how well objective word frequencies approximate strengths of lexical representations across individuals at varying experience levels. A well-known property of vocabulary growth is that the proportion of rare words increases as the size of the vocabilary increases (e.g., Baayen, 2001). This suggests that large corpora will tend to overestimate the token counts of rare words in individual vocabularies and that this overestimation will be increasingly severe if vocabularies are small, as is likely for less experienced readers. We test this intuition using SUBTLEX, the 50 million-token corpus of subtitles to the US media and film (Brysbaert & New, 2009), and taking samples of varying sizes to mimic vocabularies of individuals with varying reading experience. We opted for this corpus because it best represents *oral* language– the modality in which both poor and good readers may have similar exposures, and also because the frequency counts from this corpus explain more variance in behavioral data than five other widely-used corpora (Brysbaert & New, 2009). Note that in using this corpus, we do not intend to imply that the SUBTLEX corpus, or any specific sample from it, are faithful representations of either the vocabulary size or content of an average US English speaker at any experience level. We maintain (and confirm across comparisons with several different corpora in Study 3) that word frequency effects have more to do with the nature of sampling than with the content or the size of a corpus.

Study 1: Stability of frequency estimates as a function of sample size

The accuracy of corpus-based frequency counts for hypothetical individual vocabularies was first tested against highest- and lowest-frequency words in the SUBTLEX corpus. A vector of $5*10^7$ word tokens was created, in which each token was represented the same number of times that it occurs in the corpus (for a detailed description of word selection see Study 2, which uses the same word list). Then we took random samples of 10^5, $5*10^5$, 10^6, $5*10^6$, 10^7, $2*10^7$ tokens from the vector, 1000 samples of each size. Table 1A reports the average number of word types in the sample of each size, and in the SUBTLEX corpus.

For each sample size, we computed the average relative frequency of the 10 most frequent words, including *s* and *t* derived from the possessive (*John's*) and negation (*don't*) forms, which SUBTLEX recognizes as separate words[2]. Table 1B reports the relative frequencies of these words in the entire 50-million SUBTLEX corpus, as well as across our samples. Each of the top 10 words had extremely stable estimates of frequency relative to the respective sample size (well within 3% of the relative change from the whole-corpus estimate), even in a sample that had the size of a typical academic paper (10,000 words). This result suggests that the relative amount (percentage of total words) of exposure to words that have high corpus-based frequencies would be virtually identical in individuals that vary drastically in their experience with the printed word.

The pattern was different in the low-frequency range: Part C of Table 1 reports the number of words that have a very low frequency of occurence (1–5 occurences) in SUBTLEX (31,156 word types), as well as the average percent of those words that occur in a sample of each size. Only an average of 0.04% of the 31,156 low-frequency SUBTLEX words occur in a 10,000 word sample, whereas the average percentage reaches 4.43% in a one million-word sample, and is ten times as large in a ten-million word sample (43.41%). This pattern corroborates the observation of Brysbaert and New (2009): "Whereas frequency counts for high-frequency words reach a stable level at a corpus size of 1 million, low-frequency words seem to require a corpus size of at least 16 million words for reliable estimates." It also confirms our intuition that frequency counts obtained from large corpora tend to

[2]The actual words are not important here, as the point has to do with the comparison of sample sizes. We repeated the simulation with 10 full lexical items (rather than "s" or "t") and, separately, with the 10 top content words (rather than pronouns, prepositions or determiners), and the patterns of results were the same as presented in Table 1B.

overestimate the likelihood of occurence of rare words in smaller samples by assigning larger-than-zero frequencies to a large percentage of words that are not part of an individual's vocabulary: the tendency to over-represent words is stronger as the sample size decreases.

Another implication of this analysis, well described in the literature on vocabulary growth and instruction (cf. Baayen, 2001 and references therein; Cunningham, 2005), is that doubling the sample size (through increasing one's exposure to print) will not lead to doubling the relative frequency of all words in the sample: the increment in relative frequency will be negligibly small for high-frequency words and much larger for low-frequency words. We quantified this intuition by comparing relative frequencies of 500 words in a sample of 5 million tokens (averaged over 1000 samples) and a corpus of 50 million tokens, thus, a 10-fold increase in vocabulary size. The pool of 500 words was created by randomly selecting 50 words from 10 frequency classes based on deciles of log-transformed frequency counts of the SUBTLEX corpus, thus representing the entire range from lowest (Class 1) to highest (Class 10) frequncies attested in the corpus. (For full details of the pool creation and distributional statistics of frequency classes, see Study 2 below.) For each frequency class, we computed the mean ratio of the word's relative frequency in the whole corpus and its average relative frequency in the 10% sample, see Table 2.

The distribution of ratios between samples differing in size by a factor of 10 was in line with data on extremely frequent and extremely infrequent words reported in Table 1. Relative frequencies of higher-frequency words (class 5 onwards) did not increase along with the sample size, as their average ratios oscillated around 1. However, the increment in relative frequency for words in classes 1–4 reported as the ratio in Table 1 was about twice as large as the 10-fold increase in the sample size, from 5 to 50 million tokens. (The same qualitative pattern was observed with samples of different size, as well as with samples taken from other corpora.) If we adopt a common assumption that a relative frequency of the word in an individual vocabulary is a valid index of the strength of the word's entrenchment in one's mental lexicon, it follows from the decreasing ratios of relative frequencies that individuals with extensive reading experience have similarly strong representations of common words and much stronger representations of rare words as compared to ones developed by individuals with limited reading experience. We further confirm this simulation-based result experimentally in Study 2.

To sum up, this simulation study sheds light on one of the research questions outlined above: corpus-based frequencies are likely to systematically vary in their accuracy of estimating word occurrences in lower-frequency words and increasingly so for smaller-size vocabularies (typical of less experienced readers). For the purposes of the present argument, this implies that corpus-based frequency estimates are not at all reflective of poor readers' true experience with a word, nor can they bring forward the systematically different experiences with common and rare words that readers of varying experience may have. We also made several predictions about the strengths of lexical representations in individuals that vary in their reading experience, based on simulated relative frequencies of words in individual vocabularies. In what follows, we test whether our predictions hold if another measure of a word's mental representation were used, namely, the word's subjective frequency ratings.

Study 2: Subjective frequency as a function of corpus frequency and reading experience

In this study we examine the relationship between frequency estimates derived from corpora versus those acquired through subjective report. As Study 1 suggests that vocabulary size has little impact on the proportion of high frequency words represented, we expect that more and less experienced readers will differ little in their frequency estimates for words that have

a high objective frequency. Furthermore, since corpus-based frequency counts tend to overestimate experience with rare words for poor readers relative to good ones, we expect that words of decreasing objective frequency would elicit consistently and increasingly lower ratings of subjective frequency from poor readers as compared to good readers (see Balota et al., 2001 for a similar approach to subjective frequency estimates in young vs old readers). We explored these predictions in a norming study that collected subjective frequency ratings for an entire range of corpus-based frequencies in a population that varied in their experience with the printed word, as indexed by their level of education.

Method—A list of words was compiled that occured in all of the following corpora: SUBTLEX (Brysbaert & New, 2009), the British National Corpus (henceforth, BNC), the English component of the CELEX lexical database (Baayen, Piepenbrock, & van Rijn, 1983), the Contemporary Corpus of American English (henceforth, COCA; Davies, 2008) and, separately, the Spoken subcorpus of the Contemporary Corpus of American English (henceforth, COCA-Spoken). The fact that the words were attested with larger-than-zero frequency counts in all these corpora ensured the consistency of word spelling as well as enabled the cross-corpora comparison presented below. The resulting pool included 21,702 words. We further assigned each word from the pool to one of 10 frequency classes: the frequency classes were based on the deciles of the log-transformed frequency counts of the SUBTLEX corpus. From each of the 10 frequency classes, 50 words were randomly selected. We manually checked words in the resulting random sample and substituted abbreviations and misspelled words with an equal number of words randomly selected from the same frequency class as the removed word. The process of word selection was iterated until a total of 500 words, 50 per frequency class, was selected. The resulting word list is the same as that used in Study 1.

To obtain ratings of subjective frequency for the 500 words, we used the web-based portal designed for “crowdsourcing” tasks, the Amazon Mechanical Turk (https://www.mturk.com/mturk/). The portal supports publishing experimental stimuli, and affords access to a pool of tens of thousands participants, from whom responses can be collected in a fast, cheap and reliable way: see Mason and Suri (2012); Munro et al. (2010), Schnoebelen and Kuperman (2010), Snow et al. (2008) for a detailed description of the web portal, a comparison of results obtained from the Amazon Mechanical Turk and conventional laboratory settings, and good practices of data collection and cleaning. For the purposes of our subjective frequency task, we randomly assigned the 500 target words to 25 equal-size lists. Each list was presented as one html-coded web page that had the following format: instructions, a list of 20 words with a drop-down menu that allowed participants to select a rating for each word, and a number of follow-up questions on the demographic, educational and linguistic background of the respondent. Significant aspects of the method, including the instructions and rating scale, were borrowed (with minor modifications) from the large-scale norming study of Balota et al. (2001), who collected subjective frequency ratings for over 2,000 words. The instructions read:

> Words differ in how commonly or frequently they have been encountered. Some words are encountered very frequently, whereas other words are encountered infrequently. The purpose of this study is to rate a list of 20 words with respect to frequency. We believe that your ratings will be important to future studies involving word recognition. When judging how frequently you encounter this word, you should select one of the following options: never, once a year, once a month, once a week, every two days, once a day, or several times a day.

An example presented the word “apple” with the drop-down menu to the right and the selected option “several times a day”.

The main body of the task was a list of 20 words, each supplemented by a drop-down menu with a blank line as a default selection and options from 1 (never) to 7 (several times a day) as responses. Further questions asked about the respondent's gender and age. We also asked participants to select "which educational level describes you best": the options included "No High School, Some High School, High School Graduate, Some college/no degree, Associate degree, Bachelors degree, Some graduate school, Completed graduate degree, Not listed, Declined to answer". We further asked "Which state did you live in the most between birth and 7 years old?" and, separately "What is your native language(s)? (For example, "English, Spanish")". The last question encouraged responders to list multiple languages as native ones, even though we confined our analysis to the data from self-reported monolingual speakers of US English, see below.

Twenty-five lists with 20 words each were published for data collection on the Amazon Mechanical Turk web portal. The reward for completing a list was $0.10, in line with the accepted practice of offering approximately half a penny per question (Mason & Suri, 2012). Using the Mechanical Turk settings, we limited the total number of responders to 40 per list. A particular participant could complete as many as 20 lists, but could not complete any single list twice.

Results—Collection of 20,000 ratings of subjective frequency (40 responses × 20 words × 25 lists) took approximately 6 hours. In the data cleaning phase, we removed all responses by individuals who took less than one minute to complete the task (less than 1% of responses). We further removed those responders who failed to provide a subjective frequency rating to any word in any list, or failed to indicate their level of education, country/state of early childhood, or their native language(s) (about 10% of responses). We further removed responses by individuals who indicated multiple languages or any language other than English as their native language, as well as responses by individuals who indicated any other location than one of the US states or predominantly English-speaking Canadian provinces. Our decision to ask about the birthplace and native language, rather than restrict the population of participants to monolingual speakers of English, was designed to encourage honesty and accuracy of responses. Importantly, we paid the reward regardless of whether the participant was removed from consideration on the grounds of native language or early childhood location, so there was no incentive for deception on the participants' part.

The trimming procedures reduced the number of responses to 10,198 or approximately 51% of the original data pool, with the vast majority of the data loss due to our monolingual English requirement. There were 69 participants in total, who contributed from 20 (1 list) to 440 (22 lists) responses to this pool: the median number of responses per participant was 40 responses and the mean was 148 (sd = 188). The distribution of responses per frequency class was relatively even after the data trimming, see Table 3, and all words received between 14 and 24 ratings (down from the total of 40 ratings before the data trimming):

Table 4 summarizes the distribution of educational levels over the pool of responses in our norms of subjective frequency. We took the participant's level of formal education as a proxy of his/her exposure to printed materials and reading experience: this assumption is supported by the nearly perfect correlation between years of education and the score in the vocabulary test administered to 968 responents to the 1989 U.S. General Social Survey (National Opinion Research Center, 1989). While much more varied and specific tests of verbal skill exist, we opted for education level as a measure that can be obtained by an experimenter without much effort. A more detailed characterization of participants' language proficiency is a desirable component for future research.

To estimate the effect of educational level on subjective frequency ratings we split our cohort of participants into two groups: "LoEd" with educational level equal to or lower than the Associate degree (3155 responses), and "HiEd" the educational level equal to or above Bachelors degree (7043 responses). Table 5 reports descriptive statistics of the responses and Figure 1 presents average subjective frequency ratings per frequency class (based on deciles of the log-transformed SUBTLEX frequency counts), plotted separately for the LoEd and HiEd populations. The two cohorts did not differ in terms of responders's age: HiEd: mean = 32.1 (sd = 9.6); LoEd: mean = 35.0 (sd = 9.0); t(69.8) = 0.58, p = 0.56.

Figure 1 plots non-linear functional relationships between subjective ratings and frequency categories defined according to objective (corpus) word counts. Distinct patterns were observed for the LoEd vs HiEd cohorts. The amount of variance that the nonlinear function of objective frequency class explained in the mean subjective frequency ratings was 0.895 for the LoEd ratings and 0.901 for HiEd ratings as revealed by linear regression models. Nonlinearity was modeled as restricted cubic splines with 3 knots (function rcs in library Design in R; see Harrell, 2001 for details).

Mean subjective frequency ratings occupied a smaller range of values in the HiEd cohort (2.68 through 5.56) as compared to the LoEd cohort (1.86 through 5.39). Mean ratings were also very similar between cohorts in the top frequency class 10, which represents highly frequent function and content words (e.g, *there, yet, apartment, husband*). Yet the ratings were lower for LoEd responders in all other frequency classes and the contrast between LoEd and HiEd ratings increased as the frequency class decreased. This pattern was in full accord with the predictions we derived from the results of our simulation in Study 1: corpus frequency accurately represented the highest-frequency words across all vocabulary sizes and experience levels, yet it overestimated lower-frequency words in all, and more strongly so, in smaller-size vocabularies. Between cohorts, classes 1–3 showed the maximum contrast in mean subjective frequency ratings, on the order of 0.8 points on a 7-point scale. Also, the mean ratings that HiEd responders showed for frequency classes 1–3 were virtually the same as the mean ratings of LoEd responders to words from frequency class 7. To make the comparison more concrete, there is a difference of one order of magnitude in the average corpus frequency for frequency class 3 (0.07 per million) and class 7 (0.7 per million).

The three lower frequency classes 1–3 did not show an appreciable difference in mean subjective frequency within either LoEd and HiEd cohorts. This may indicate the existence of a lower threshold on responders' sensitivity to corpus frequency: that is, the lower 30% of the frequency list in SUBTLEX appears to be represented equally weakly in the mental lexicon within either cohort. The practical implication is that behavioral responses to words within an experience-matched group are not expected to differ if words occur less than 0.1 times per million. Likewise, subjective ratings to words in frequency classes 4–6 do not reliably ($p > 0.1$ in all multiple comparison t-tests with the Bonferroni correction) vary within either cohort (average corpus frequencies between 0.12 and 0.37 occurrences per million). Yet, as Figure 1 demonstrates, subjective estimates of word frequency reliably vary by class over a much broader range, with an especially steep positive trend found over the higher frequency classes– 6 (average SUBTLEX frequency of 0.37 per million) to 10 (average SUBTLEX frequency of 680 per million). The observed variability across most of the frequency range in both cohorts strengthens the argument of Keuleers, Diependaele, and Brysbaert (2010) that experimental studies tend to under-utilize the frequency range available in a language by selecting words with frequencies that exceed 10 per million or an even higher threshold. For the 50-million SUBTLEX corpus imposing a threshold of 10 per million would confine the word selection to classes 9 and 10, and would ignore about 80% of the language vocabulary.

To sum up, our data indicate that one and the same objective frequency elicited reliably different estimates of subjective frequency depending on how high the corpus frequency was and how experienced with reading (i.e., educated) the individual was. We note that working out an optimal scale for subjective frequency (currently ranging from *never* to *several times a day*) and an optimal breakdown of corpus frequencies into frequency bins is a topic for further research. To the extent that subjective frequencies capture the entrenchment of lexical representations in an individual, these patterns revealed that words with high objective frequencies appear to be well established across both cohorts, while words with lower corpus frequencies are largely absent from vocabularies of the LoEd cohort but less so from those of the HiEd cohort.

This experience-driven difference is consistent with our predictions in Study 1 regarding the ratio of relative word frequencies in samples of different sizes. To quantify the link between Studies 1 and 2, we correlated the ratio of relative frequencies from Study 1 with the ratio of mean subjective frequency ratings in the HiEd cohort to those supplied by the LoEd cohort. The Spearman correlation of the two ratios across all frequency classes was positive and very strong: $\rho = 0.82$; $p = 0.006$. (The correlation remained similarly strong for all pairs of sample sizes and corpora that we tested: $0.75 < \rho < 0.85$, all $p < 0.01$). That is, a stronger discrepancy between relative word frequencies in a large and a small corpus sample came with a stronger discrepancy in subjective judgments of the individual exposure to the word, collected from individuals with ostensibly larger and smaller vocabularies. This implies that words with a given objective frequency elicit human judgments that differ systematically as a function of the responders' reading experience evidenced in vocabulary size. This points to the possibility that widely reported frequency-by-skill interactions may be spurious and a product of the choice of frequency estimates rather than a genuine behavioral pattern. We take up this point further in Study 4.

Study 3: A cross-corpora comparison

Several recent cross-corpora studies (e.g., Balota et al., 2004; Brysbaert & New, 2009; Burgess & Livesay, 1998) have made the case that the size of the corpus and the linguistic material it covers influence the adequacy of corpus-based word frequency estimates as predictors of behavioral observations. This consideration may be particularly important if a corpus is used for frequency estimates that is predominantly based on scientific or non-fiction literature that poor readers may have little experience with: a spoken corpus or a corpus of subtitles might make a more appropriate data source for individual-differences research. Here we test whether our observations are an artifact of the particular corpus we have chosen.

We identified 500 words from our Study 1 sample, which were taken from the SUBTLEX corpus, that also occurred in four additional (sub-)corpora: SUBTLEX, CELEX, BNC, COCA and COCA-Spoken. These corpora were chosen to vary in size, language variety and coverage of genres (further explanation below), in order to serve as a comprehensive testbed for the relationship between objective and subjective frequencies. We examined the relationship between subjective frequencies from our crowd-sourcing study and objective frequencies in each corpus by plotting them against each other for both educational levels (see Figure 2). Recall that our selection of words involved sampling 50 items from the 10 frequency classes defined as deciles of the log-transformed frequency list of SUBTLEX. Since frequency classes are not aligned across corpora, we ensured comparability of the corpora by plotting raw subjective word frequencies against log (base 10) word frequencies per million, rather than plotting subjective frequencies averaged per frequency class against the frequency class as in Figure 1. The trend lines are plotted separately for LoEd and HiEd cohorts for SUBTLEX (50 million words of speech or read speech as documented in the US 3lm and media subtitles; Panel A), the British National Corpus, or BNC (100 million word

corpus of British English with spoken and written subcorpora; Panel B), CELEX (18 million words of British English based on written sources; Panel C); the Corpus of Contemporary American English, or COCA (400-million word written and spoken corpus; panel D), and the spoken component of COCA (87 million word; Panel E).

The qualitative data patterns reported in Figure 2 were virtually identical across corpora. Words with the highest corpus frequency invariably elicited similarly high subjective frequency ratings in both cohorts, while the LoEd cohort showed consistently and increasingly lower subjective frequency ratings in words with lower corpus frequencies. The reliability of this frequency by skill interaction was confirmed in statistical regression models fitted to subjective frequency as a dependent variable for each of the five corpora (all ps < 0.01; in the interest of space we only present the model for SUBTLEX in Table 6; words with a log objective frequency below and above 3 standard deviations from the mean of the log-transformed objective frequency (1.1% of data points) were removed prior to fitting).

The cross-corpora comparison made it clear that none of the considered corpora offered frequency counts that were differentially sensitive to the strengths of lexical representations in readers of varying educational level. We conclude that – for the purpose of approximating those strengths via corpus-based frequencies – there is no qualitative advantage in using one corpus over any other.

Study 4: Direct comparisons of subjective vs. objective frequency

We proposed that subjective frequency ratings are more accurate estimators of the strength of the word's representation in the individual mental lexicon than objective corpus-based frequencies, especially when individual differences in reading experience are the subject of interest. We furthermore predict that (i) subjective frequency will explain more variance in reading behavior than corpus frequency, and (ii) the interaction of skill with word frequency would be attenuated if experience-specific subjective frequency ratings and not corpus frequencies are used. We tested these hypotheses by pitting predicted values of a word's subjective frequency and the respective objective word frequency values against two types of behavioral measures: online eye-movement measures of reading and ii) lexical decision measures, both obtained from readers that vary in their reading experience. The necessity for using predicted values stems from the fact that the words used in these previously conducted experimental tasks did not fully coincide with the 500 words for which we collected actual subjective frequency estimates in Study 2. Predictions were derived using the statistical model in Table 6 to estimate subjective frequencies for target words the model had not seen, based on a non-linear function of log objective frequency (from SUBTLEX) and participant education level: function predict.lm in library stats of the statistical package was used. For parity, we also calculated frequency estimates for all words in the eye-tracking experiment based on objective frequencies of the 500 target words: all patterns reported below were confirmed in the models that incorporated predicted rather than observed subjective frequencies. As Figure 3 illustrates, the predicted values faithfully replicated the functional relation between the alternative measures of frequency even for the unseen words from the eye-tracking experiment of Kuperman and Van Dyke (2011), see below.

Eye-tracking data—A detailed description of method, procedure and participants in the eye-tracking study is available in Kuperman and Van Dyke (2011), which reports effects of various skill measures on eye-movements, but was not concerned with subjective frequency. We use the same eye-movement record in our current evaluation, but review only basic facts of the procedure. Eye movements were collected for 81 English sentences read silently in isolation by 71 participants. Participants (43 females; 28 males) belonged to the age group of

16–24 (mean 20.8; SD = 2.6), and were not college-bound, i.e., their level of formal education did not exceed the equivalent of a high school diploma. All were native English speakers, and none had a diagnosed reading or learning disability. The eighty-one sentences served as fillers in three eye-tracking experiments, all administered to the same cohort of participants. Prior to participating in the eye-tracking experiment, the participants also undertook a battery of 18 reading ability tests. We opted to use scores on a single test, the Woodcock Johnson Word Identification task as an index of verbal proficiency. Of all measures in our battery, this task best characterizes both the strength of individual word representations in the participant's mental lexicon and decoding skill, so we expected this task to capture the same aspect of reading proficiency that word frequency captures. In addition, we expected that this measure would provide a good approximation of relative vocabulary size. This is based on our previous work (Van Dyke, Johns, & Kukona, submitted) with a comparable cohort of participants in which the Woodcock Johnson Word Identification task correlates with the Peabody Picture Vocabulary Test at $r = .75$, and at $r = .54$ after partialling out IQ).

The Woodcock-Johnson assessment followed standard administration procedures: participants read aloud a list of individual words, divided into sets of 6–8 words, each set increasing in difficulty. No contextual support is provided; participants were not expected to know the meaning of the words, but simply to pronounce them correctly using their prior experience with the word or their knowledge of letter-sound correspondences. Seventy-six trials are possible, however participants begin in the middle of the list and move backward if errors occur in order to establish their baseline. Participants then advance through the list until they make 6 errors in a row. Difficulty ranges from words like "achieved", "tremendous", "systematic" in the initial set (from the middle of the list), to "homogenization', "indissolubly" and "ubiquitous" in the final set. Average reliability of this task across the age range of our study participants is reported as .90 (McGrew & Woodcock, 2001); test duration is 5–8 minutes.

Scores in the task ranged from 500 (grade equivalent of 5th grade) to 588 (grade equivalent above the second year of graduate school), with a mean of 549 and a median of 545 (the grade equivalent of approximately the first year of college) and standard deviation of 22.6 (4.4 grades). Performance of the above-median sub-population thus was on par with the word identification skills expected for individuals with an associate degree (typically awarded after two years of college-level education) or higher. Based on this, we matched the HiEd cohort in the norming Study 2 (Bachelors degree or higher) to the subgroup of participants who performed above the median in the word identification task, and the LoEd cohort (No High School to Associate Degree) to the subgroup who performed below or at the median in this task. This matching overestimated the level of reading experience - and possibly inflated the subjective frequency ratings - for some participants in the below-median subgroup. Yet we opted for this method as a first approximation to establishing the respective roles of subjective and objective frequency as co-determiners of word recognition. We discuss the implications of our matching procedure below.

For the eye-tracking task, participants were seated in front of a 17-in. display with a refresh rate of 85.03 Hz with their eyes approximately 64 cm from the display. They wore an EyeLink II head-mounted eye tracker (SR Research, Mississauga, Ontario, Canada), sampling at a rate of 250 Hz from both eyes. Sentences were presented one at a time on a single line, with a maximum of 90 characters, using a monospace font. Participants were instructed to read each sentence for comprehension and told that they would be required to answer a comprehension question. Comprehension questions occurred on 55% of trials. For further details on how the battery of skill measures and the eye-tracking experiments were administered, including further details regarding participants, stimuli, calibration procedure

for the eye-tracker, as well as the presentation of stimuli and comprehension questions, see Kuperman and Van Dyke (2011).

We tested the amount of variance explained by predicted subjective frequency ratings and by log corpus frequencies against a dataset that included 90,627 behavioral responses of 71 participants to 853 unique words in 81 sentences. We fitted two models, one with a nonlinear function of Predicted Subjective Frequency (*PredSubjFreq*) and the other with a nonlinear function of Corpus Frequency (*CorpFreq*), to each of the following dependent measures: single fixation duration, gaze duration, regression path time and total fixation time[3]. Nonlinearities in critical predictors were modeled using restricted cubic splines with 3 knots. Additionally, all models included word length as a robust predictor of eye-movement latencies (see references in Rayner, 1998): the nonlinear relationship of this predictor to dependent variables was modeled using restricted cubic splines with 5 knots, the maximum number of knots that showed consistently reliable effects. The resulting models revealed strong non-linear effects of both subjective and objective frequencies, and of word length (all p*s* < 0.001; models not shown). In single fixation duration, the advantage of predicted subjective frequency over corpus frequency in the adjusted amount of explained variance was about 2.25% (4.00 vs 1.75); in gaze duration, the advantage was about 3.86% (9.44 vs 5.58), while in total fixation time predictive subjective frequency performed slightly worse – a decrease of 0.71% – than log corpus frequency (2.88 vs 3.59). We conclude that skill-matched subjective frequency ratings perform on par with customary corpus-based metrics in global eye-movement measures such as total fixation time, and perform much better than log corpus frequencies in early measures, which are taken to be more indicative of word identification (e.g., single fixation and gaze duration.). We also fitted linear mixed effects models to all dependent variables with random intercepts for subject and word, and random subject-specific slopes of subjective or objective frequency and fixed effects as outlined above. In all cases, models with subjective frequency explained more variance, and the standard deviation of the subject-specific slopes was smaller for subjective frequency than for objective frequency. This implies that smaller individual adjustments to the overall effect of subjective frequency are required, and thus subjective frequency is a more accurate measure than its corpus-based counterpart.

The interaction of word frequency by skill: We noted in the introduction that an oft-reported interaction of word frequency by skill has had a significant impact on theorizing about the mechanisms of word recognition in skilled vs. less skilled readers. However, a number of researchers have questioned the validity of this interaction, suggesting that it may be an artifact of the base-rate effect (Butler & Hains, 1979; Faust et al., 1999; Yap et al., 2012). To rule out the latter possibility, we applied the z-score standardization and a non-parametric standardization (subtracting the distribution's median from the value and dividing the difference by the interquartile range) to the participant-specific distribution of eye-movement latencies in our data for each dependent measure. Across measures, the canonical frequency-by-skill interaction was retained, in line with the speeded naming results and contra the lexical decision results of Yap et al. (2012). Even after overall processing times were matched between participants and hence between the LoEd and HiEd cohorts, the cohort of more experienced HiEd readers demonstrated weaker effects of objective frequency on standardized response times: we conclude that the base-rate effect was not the reason for the frequency-by-skill interaction in our eye-tracking data and further report non-standardized eye-movement latencies.

[3]The former two measures are more commonly associated with processes implicated in word identification, while the latter two are typically interpreted as correlates of sentence-wide processes: for formal definitions see e.g., Kuperman and Van Dyke (2011). It is important to recognize, however, that there is no one-to-one correlation between measures and stages of word or sentence processing (Rayner, 2009).

We considered a second reason that this interaction may be spurious; namely, that it is an artifact of the frequency estimates themselves, which are typically derived from corpora. To evaluate whether the use of subjective frequency estimates would affect the presence of this interaction, we fit two models to log gaze duration as a dependent variable. One model included the critical interaction of experience (LoEd vs HiEd) with a non-linear function of subjective frequency, and the other an interaction of that same factor with the non-linear function of corpus frequency: Nonlinearities were modeled with the help of restricted cubic splines with 3 knots. Both models had a restricted cubic splines function of word length (5 knots) as an additional predictor.

Statistical models fitted to log gaze duration confirmed our hypothesis. The model with predicted subjective frequency (labeled as *PredictSubjFreq* below) as a critical predictor showed that it did not enter into a reliable interaction with experience level ($p = 0.96$), Table 7. The interaction did not reach statistical significance either in the model with predicted subjective frequency as a linear predictor ($p > 0.5$; model not shown.) At the same time, a reliable interaction was observed between log SUBTLEX frequency (labeled as *CorpFreq* below) and experience level in the model for log gaze duration, as predicted by the previous literature ($ps < 0.003$ for both restricted components), Table 8.

The contrast between the the model with subjective vs. corpus-based frequency presented as Tables 7 and 8 comes out clearly in Figure 4. We plotted lowess smoother trend lines for gaze duration as a function of log SUBTLEX objective frequency (panel A) and predicted subjective frequency (panel B): both panels present separate trend lines for low- and high-education levels.

The effect of corpus frequency was stronger overall for the LoEd cohort than for the HiEd cohort, with a 200 ms (400 ms to 200 ms) reduction in the model-estimated gaze duration between the least and most frequent words in the LoEd cohort, and a 115 ms reduction (300 to 185 ms) in the HiEd cohort. The between-cohorts contrast for lowest-frequency words was then on the order of 100 ms, and about 15 ms for the highest-frequency words, see Panel A of Figure 4. The pattern in Panel B was markedly different. First, the range of subjective frequencies was smaller for the HiEd group, which followed directly from the smaller range of values shown by the participants in the HiEd cohort in the norming task (Study 2): we note that regression models still estimate their parameters for the entire range of subjective frequency values for both cohorts. Second, the slopes of the trend lines and model fits were nearly parallel, at least for the interval between values 3 and 7 of subjective frequency that are defined for both cohorts. This similarity in slopes explains the unreliable interaction effect in the regression model in Table 7. We note that for all dependent variables models with subjective frequency explain as much variance as the ones with the experience by corpus frequency interaction (differences in the amount of explained variance well within 0.5%): yet latter models are less parsimonious in that they invest three degrees of freedom to implement the two main effects and an interaction term, as opposed to one degree of freedom required for the main effect of subjective frequency.

Models with the same sets of predictors as in Tables 7 and 8 fitted to single fixation duration showed statistically significant interactions of both log corpus frequency by experience [lCorpFreq x LoEd: $\hat{\beta} = -0.039$; $SE = 0.009$; $p < 0.0001$] and a non-significant interaction of subjective frequency by experience [PredictSubjFreq x LoEd: $\hat{\beta} = -0.013$; $SE = 0.009$; $p = 0.125$] (non-linearity of critical predictors was not supported by any model). Models fitted to regression path time revealed a marginally significant interaction of log corpus frequency by experience [rcs1 × LoEd: $\hat{\beta} = -0.052$; $SE = 0.009$; $p < 0.001$; rcs2 × LoEd: $\hat{\beta} = 0.023$; $SE = 0.012$; $p = 0.052$], whereas the interaction of subjective frequency by experience was nowhere near the 0.05-threshold of statistical significance [rcs1 × LoEd: $\hat{\beta} = -0.023$; $SE =$

0.009; $p = 0.008$; rcs2 × LoEd: $\hat{\beta} = 0.0005$; $SE = 0.011$; $p = 0.963$]. In models fitted to total fixation time, the interaction of log corpus frequency by experience was significant [rcs1 × LoEd: $\hat{\beta} = -0.046$; $SE = 0.007$; $p < 0.0001$; rcs2 × LoEd: $\hat{\beta} = 0.064$; $SE = 0.010$; $p < 0.0001$], while the interaction of subjective frequency by experience was not [rcs1 × LoEd: $\hat{\beta} = 0.050$; $SE = 0.011$; $p < 0.0001$; rcs2 × LoEd: $\hat{\beta} = -0.024$; $SE = 0.014$; $p = 0.098$]. To sum up, subjective frequency did not enter into reliable interactions with experience, while log corpus frequency reliably interacted with experience throughout the eye-movement record. Taken together, these patterns suggest that the frequency by skill interactions reported across a number of studies are specific to the method of estimating word frequency via counts of word occurrence as attested in linguistic corpora. In our data, a unit of change in the individual familiarity with a specific word (arguably indicative of that word's entrenchment in the individual mental lexicon), corresponds to the same contrast in behavioral correlates both in more experienced and less experienced readers. Furthermore, the strong main effects of experience that we observed across the entire frequency range confirm the role of experience in developing skills that facilitate processing of any and all words.

Lexical Decision data—To further corroborate results with the eye-tracking data, we contrasted predicted subjective frequency and objective frequency measures as predictors in the lexical decision task. We utilized the British Lexicon Project (Keuleers et al., 2011), which reports lexical decision latencies and accuracy for 28,739 mono- and disyllabic English words for 78 participants. Self-reported educational level, ranging in this data set from high school level to the Ph.D. degree, was taken as a proxy of individual reading experience. The criterion for including a particular participant in the British Lexicon Project database into the HiEd cohort was similar to the one in our norming Study 2, a bachelor's or higher degree. There were 49 participants and 441,261 responses in the LoEd cohort, and 29 participants and 292,706 responses in the HiEd cohort. As described above, we took SUBTLEX frequency counts as objective estimates of the reader's familiarity with words, and we predicted subjective frequencies of unrated words based on the objective frequencies and the educational level of participants. The resulting predictors were pitted against 337,857 lexical decision latencies of correct responses to existing words in Keuleers et al.'s data. Models were fitted with the nonlinear effect of word length and a nonlinear effect of either predicted subjective or objective frequency: For all predictors, non-linearities were modelled with restricted cubic splines with 3 knots. Once again, experience-matched subjective frequency showed an advantage in the amount of explained variance of about 0.57% as compared to objective frequency (7.45 vs 6.88 percent of explained variance; model not shown). The advantage was replicated in linear mixed effects models with subjective or objective frequency as both the fixed and the random effect (see above); likewise there was less individual variability in the effect of subjective frequency on latencies, as demonstrated by a smaller standard deviation of the subject-specific random slopes.

While the advantage of subjective frequency appears small, it is substantial given the amounts of variance explained by a number of other investigations into factors that affect lexical decision data (e.g., Balota et al., 2001, 2004; Brysbaert & Cortese, 2011; Cortese & Khanna, 2007). Moreover, it is gained against objective counts from arguably the best available corpus, SUBTLEX (Brysbaert & New, 2009). The choice of corpus is particularly important in light of the finding that increasing the quality of objective frequency demonstrably decreases the amount of variance explained by subjective frequency estimates (Brysbaert & Cortese, 2011). The fact that subjective frequency explains as much (and even more) variance than objective frequency is particularly interesting because this has often not been the case (e.g., Balota et al., 2004). Our result seems to be a direct result of the fact that our subjective frequencies were matched for education level of our participants. To confirm

this, we fitted two models with (a) word length and nonlinear functions of either (b) subjective frequency norms *not* sensitive to reading experience (collected by Cortese & Khanna, 2007) or (c) log SUBTLEX frequency counts to mean lexical decision latencies reported for 2,407 words in the English Lexical Project (Balota et al., 2004). Nonlinearities were modeled with restricted cubic splines with 3 knots. The model with (a) and (b) as predictors explained 38% of variance, while the model with (a) and (c) explained 40% of variance in response times, that is, 2% in favor of corpus frequencies. Taken together, these results point to the usefulness of experience-sensitive subjective norms as co-determiners of reading behavior, especially in investigations of less-skilled reading.

Word Frequency by Skill in Lexical Decision: Figure 5, Panel A reveals the familiar objective frequency by experience interaction in our lexical decision data, with more experienced readers demonstrating overall faster processing times and a weaker effect of word frequency: the interaction of non-linear word frequency by cohort is significant ($p < 0.01$, model not shown); the non-linearity is modelled with restricted cubic splines with 3 knots. We further tested whether this interaction would hold when the between-subjects differences in the mean RT are removed: this would show the robustness of the interaction to the base-rate effect, i.e. the fact that longer RTs tend to show effects of a larger magnitude (see Yap et al., 2012). Figure 5, Panel B plots standardized (per subject) RTs against objective word frequency. A reverse pattern of the frequency by skill interaction was observed: relative to their overall lexical decision latencies, more experienced readers were more affected by differences in corpus word frequency than less experienced readers, especially in the low-frequency range ($p = 0.048$; model not shown).

This reverse pattern emerged even more prominently when subjective frequency estimates were pitted against standardized RTs, see Figure 5, panel C. Again, when the overall processing speed was controlled statistically, more experienced readers were more affected by subjective frequency estimates than less experienced readers, in line with reports of Butler and Hains (1979), Llewelen et al. (1993), Sears et al. (2008) and Yap et al. (2012). Given the metalinguistic nature of the lexical decision task, we are reluctant to fully attribute the reversal of the interactive pattern to the word recognition component of the task, especially since word naming latencies (Yap et al., 2012) and eye-movement latencies in the present study do not show such a reversal (for relevant aspects of the cross-task comparison see e.g., Kuperman, Drieghe, Keuleers, & Brysbaert, in press). Yet the findings from the lexical decision task are consistent with our argument that an advantage in experience does not lead to a disproportionate advantage with low-frequency words, nor does it correlate with a lesser reliance on such lexical properties as word frequency, contra earlier discussions of the frequency by skill interaction.

General Discussion

Our inquiry into the nature of alternative measures of individual experience with specific words makes two contributions. First, we demonstrate that subjective frequency ratings can yield more accurate estimates of word familiarity across an experiential range than those based on corpus frequency counts. Second, our choice of a measure that is more sensitive to individual reading experience suggest a need to reinterpret current accounts of the interplay between the quality of specific lexical representations, as indexed by frequency, and reading skill. In what follows, we elaborate on both issues.

Subjective and objective measures of word frequency

We established that subjective frequency ratings given to 500 words representing the entire objective frequency range have a systematically different effect on reading behaviors in more versus less experienced readers. Words with the highest objective frequencies elicited

high subjective ratings for all participants, while ratings for the lower frequency bands were strongly influenced by reading experience; ratings of less experienced readers decreased as the objective frequency of rated words decreased (Table 5 and Figure 1 in Study 2). We also demonstrated that vocabulary size, which represents a key difference between good and poor readers, differentially affects the accuracy of corpus-based frequency estimates according to the rarity of the word in the language. Relative frequencies of very common words were excellently approximated by relative word frequencies in a large corpus across sample sizes. However, relative frequencies of rarer words were systematically overestimated by corpus-based relative frequency counts, and increasingly so for small samples, which are analogous to limited vocabularies, and by extension, limited reading experience. In fact, the varying contrasts in the relative frequency of words that was driven by differences in sample sizes, and the contrasts in subjective frequency ratings for those same words driven by different levels of reading experience correlated very strongly at $\rho = 0.82$, $p = 0.006$. The words were chosen to represent the entire frequency range of what is considered the best available corpus of the English language, SUBTLEX (Brysbaert & New, 2009), yet all our observations held true when objective frequency counts were obtained from four other corpora (Study 3).

A further outcome is the observation that subjective frequency ratings collected from more and less experienced readers explained more variance than corpus frequencies in several behavioral measures when the ratings were matched for experience level of the participants. Although we recognize that further replication will be important, the current results point to the superiority of the subjective frequency measure over the standard method which assesses frequency from corpus-based counts. Our findings show that these counts are inaccurate estimators, especially for rare words and for less skilled readers. Thus, when research questions are concerned with assessing reading ability, the method for deriving frequency counts must be carefully considered. A potential alternative to our subjective method of frequency estimation is compiling corpora that thoroughly sample texts by the skill level of either text producers or comprehenders. The grade-annotated Zeno corpus of school materials and the spoken subset of BNC annotated for the speaker's education level exemplify this approach: unfortunatly, neither of the corpora is large enough to enable the representative coverage of frequency distributions by skill level.

We recognize that our present plea for the use of subjective frequency ratings must be weighed against the criticisms presented by opponents of this lexical measure. The main points of criticism are that: (i) subjective measures explain less variance than objective ones, and especially so, when good objective measures are selected (see discussion of what constitutes "good" measures in Zevin & Seidenberg, 2002 and Brysbaert & Cortese, 2011), (ii) experience-driven individual differences in subjective frequency are in fact differences in some other linguistic characteristics of words, and (iii) subjective judgments are confounded with multiple lexical properties that do not per se reflect the strength of lexical representations but might nonetheless modulate the correlation between subjective frequency ratings and the behavioral responses to word recognition. Our analyses clearly show that point (i) is incorrect; Study 4 demonstrates that subjective frequency measures actually accounted for *more* variance across tasks. In what follows, we discuss points (ii) and (iii).

It is a logical possibility that individual differences in subjective frequency, and the effect of this measure on reading behavior, is an effect of some other subjective measure in disguise (Stadthagen-Gonzales & Davis, 2006). This possibility is supported by the fact that subjective frequency correlates with a number of lexical variables, including age-of-acquisition (AoA) and imageability (Baayen et al., 2006; Brysbaert & Cortese, 2011; also see below). For instance, the advantage in subjective frequency that experienced readers

have may stem from their earlier encounter (earlier AoA) of rarer words. An earlier AoA of the word – according to some theories (reviewed e.g., in Juhasz, 2005) – translates into a higher resting activation level for that word and facilitates its recognition. The data of Kuperman, Stadthagen-Gonzales and Brysbaert (in press) rules out AoA as a possible source of the individual differences in subjective frequency, however. Kuperman et al. collected AoA ratings for 30,000 English content words from the SUBTLEX corpus using the same crowdsourcing methods employed here. Participants were asked to enter the age (in years) at which they thought they had learned the word as well as demographic information related to their education level. We found that, unlike subjective frequency norms, AoA ratings did not vary significantly by education level of responders, nor did the correlations of AoA ratings with corpus frequencies. Thus, the experience-driven differences in subjective frequencies of words are not explained by the differences in the age when these words were acquired.

With respect to point (iii), Baayen et al. (2006) reported correlations between subjective frequency ratings and a number of objective lexical measures. Aside of corpus word frequency, these included the ratio of a word's frequency in a written corpus to its frequency in a spoken corpus, word category (noun vs verb), the noun-to-verb frequency ratio for words that can figure as both nouns and verbs (*help*); orthographic density; and inflectional and derivational entropies (see Figure 4 in Baayen et al., 2006). These confounding factors were among the ones that affected lexical decision latencies: this led Baayen et al. to warn against using subjective frequency as a predictor of behavior, as this measure is the "off-line inverse of visual lexical decision". Similar considerations led Thompson and Desrochers (2009) to investigate objective and subjective lexical measures that share variance with subjective frequency ratings to 6,202 French words. The ratings were argued to share "bias variance" with measures like imageability, orthographic neighborhood, and spoken word frequency. Thompson and Desrochers concluded that both subjective and objective estimates of word frequency are flawed (the latter due to the sampling error in the low-frequency range which we discuss above) and are best used together for the purposes of characterizing one's familiarity with words, see also Balota et al. (2001). We believe that taking objective word frequency as a benchmark (even as critically as done in Thompson and Desrochers, 2009) leads to a potential methodological oversight, namely, an implicit assumption that objective word frequency is uncorrelated with all those lexical properties that "contaminate" subjective judgments and is, at least for some frequency bands, an unbiased estimator of the individual's familiarity with the word. The question arises: What if subjective frequency ratings are higher for verbs than for nouns because verbs are, on average, more frequent than nouns? What if subjective frequency ratings are higher for words with larger inflectional and derivational entropies et cetera, because such words tend to be more frequent? If subjective frequency ratings are a good approximation of corpus-based frequency counts, they would show correlations with lexical measures that would have the same polarity as the correlations between those measures and objective frequency counts.

We examined this series of questions by fitting two multiple regression models to the data reported in Balota et al. (2004) and reanalyzed in Baayen et al. (2006), available as dataset english in library languageR in the statistical software R. One model replicated Baayen et al.'s model with item-average subjective frequency ratings as a dependent variable and word category, the noun-to-verb frequency ratio, orthographic density, inflectional and derivational entropies, and spoken frequency counts based on the BNC corpus, as predictors. The second model had the same set of predictors, but written word frequency as a dependent variable. (We opted for spoken frequency instead of the written/spoken frequency ratio used in Baayen et al.'s original model to avoid circularity in the definition of the second model.) All predictors were statistically significant ($p < 0.05$) in both models. Moreover the polarity of regression coeffcients was identical across the two models, except for the effect of

inflectional entropy which was positively correlated with subjective frequency ratings and negatively with objective frequency counts. The sets of regression coefficients from the two models correlated at $r = 0.97$, $p < 0.0001$. We conclude that subjective frequency ratings are no more contaminated by correlations with lexical properties than objective frequencies are. In fact, with a single exception of inflectional entropy, the correlation structure of the subjective estimator faithfully mirrors that of the objective estimators: raters introspectively judge verbs – as well as words with a higher spoken frequency, words with a smaller orthographic density, words with a larger noun-to-verb frequency ratio, and words with a greater derivational entropy – to be more frequent because they *are* more frequent in written English. While we did not apply the same procedure to the data in Thompson and Desrochers' (2009), Table 1 in their paper indicates that all correlations between subjective frequency ratings and "contaminating" lexical measures, on the one hand, and objective frequency counts and those same measures have the same polarity. Presumably, even for French the pattern holds: raters judge more image able words and words with a higher number of orthographic neighbors as more frequent because they are indeed more frequent in the language.

The method of data collection that we employ here allows for a targeted estimation of the strength of lexical representations in populations varying along any of the dimensions that are pertinent to psycholinguistic research: age, skill, experience, education, L2 ability, clinical conditions, and others. Importantly, this approach is conceptually different from the studies that use as their diagnostics the frequency by skill interaction, or even the variability of subject-specific adjustments to the word frequency effect. The implicit assumptions of those studies is that there is a single word frequency distribution spanning over levels of skill and over individuals, and also that groups and individuals only vary in how strongly this distribution affects their word recognition performance. In contrast, the use of a subjective measure implies and operationalizes inherently different word frequency distributions across individuals or groups. It is further possible that individuals vary not only in their frequency distributions, but also in how strongly these distributions affect their performance. However, this source of variability is not confirmed by the present study. To sum up, we believe that the advantage of subjective lexical measures that we report here should be made even more prominent, especially in individual-differences research.

The role of experience and skill in word recognition

One of our critical findings is that the widely reported frequency-by-skill interaction is apparently an artifact of the commonly employed method of estimating quality of lexical representations via objective corpus-based word frequency counts. When experience-matched subjective frequency ratings were used as predictors of eye-movement measures in Study 4, they did not enter into a reliable interaction with our index of word identification skill, the skill that arguably comes closest to estimating the individual's experience with words. Thus, instead of observing a typical pattern of weaker effects of word frequency on good readers and its stronger effects on poor readers, we observed a strong main effect of the word identification skill, with poor scorers being generally slower than good ones, and parallel regression lines indicating an equally strong effect of the change in subjective frequency ratings on the change in behavioral latencies in good and poor readers.

To more completely understand the practical implications of using suboptimal frequency estimates consider a fictitious factorial experiment, in which groups of words are matched on low vs. high corpus frequency and then presented to readers of varying experience levels; much like the many published word-recognition experiments. Suppose that words in the low-frequency experimental condition were chosen from frequency class 5 and those in the high-frequency condition from frequency class 8. For our respondents and for the SUBTLEX-based frequency classes, this difference in corpus frequencies would amount to a

contrast between 2.29 and 3.22 (or 0.93 points) in the class-average subjective frequency norms for less experienced readers, and 3.15 and 3.83 (or 0.68 points) for more experienced readers, a relative change of about 27%. (The relative change is only slightly reduced (19%) if word selection is restricted to the customarily used frequency classes 9 and 10.) If the indices of reading behavior, such as eye-movement latencies, are approximated using *objective frequency*, the 27% relative change could inflate a 40 ms LF-HF contrast for experienced readers into a contrast of about 51 ms for poor ones, and a 100 ms contrast for experienced readers into a 127 ms contrast for poor ones. The apparent interaction of corpus frequency by skill would then be at least partially due to an inaccurate frequency estimator, namely, one that is not sensitive to experience-related differences in the quality of lexical representations. Thus, if a research question requires words to be matched on frequency, subjective estimates attuned to education levels or even more fine-grained indices of verbal proficiency may be a prefered measurement.

That the frequency-by-skill interaction was unreliable when a more accurate estimator of frequency was chosen, calls into question a theoretical perspective that has been unchallenged for decades and supported by dozens of studies across languages, experimental paradigms and populations (see references in the Introduction and counterevidence in Yap et al. (2012), Llewelen et al. (1993) and Sears et al. (2009)). The perspective holds that accrued reading experience (a) increases the number of lexical representations and strengthens the quality of those representations in the mental lexicon of an individual, and (b) facilitates the development of reading skill by automatizing mechanisms of lexical processing, including decoding, retrieval of the word's meaning, and reading comprehension. Crucially, this perspective also assumes that (c) the degree to which the quality of the word's mental representation is used towards recognition of that word is contingent on how automatic lexical processing is in the individual. It follows from (c) that the more skilled an individual is in generic mechanisms of word processing the less reliant he/she will be on the actual lexical characteristics of that word (see above the quotation from Yap et al., 2012). While we concur with (a) and (b), our data run counter to (c). The data indicate that one and the same contrast X in an individual's familiarity with a pair of words elicits the *same* contrast in behavioral response latencies in more experienced readers and less experienced readers. This implies that neither experience, nor changes in vocabulary size and other dimensions of reading ability that experience gives rise to, change the extent to which specific aspects of word form are utilized in word recognition. A superior skill for automatic processing gained via experience does not undermine the impact of the word's representation on its recognition speed either: it only reflects in the main effect of skill on the processing speed, with more skilled readers being equally faster in processing *any* word than the less skilled readers, not disproportionally faster in processing lower-frequency words.

A possible counterargument to our claims is as follows: "Word frequency is only one of many lexical properties that show different effects on word recognition as a function of increasing experience-driven automaticity of processing. While word frequency may be reasonably argued to vary from one reader to another, other properties, like orthographic word length, are genuinely objective. The number of letters in a word is independent of the reader's experience and hence the robust differential effects of word length on word recognition times in good vs poor readers cannot be explained away by any subjective measure of length"[4]. We address this criticism by directly investigating word length effects

[4]A need for a subjective measure of word length is not inconceivable. A recent study of Rayner, Slattery and Bélanger (2010) makes a case that the perceptual span of effective vision is larger for faster readers than for slower ones, and thus any given word may be subjectively longer (have more letters outside of the individual reader's perceptual span) for slower than for faster readers, and - given the observed correlation between experience and reading speed - may also be subjectively longer for less experienced than for more experienced readers.

in our data, and show that word-length by skill interactions arise due to a base-rate artifact, as suggested by several other researchers (Butler & Hains, 1979; Faust et al., 1999; Yap et al., 2012).

Considering our eye-tracking data first, we plotted partial effects of word length on raw gaze duration in milliseconds for the LoEd and HiEd cohorts in the left panel of Figure 6 (for definitions of cohorts see Study 4 above). For words longer than four letters there is a clear processing advantage for more experienced readers across the word length range, and the advantage is also stronger the longer the word in the range of 5–10 letters. The right panel of Figure 6 presents the effects of word length on gaze durations that were z-transformed for each subject: identical results were obtained with the non-parametric transformation based on the median and the interquartile range rather than the mean and standard deviation. As argued, for instance, in Yap et al. (2012) behavioral latencies transformed in such a way remove the correlation between the average latency and the magnitude of lexical properties. The right panel of Figure 6 clearly shows a weak numerical advantage of more experienced readers over less experienced ones, while the 95% confidence intervals indicate that the advantage is not statistically reliable. That the interaction between word length and experience/skill was statistically unreliable was confirmed by high p-values in the statistical models fitted to z-transformed first fixation duration, single fixation duration, gaze duration and total fixation time (all ps> 0.1; models not shown; non-linearities were modelled with the restricted cubic splines with 3 knots.) The respective models fitted to *raw* eye-movement latencies showed reliable (though demonstrably spurious) interactions between word length and experience.

Consideration of lexical decision latencies from the British Lexicon Project (Keuleers, Lacey, Rastle, & Brysbaert, 2010) corroborated this result. More experienced readers in the British Lexicon Project appeared to show a substantial, gradually increasing advantage when processing increasingly longer words, see left panel of Figure 7. Yet the curves of the word length effect were virtually indistinguishable when the lexical decision latencies were z-transformed per subject, removing the base-rate effect (see right panel of Figure 7.) Statistical models confirm the statistical significance of the word length by experience interaction in the untransformed data, and the lack thereof at the 0.01-level in the z-transformed data (models not shown).

Based on these two analyses, we conclude that - at least for healthy adult speakers of English - interactions of word length with experience are artifacts of the base-rate, arising because longer response times tend to show stronger lexical effects. Statistical removal of the base-rate effect also removes or even reverses the interactions, thus supporting our view that more experienced readers invest the same (if not larger) effort relative to their overall performance into reading a longer or less frequent word as the less experienced readers do. This, again, is evidence that runs counter to the notion that skilled readers rely less on lexical characteristics of words than poorer readers, who are more affected by longer or less frequent words (for related critique, see also Besner, Stolz & Boutilier, 1997).

Conclusions

If experience does not modulate utilization of lexical representations, what is its impact on word recognition? As argued above, experience changes the quality of lexical representations, and does so differently for different words and different individuals. Some aspects of this relationship are well-described, including the logarithmic relationship between word frequency of occurence and behavioral correlates of word recognition: ten exposures to an infrequent word may have a similarly strong impact on the quality of that word's mental representation as 100 exposures to a word that is well entrenched in one's mental lexicon (e.g., Murray & Forster, 2004; for independent effects of word frequency and

repetition see also, Raney & Rayner, 1995). Importantly, it may not be simply the number of exposures to a word – larger for good readers, smaller for poor ones, due to their differences in reading experience – that would give rise to individual variability. It may be that poor readers are not able to use the exposures they do get to create the kind of high quality lexical representations that skilled readers have (Perfetti et al., 2005). For example, readers who make fewer phonological discriminations due to poor phonological processing skills will not end up with the same quality of lexical representation after 100 exposures than someone without phonological problems would end up with, even if their level of reading experience is matched. The same holds true for readers with a limited learning capacity or a compromised long-term lexical memory, or any other behavioral or organic characteristic that impedes the entrenchment of mental lexical representation: in all these cases the readers would have to have a larger number of exposures to a word than readers without those characteristics to create a representation of the same quality. None of these scenarios can be accounted for by general-use corpora, however large and genre-balanced they are. In sum, as suggested by MacDonald and Christiansen (2002), the quality of lexical representations in an individual are jointly determined by biological differences in a number of verbal and broad cognitive skills together with the amount of exposure to printed words (reading experience). All these considerations point to the need to use more fine-grained measures of word experience, ones that are sensitive to both biological differences and levels of reading experience, especially where research goals involve understanding the factors that contribute to variation in reading skill. The present paper is a step in this direction.

Acknowledgments

Thanks are due to Marc Brysbaert, Kevin Diependaele and an anonymous reviewer for their comments on an earlier draft of this paper, and to Emmanuel Keuleers for providing access to the British Lexicon Project data. This work was supported by the NSERC Discovery grant RGPIN/402395-2012 to Victor Kuperman, and the following grants from the NIH National Institute of Child Health and Human Development: HD 058944 to Julie Van Dyke (PI), HD 056200 to Brian McElree (PI), HD 040353 to Donald Shankweiler (PI). Partial support also came from the NIH National Institute on Deafness and Other Communication Disorders via Grant DC 07548 to Carol Fowler (PI).

References

Allen P, Madden D, Slane S. Visual word encoding and the effect of adult age and word frequency. Advances in psychology. 1995; 110:30–71.

Ashby J, Rayner K, Clifton C. Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. The Quarterly Journal of Experimental Psychology Section A. 2005; 58A:1065–1086.

Baayen, RH. Word frequency distributions. Dordrecht: Kluwer Academic Publishers; 2001.

Baayen, RH. Analyzing linguistic data: A practical introduction to statistics using R. Cambridge: Cambridge University Press; 2008.

Baayen RH, Feldman LB, Schreuder R. Morphological influences on the recognition of monosyllabic monomorphemic words. Journal of Memory and Language. 2006; 55:290–313.

Baayen, RH.; Piepenbrock, R.; Gulikers, L. The CELEX lexical database (cd-rom). University of Pennsylvania; Philadelphia, PA: 1995. Linguistic Data Consortium

Balota D, Cortese M, Sergent-Marshall S, Spieler D, Yap M. Visual word recognition for single-syllable words. Journal of Experimental Psychology:General. 2004; 133:283–316. [PubMed: 15149254]

Balota D, Ferraro F. Lexical, sublexical, and implicit memory processes in healthy young and healthy older adults and in individuals with dementia of the Alzheimer type. NEUROPSYCHOLOGY. 1996; 10:82–95.

Balota D, Pilotti M, Cortese M. Subjective frequency estimates for 2,938 monosyllabic words. Memory & Cognition. 2001; 29(4):639. [PubMed: 11504012]

Balota DA, Chumbley JI. Are lexical decisions a good measure of lecical access? The role of word frequency in the neglected decision stage. Journal of Experimental Psychology: Human Perception and Performance. 1984; 10:340–357. [PubMed: 6242411]

Besner D, Stolz J, Boutilier C. The stroop effect and the myth of automaticity. Psychonomic Bulletin & Review. 1997; 4(2):221–225. [PubMed: 21331828]

Brysbaert M, Cortese M. Do the effects of subjective frequency and age of acquisition survive better word frequency norms? The Quarterly Journal of Experimental Psychology. 2011; 64(3):545–559. [PubMed: 20700859]

Brysbaert M, Ghyselinck M. The effect of age of acquisition: Partly frequency related, partly frequency independent. Visual Cognition. 2006; 13(7–8):992–1011.

Brysbaert M, New B. 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. 2009; 41(4):977. [PubMed: 19897807]

Burgess C, Livesay K. 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 and Computers. 1998; 30:272–277.

Butler B, Hains S. Individual differences in word recognition latency. Memory & Cognition. 1979

Chafin R, Morris R, Seely R. Learning new word meanings from context: A study of eye movements. Journal of Experimental Psychology: Learning, Memory and Cognition. 2001; 27:225–235.

Chateau D, Jared D. Exposure to print and word recognition processes. Memory & Cognition. 2000; 28(1):143. [PubMed: 10714145]

Colombo L, Pasini M, Balota D. Dissociating the influence of familiarity and meaningfulness from word frequency in naming and lexical decision performance. Memory & cognition. 2006; 34(6): 1312–1324. [PubMed: 17225511]

Connine C, Mullennix J, Shernoff E, Yelen J. Word familiarity and frequency in visual and auditory word recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1990; 16(6):1084.

Cortese M, Khanna M. Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. The Quarterly Journal of Experimental Psychology. 2007; 60(8):1072–1082. [PubMed: 17654392]

Cunningham, A. Vocabulary growth through independent reading and reading aloud to children. In: Hiebert, E.; Kamil, M., editors. Teaching and learning vocabulary: Bringing research to practice. Mahwa, NJ: Erlbaum; 2003. p. 45-68.

Davies, M. The Corpus of Contemporary American English (COCA): 410+ million words, 1990-present. 2008. Available online: http://www.americancorpus.org

Diependaele K, Lemhofer K, Brysbaert M. Explaining individual differences in the word frequency effect: Insights from first and second language word recognition. 2012 Submitted.

Engbert R, Nuthmann A, Richter E, Kliegl R. SWIFT: A dynamical model of saccade generation during reading. Psychological Review. 2005; 112:777–813. [PubMed: 16262468]

Faust M, Balota D, Spieler D, Ferraro F. Individual differences in information-processing rate and amount: Implications for group differences in response latency. Psychological Bulletin. 1999; 125(6):777–799. [PubMed: 10589302]

Francis W, Kučera H. Frequency analysis of English usage: Lexicon and grammar. Journal of English Linguistics. 1982; 18(1):64–70.

Frederiksen, J. Cognitive psychology and instruction. New York: Plenum; 1978. Assessment of lexical decoding and lexical skills in their relation to reading proficiency.

Gaygen D, Luce P. Effects of modality on subjective frequency estimates and processing of spoken and printed words. Perception & psychophysics. 1998; 60(3):465. [PubMed: 9599996]

Gernsbacher MA. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. Journal of Experimental Psychology: General. 1984; 113:256–281. [PubMed: 6242753]

Gilhooly K, Logie R. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. Behavior Research Methods & Instrumentation. 1980

Gordon P. Level-ordering in lexical development. Cognition. 1985; 21:73–93. [PubMed: 4092417]

Harrell, F. Regression modeling strategies. Berlin: Springer; 2001.

Hawelka S, Gagl B, Wimmer H. A dual-route perspective on eye movements of dyslexic readers. Cognition. 2010; 115(3):367–379. [PubMed: 20227686]

Juhasz B. Age-of-acquisition effects in word and picture identification. Psychological bulletin. 2005; 131(5):684–712. [PubMed: 16187854]

Juhasz B, Rayner K. Investigating the effects of a set of intercorrelated variables on eye n durations in reading. Journal of Experimental Psychology: Learning Memory and Cognition. 2003; 29(6): 1312–1317.

Juhasz B, Rayner K. The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. Visual Cognition. 2006; 13(7–8):846–863.

Keuleers E, Diependaele K, Brysbaert M. Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. Frontiers in Psychology. 2010; 1:1–174. [PubMed: 21833184]

Keuleers E, Lacey P, Rastle K, Brysbaert M. The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. Behavior Research Methods. 2010:1–18.

Kuperman V, Drieghe D, Keuleers E, Brysbaert M. How strongly do word reading times and lexical decision times correlate? combining data from eye movement corpora and megastudies. Quarterly Journal of Experimental Psychology. in press.

Kuperman V, Stadthagen-Gonzales H, Brysbaert M. Age-of-acquisition ratings for 30 thousand english words. Behavior Research Methods. in press.

Kuperman V, Van Dyke J. Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. Journal of Memory and Language. 2011; 65:42–73. [PubMed: 21709808]

LaBerge D, Samuels S. Toward a theory of automatic information processing in reading. Cognitive psychology. 1974; 6(2):293–323.

Landauer T, Kireyev K, Panaccione C. Word maturity: A new metric for word knowledge. Scientific Studies of Reading. 2011; 15(1):92–108.

Laubrock J, Kliegl R, Engbert R. SWIFT explorations of age differences in eye movements during reading. Neuroscience and biobehavioral reviews. 2006; 30:872–884. [PubMed: 16904181]

Llewelen M, Goldinger S, Pisoni D, Greene B. Lexical familiarity and processing efficiency: Individual differences in naming, lexical decision, and semantic categorization. Journal of Experimental Psychology: General. 1993; 122(3):316–330. [PubMed: 8371087]

MacDonald M, Christiansen M. Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). Psychological Review. 2002; 109(1):35–54. [PubMed: 11863041]

Mason W, Suri S. Conducting behavioral research on amazons mechanical turk. Behavior Research Methods. 2012; 44:1–23. [PubMed: 21717266]

McGrew, KS.; Woodcock, RW. Technical manual: Woodcock-johnson iii. Itasca, IL: Riverside; 2001.

Monsell S, Doyle MC, Haggard PN. Effects of frequency on visual word recognition tasks. Journal of Experimental Psychology: General. 1989; 118:43–71. [PubMed: 2522506]

Munro R, Bethard S, Kuperman V, Lai V, Melnick R, Potts C, et al. Crowdsourcing and language studies: the new generation of linguistic data. Naacl workshop on creating speech and language data with amazons mechanical turk. 2010

Murray W, Forster K. Serial mechanisms in lexical access: the rank hypothesis. Psychological Review. 2004; 111(3):721. [PubMed: 15250781]

O'Dowd S. Does vocabulary decline qualitatively in the old age? Educational Gerontology. 1984; 10(5):357–368.

Perfetti C, Wlotko E, Hart L. Word learning and individual differences in word learning reflected in event-related potentials. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005; 31(6):1281.

Pugh K, Frost S, Sandak R, Landi N, Rueckl J, Constable R, et al. Effects of stimulus difficulty and repetition on printed word identification: an fMRI comparison of nonimpaired and reading-disabled adolescent cohorts. Journal of Cognitive Neuroscience. 2008; 20(7):1146–1160. [PubMed: 18284344]

Raney G, Rayner K. Word frequency effects and eye movements during two readings of a text. Canadian Journal of Experimental Psychology. 1995; 49(2):151–173. [PubMed: 9183975]

Ratcliff R, Gomez P, McKoon G. A diffusion model account of the lexical decision task. Psychological Review. 2004; 111(1):159. [PubMed: 14756592]

Rayner K. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin. 1998; 124(3):372–422. [PubMed: 9849112]

Rayner K, Duffy SA. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity y, and lexical ambiguity. Memory and Cognition. 1986; 14:191–201. [PubMed: 3736392]

Rayner K, Reichle ED, Stroud MJ, Williams CC, Pollatsek A. The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. Psychology and Aging. 2006; 21:448–465. [PubMed: 16953709]

Rayner K, Slattery T, Bélanger N. Eye movements, the perceptual span, and reading speed. Psychonomic Bulletin & Review. 2010; 17(6):834–839. [PubMed: 21169577]

Reichle E, Pollatsek A, Rayner K. E–Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. Cognitive Systems Research. 2006; 7(1):4–22.

Schnoebelen T, Kuperman V. Using Amazon Mechanical Turk for linguistic research. Psihologija. 2010; 43:441–464.

Sears C, Siakaluk P, Chow V, Buchanan L. Is there an effect of print exposure on the word frequency effect and the neighborhood size effect? Journal of Psycholinguistic Research. 2008; 37(4):269–291. [PubMed: 18344000]

Shaywitz S, Shaywitz B, Fulbright R, Skudlarski P, Mencl W, Constable R, et al. Neural systems for compensation and persistence: young adult outcome of childhood reading disability. Journal of Biological Psychiatry. 2003; 54(1):25–33.

Snow R, O'Connor B, Jurafsky D, Ng A. Cheap and fast|but is it good?: evaluating non-expert annotations for natural language tasks. Proceedings of the conference on empirical methods in natural language processing. 2008:254–263.

Spieler D, Balota D. Factors influencing word naming in younger and older adults. Psychology and Aging. 2000; 15(2):225–231. [PubMed: 10879577]

Stadthagen-Gonzalez H, Davis C. The Bristol norms for age of acquisition, imageability, and familiarity. Behavior Research Methods. 2006; 38(4):598–605. [PubMed: 17393830]

Stanovich K. Toward an interactive-compensatory model of individual differences in the development of reading fluency. Reading research quarterly. 1980:32–71.

Stanovich K, Cunningham A. Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. Memory & Cognition. 1992; 20(1):51–68. [PubMed: 1549065]

Stanovich K, West R. On priming by a sentence context. Journal of Experimental Psychology: General. 1983; 112(1):1. [PubMed: 6221061]

Stanovich K, West R. Exposure to print and orthographic processing. Reading Research Quarterly. 1989:402–433.

Thompson G, Desrochers A. Corroborating biased indicators: Global and local agreement among objective and subjective estimates of printed word frequency. Behavior research methods. 2009; 41(2):452–471. [PubMed: 19363186]

Van Dyke J, Johns CL, Kukona A. Individual differences in sentence comprehension: A retrieval interference approach. 2012 Submitted.

Whalen D, Zsiga E. Subjective familiarity of English word/name homophones. Behavior Research Methods Instruments and Computers. 1994; 26:402–402.

Williams R, Morris R. Eye movements, word familiarity, and vocabulary acquisition. European Journal of Cognitive Psychology. 2004; 16(1/2):312–339.

Yap MJ, Balota DA, Sibley DE, Ratcliff R. Individual differences in visual word recognition: Insights from the english lexicon project. Journal of Experimental Psychology: Human Perception and Performance. 2012; 38(1):53–79. [PubMed: 21728459]

Zeno, S.; Ivens, SH.; Millard, RT.; Duvvuri, R. The educator's word frequency guide. Touchstone Applied Science Associates; 1995.

Zevin J, Seidenberg M. Age of Acquisition Effects in Word Reading and Other Tasks. Journal of Memory and Language. 2002; 47(1):1–29.

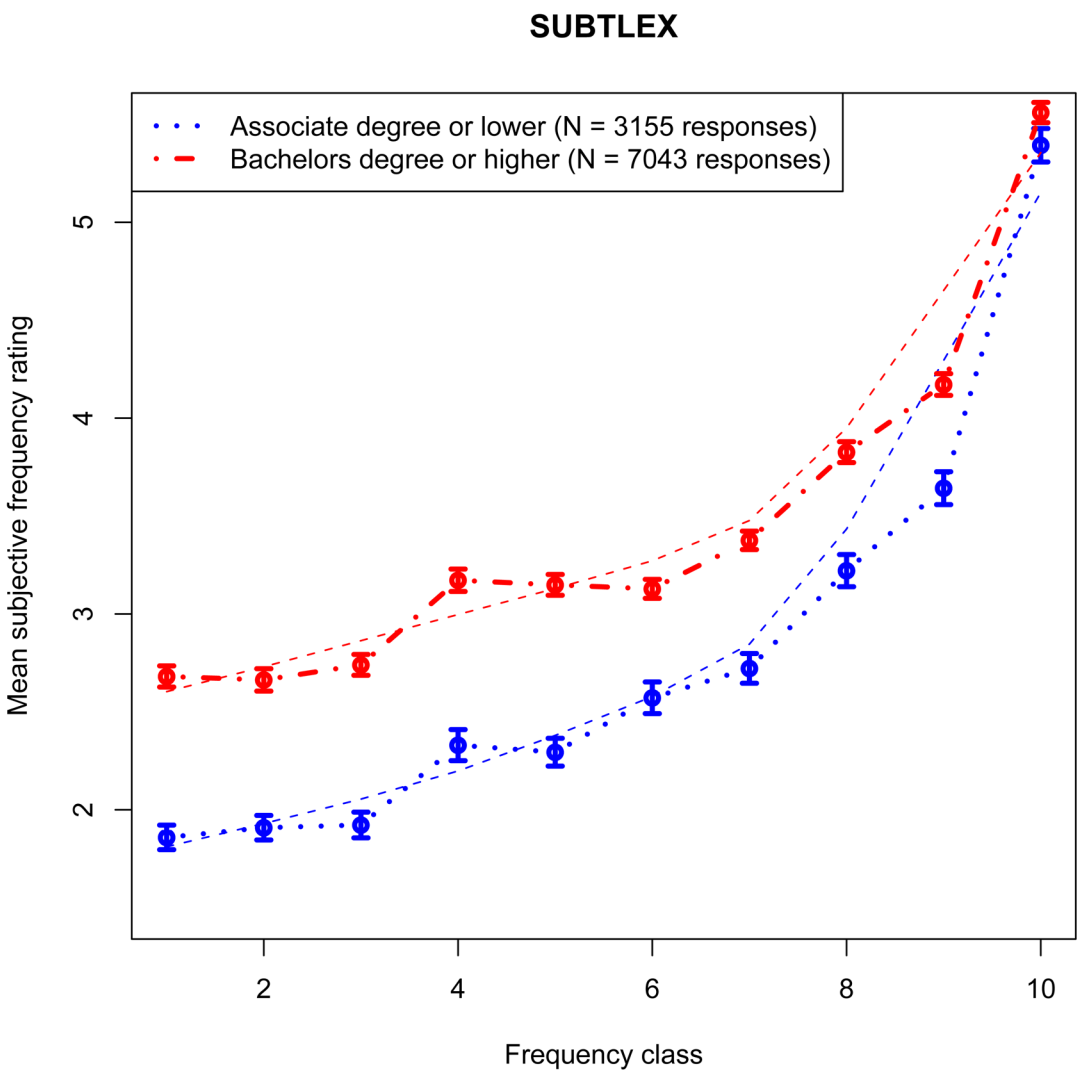

Figure 1.
Mean subjective frequency ratings per objective frequency class (based on log-transformed frequency counts in SUBTLEX). Mean subjective ratings are presented separately for the LoEd cohort with the Associate degree or lower (dotted line) and the HiEd cohort with the Bachelors degree or higher (dotdash line). Dashed lines are locally weight smoother (lowess) trend lines for respective cohorts. Error bars represent one standard error unit.

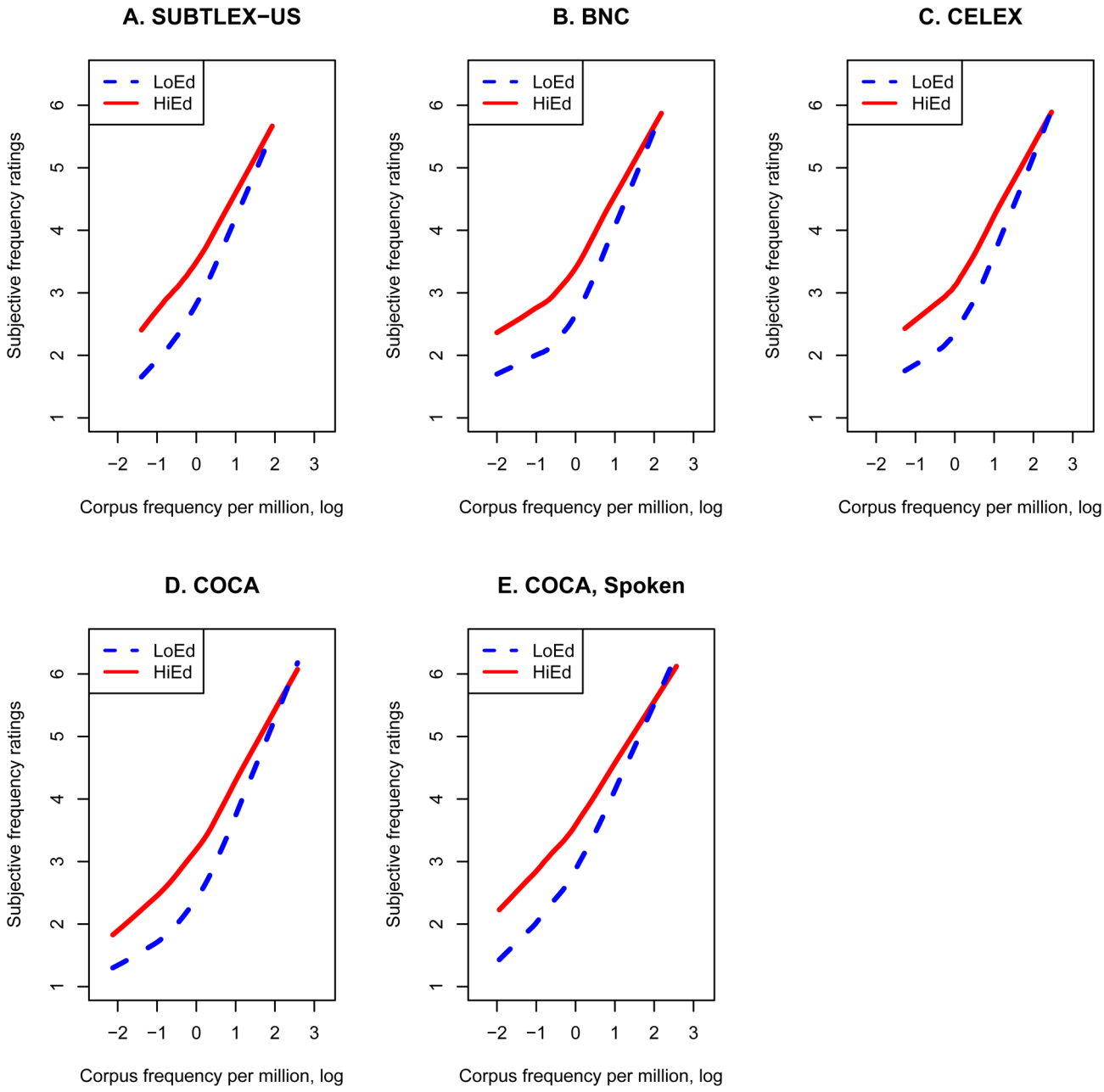

Figure 2.

Effects of corpus frequency on subjective frequency ratings presented as locally weighted smoother (lowess) trend lines for LoEd education level (dashed line) vs HiEd education level (solid line): panel A: SUBTLEX, panel B: BNC, panel C: CELEX; panel D: COCA; and panel E: the spoken subcorpus of COCA.

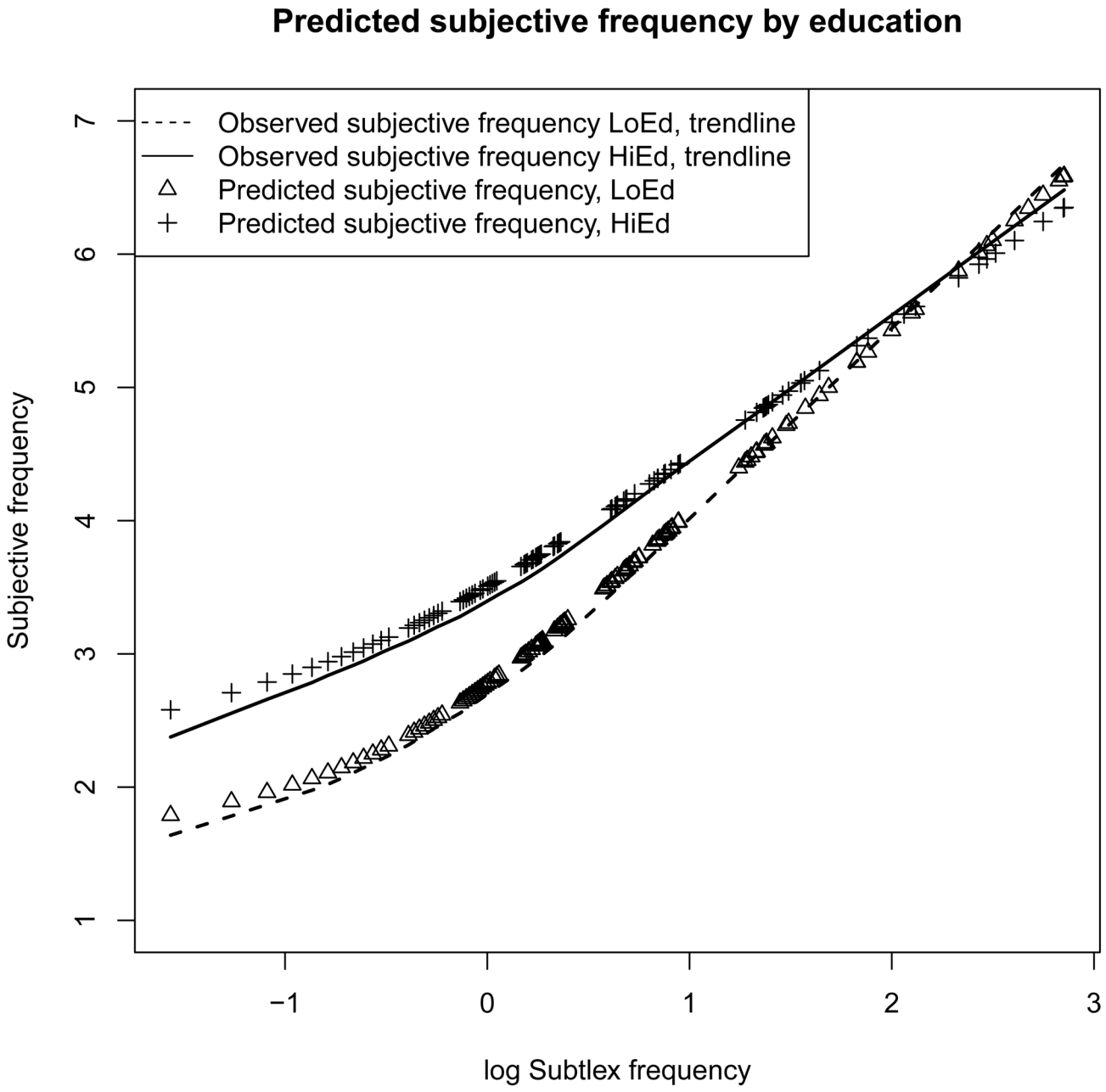

Figure 3.
Subjective frequencies predicted for 853 words in 81 stimuli sentences of Kuperman and Van Dyke (2011) as a function of word frequency in SUBTLEX and two levels of experience. Predicted values are plotted against trend lines of subjective frequency as a function of corpus frequency for LoEd and HiEd cohorts (Study 2).

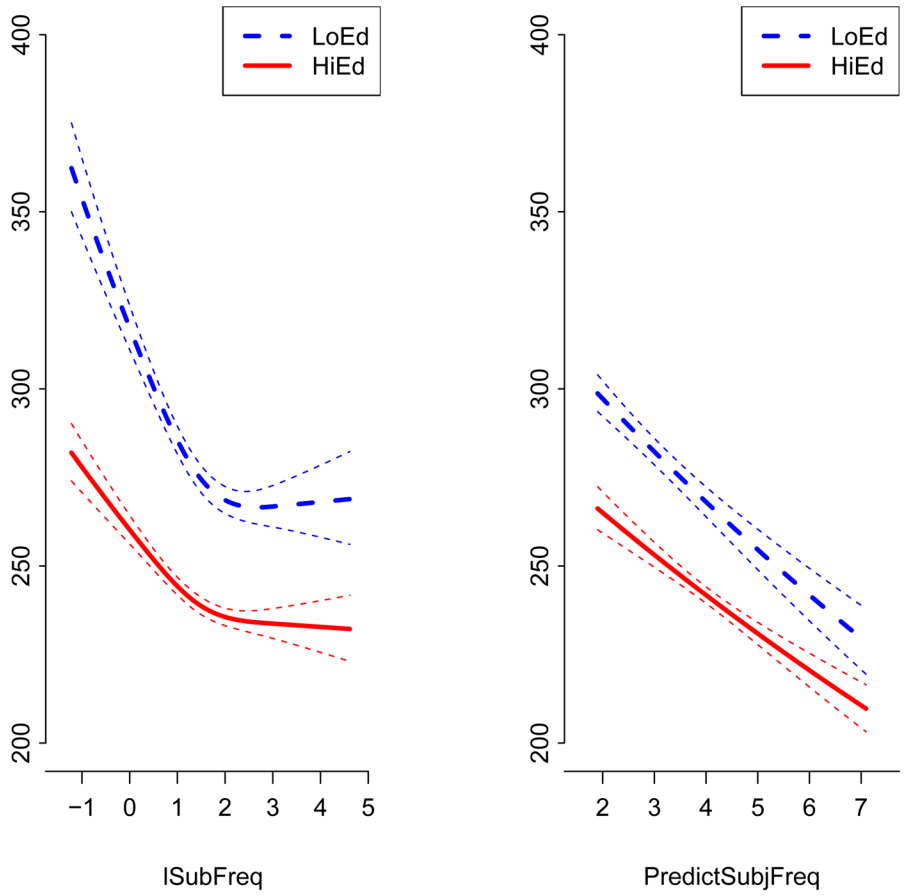

Figure 4.
Partial model-estimated effects and trend lines of observed effects of log corpus frequency (left panel), and of predicted subjective frequency on gaze duration (right panel). Results are reported for 71 participants dichotomized into low- and high-education cohorts (median grade equivalent = 13th grade, or one year of post high-school education). Model estimates of nonlinear effects are made using restricted cubic splines with 3 knots. Dashed lines represent 95% confidence intervals of model-estimated partial effects. Trend lines are produced using the lowess smoother function for low- and high-education cohorts separately.

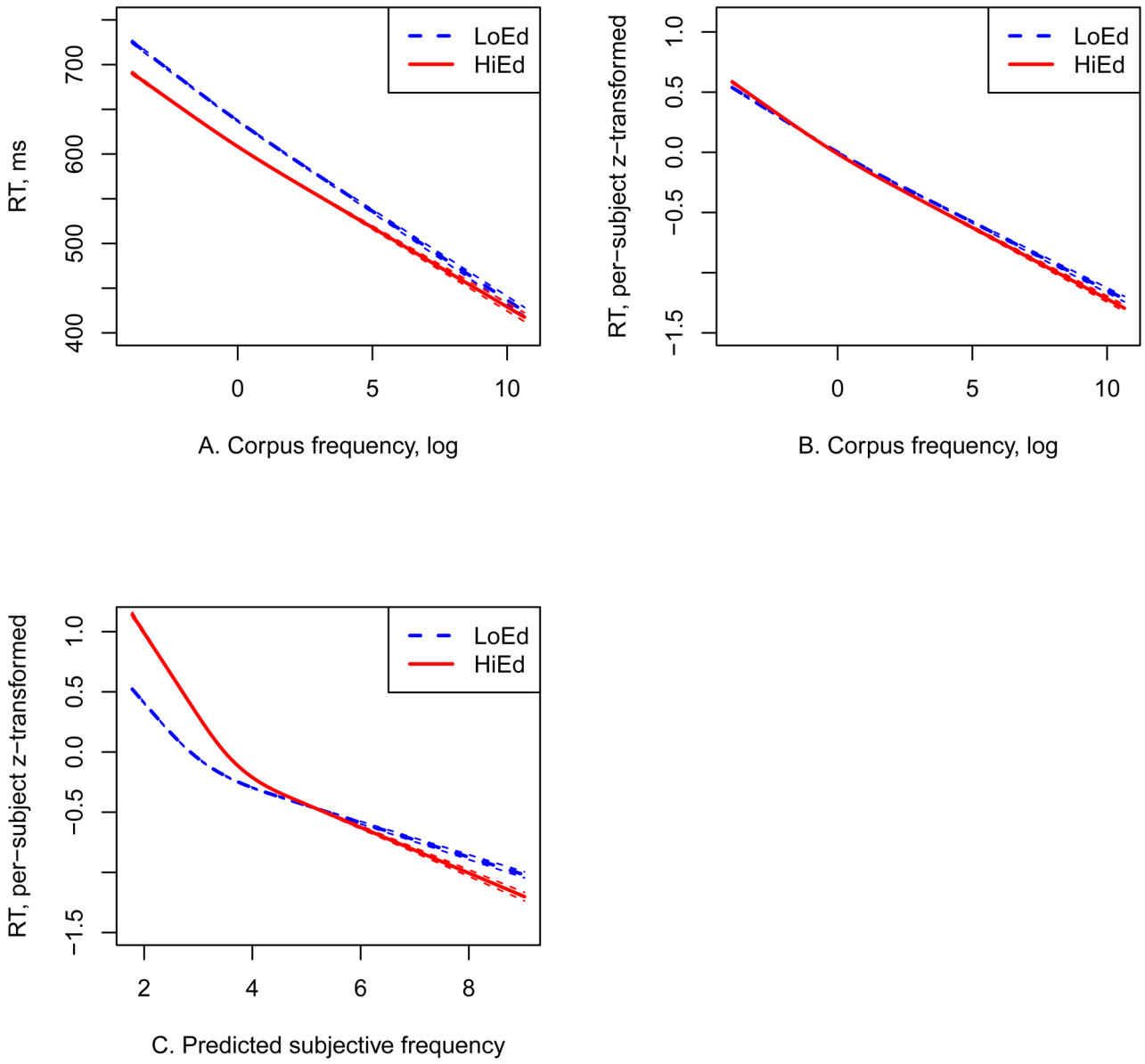

Figure 5.

Partial model-estimated effects of corpus frequency on raw lexical decision RTs (panel A) and on standardized (z-transformed per subject) RTs (panel B), as well as partial effects of predicted subjective frequency on standardized RTs (panel C). Results are reported for 78 participants dichotomized into low- and high-education (university bachelor or a higher degree) cohorts. Model estimates of nonlinear effects are made using restricted cubic splines with 3 knots. Dashed lines represent 95% confidence intervals of model-estimated partial effects.

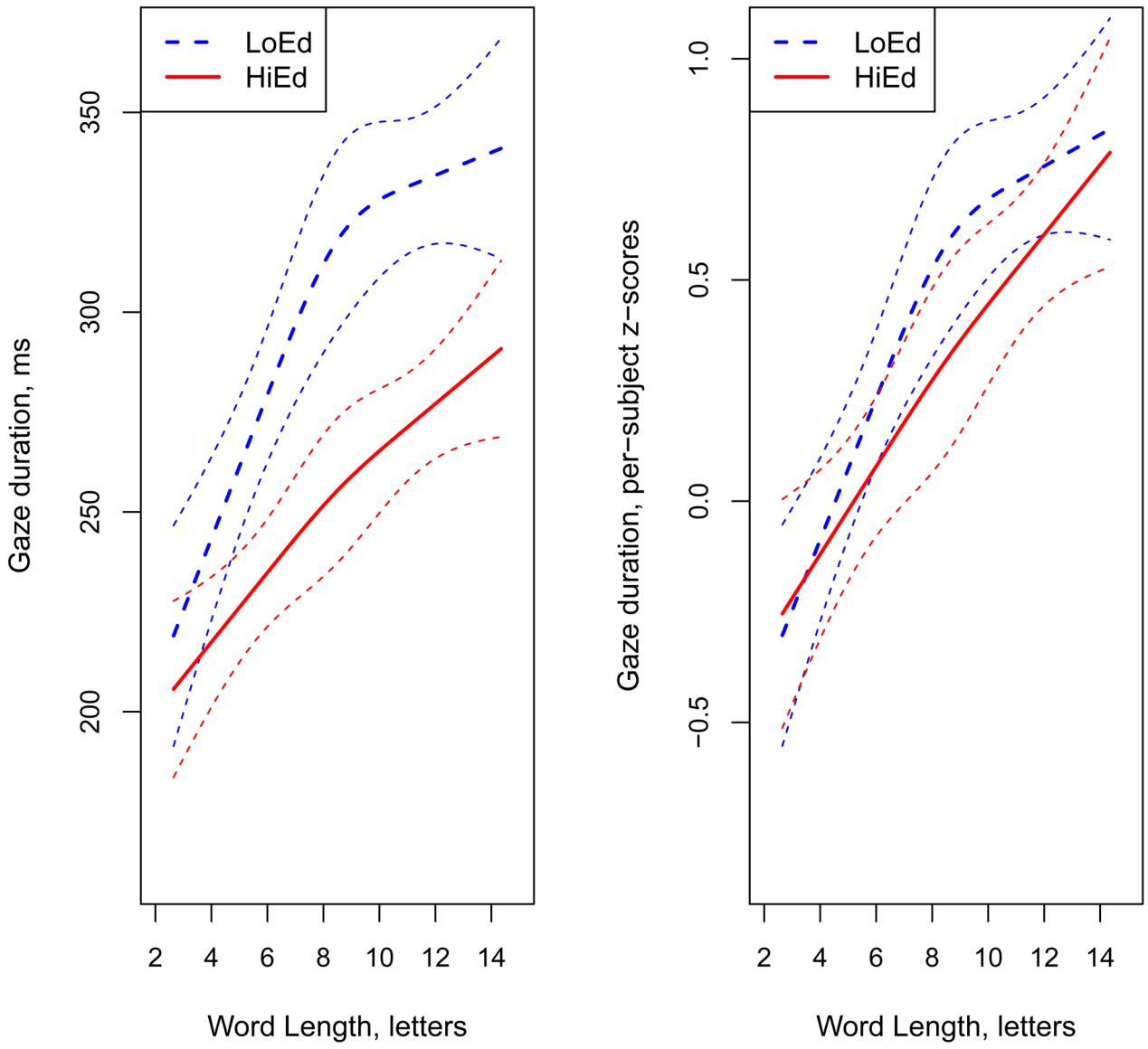

Figure 6.
Partial model-estimated effects of word length (in letters) on gaze duration (left panel)and on the z-transformed (per subject) gaze duration (right panel). Results are reported for 71 participants dichotomized into low- and high-education cohorts (median grade equivalent = 13th grade, or one year of post high-school education). Model estimates of nonlinear effects are made using restricted cubic splines with 3 knots. Dashed lines represent 95% confidence intervals of model-estimated partial effects.

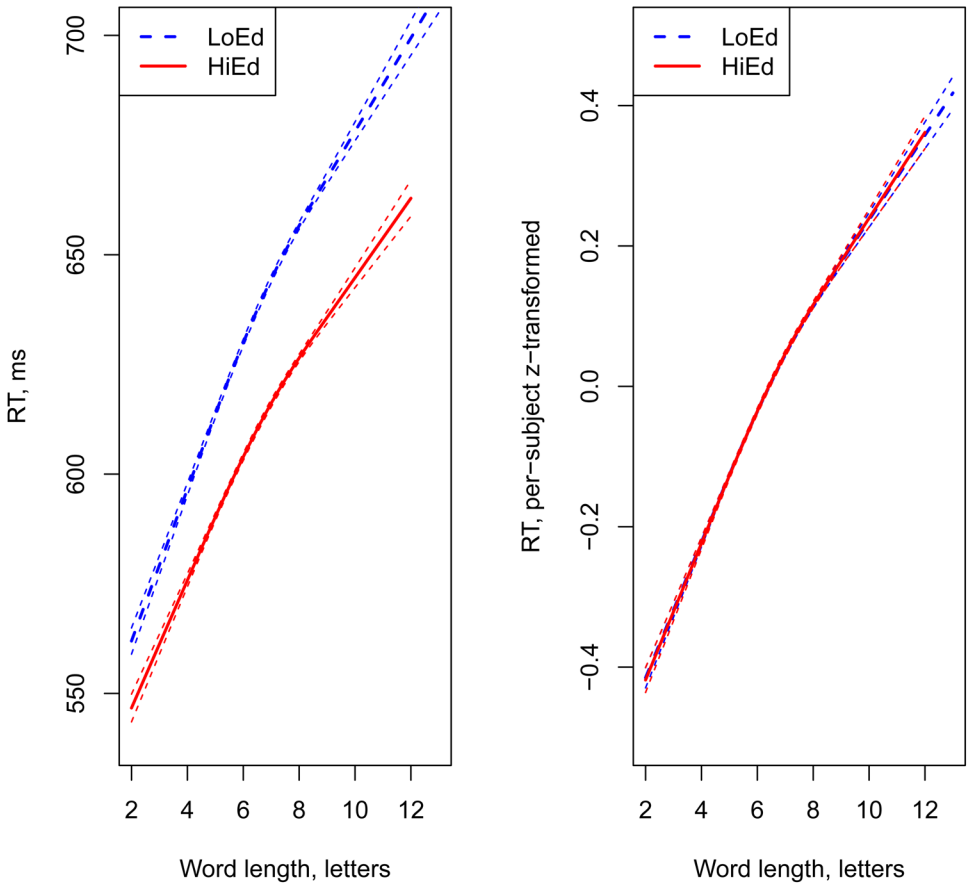

Figure 7.
Partial model-estimated effects of word length (in letters) on gaze duration (left panel) and on the z-transformed (per subject) gaze duration (right panel). Results are reported for 78 participants dichotomized into low- and high-education (university bachelor or a higher degree) cohorts. Model estimates of nonlinear effects are made using restricted cubic splines with 3 knots. Dashed lines represent 95% confidence intervals of model-estimated partial effects.

Table 1

Lexical statistics for SUBTLEX (50 million tokens) and word samples from SUBTLEX, ranging in size from 10^5 to $2 * 10^7$ tokens, averaged over 1000 samples of each size. A. Number of word types. B. Relative frequencies of 10 top frequency words. C. The number of words with 1–5 occurrences in SUBTLEX, and the average percentage of those words occuring in the sample of the given size.

	SUBTLEX: 5.0e+07	**1e+05**	**5e+05**	**1e+06**	**5e+06**	**1e+07**	**2e+07**
A. Word types	74286	2076.2	18630.4	25064.5	43501.5	52451.0	61719.7
B.	RelFreq	RelFreq	RelFreq	RelFreq	RelFreq	RelFreq	RelFreq
a	2.09	2.08	2.10	2.10	2.09	2.09	2.09
and	1.37	1.33	1.37	1.38	1.37	1.37	1.37
I	4.1	4.07	4.10	4.10	4.10	4.10	4.10
it	1.94	1.89	1.94	1.94	1.94	1.94	1.94
s	2.13	2.10	2.13	2.12	2.13	2.13	2.13
t	1.47	2.10	1.48	1.48	1.48	1.48	1.48
that	1.45	1.43	1.46	1.45	1.45	1.45	1.45
the	3.02	3.06	3.02	3.02	3.02	3.02	3.02
to	2.33	2.27	2.33	2.33	2.32	2.33	2.33
you	4.29	4.28	4.30	4.30	4.29	4.30	4.29
C.	LowFreq SUBTLEX (1–5)	Non-zero %	Non-zero %	Non-zero %	Non-zero %	Non-zero %	Non-zero %
	31156	0.04	2.13	4.34	21.77	43.41	86.73

Table 2

The ratio of a word's relative frequency in the 50-million token SUBTLEX corpus to its relative frequency in a sample of 5 million tokens (relative frequencies averaged over 1000 samples). Ratios are averaged per frequency class (1 - lowest frequency, 10 - highest frequency) and are based on a pool of 500 words, with 50 words per frequency class.

Frequency class	1	2	3	4	5	6	7	8	9	10
Between-sample ratio	2.234	2.083	1.672	1.344	1.020	1.020	0.996	0.998	1.012	1.003

J Exp Psychol Hum Percept Perform. Author manuscript; available in PMC 2014 June 01.

Table 3

Distribution of responses by frequency class.

Frequency class	1	2	3	4	5	6	7	8	9	10
Number of responses	1017	1014	1028	986	1030	1014	1026	1049	1021	1013

Table 4

Distribution of educational levels by the number of responses in the subjective frequency norming study.

Educational level	Number of responses
No High School	20
Some High School	0
High School Graduate	500
Some college/no degree	1797
Associate degree	838
Bachelor's degree	5223
Some graduate school	840
Completed graduate degree	980

Table 5

Values are reported per frequency class for the mean SUBTLEX corpus frequency (also per million in parentheses), median SUBTLEX frequency, and - separately for LoEd and HiEd cohorts - mean subjective frequency ratings (and standard deviations in parentheses), and numbers of observations. Corpus size of SUBTLEX is 50-million tokens.

FreqClass	MeanFreq (per million)	MedianFreq	MeanSubjFreq, LoEd	MeanSubjFreq, HiEd	N, LoEd	N, HiEd
1	1 (0.02)	1	1.86 (1.11)	2.68 (1.44)	312	705
2	2 (0.04)	2	1.91 (1.14)	2.66 (1.50)	328	686
3	3.46 (0.07)	3	1.92 (1.15)	2.74 (1.43)	309	719
4	5.95 (0.12)	6	2.33 (1.40)	3.17 (1.49)	312	674
5	9.84 (0.20)	10	2.29 (1.28)	3.15 (1.41)	323	707
6	18.21 (0.37)	18	2.57 (1.44)	3.13 (1.28)	320	694
7	34.45 (0.70)	34	2.72 (1.35)	3.38 (1.26)	313	713
8	70.65 (1.44)	68	3.22 (1.45)	3.83 (1.45)	312	737
9	222.43 (4.54)	207	3.64 (1.49)	4.17 (1.48)	316	705
10	33351.41 (680.64)	1795	5.39 (1.51)	5.56 (1.37)	310	703

Table 6

Linear regression model: subjective frequency ratings are predicted by the interaction of education level by log SUBTLEX frequency (logCorpFreq). The nonlinear effect of log frequency was modeled as restricted cubic splines with 3 knots: the model reports the n-th restricted component of predictor X as rcs(X)n. HiEd value of the factor Education is the reference level. Adjusted R-squared: 0.31.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.2370	0.0534	60.62	0.0000
LoEd	−0.9278	0.0965	−9.61	0.0000
rcs(logCorpFreq)1	0.4194	0.0527	7.95	0.0000
rcs(logCorpFreq, 3)2	0.4510	0.0581	7.76	0.0000
LoEd:rcs(logCorpFreq)1	−0.0865	0.0951	−0.91	0.3631
LoEd:rcs(logCorpFreq)2	0.3316	0.1049	3.16	0.0016

Table 7

Linear regression model for log gaze duration with the critical interaction between experience (labeled as LoEd vs HiEd) and predicted subjective frequency (*PredictSubjFreq*): the critical interaction is presented in bold. The nonlinear effect of frequency was modeled as restricted cubic splines with 3 knots, the effect of word length as restricted cubic splines with 5 knots: the model reports the n-th restricted components of predictor X as *rcs(X)n*.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.3790	0.0378	115.96	0.0000
LoEd	0.2333	0.0327	7.14	0.0000
rcs(PredictSubjFreq)1	−0.0352	0.0083	−4.26	0.0000
rcs(PredictSubjFreq)2	0.0137	0.0096	1.44	0.1513
rcs(WordLength)1	0.3114	0.0056	55.23	0.0000
rcs(WordLength)2	−3.2231	0.1112	−28.98	0.0000
rcs(WordLength)3	6.3621	0.2688	23.67	0.0000
rcs(WordLength)4	−3.0232	0.1811	−16.69	0.0000
LoEd:rcs(PredictSubjFreq)1	**−0.0234**	**0.0089**	**−2.64**	**0.0082**
LoEd:rcs(PredictSubjFreq)2	**0.0005**	**0.0113**	**0.05**	**0.9629**

Table 8

Linear regression model for log gaze duration with the critical interaction between experience (labeled as LoEd vs HiEd) and log corpus frequency (*CorpFreq*), based on the 50-million token SUBTLEX: the critical interaction is presented in bold. The nonlinear effect of frequency was modeled as restricted cubic splines with 3 knots, the effect of word length as restricted cubic splines with 5 knots: the model reports the n-th restricted components of predictor X as *rcs(X)n*.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.2872	0.0217	197.14	0.0000
LoEd	0.2613	0.0084	31.24	0.0000
rcs(CorpFreq)1	−0.0347	0.0044	−7.93	0.0000
rcs(CorpFreq)2	0.0123	0.0057	2.17	0.0303
rcs(WordLength)1	0.3133	0.0056	55.87	0.0000
rcs(WordLength)2	−3.1997	0.1108	−28.88	0.0000
rcs(WordLength)3	6.2506	0.2682	23.30	0.0000
rcs(WordLength)4	−2.9090	0.1810	−16.07	0.0000
LoEd:rcs(CorpFreq)1	**−0.0551**	**0.0059**	**−9.31**	**0.0000**
LoEd:rcs(CorpFreq)2	**0.0233**	**0.0078**	**2.99**	**0.0028**