Reassessing Word Frequency as a Determinant of Word Recognition for Skilled and Unskilled Readers

Victor Kuperman
McMaster University

Julie A. Van Dyke
Haskins Laboratories, New Haven, Connecticut

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 × 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 reliable he or she will be on the actual lexical characteristics of that word.

Keywords: word frequency, lexical decision, eye-movements, individual difference, reading skills.
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, Tse, & Balota, 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 (RTs) to the word will elicit larger effects of word frequency (see also Diependaele et al., 2012). 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 × 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 × 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 nonparametric transformations, cannot explain away all instances of the Frequency × 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 mentioned in this section, 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 the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995); 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 measurement accuracy. Finally, speed and
accuracy of word recognition behavior are asymptotically bound
(physically bound at the low end and constrained to 100% accu-
racy at the high end), and these limits have implications for the
functional relation of performance indices with word frequency.1
This article 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 that may become weaknesses in the
exploration of individual differences, where accurate estimation of
participant-specific experience is essential. It is precisely this
consideration that 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, Kireyev, & Panaccione,
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

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 (Thompson & Desro-
chers, 2009). These latter measures appear beneficial for
individual-differences research, as they aim to capture character-
istics 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, that is, a corpus. Most
commonly, two aspects of individual experience with words are
addressed: (a) the degree of a word’s entitlement 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, opera-
tionalized 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; Chaf-
fin, Morris, & Seely, 2001; Colombo, Pasini, & Balota, 2006;
Connine, Mullennix, Shernoff, & Yellen, 1990; Gaygen & Luce,
1998; Gernsbacher, 1984; Gilhooly & Logie, 1980; Gordon, 1985;
Juhasz, 2005; Juhasz & Rayner, 2003, 2006; Schreuder, 2006; Balota
et al., 2001; Thompson, 2008; Williams & Morris, 2004). Other
concerns have been raised regarding the use of subjective lexical
measures (Baayen, Feldman, & Schreuder, 2006; Balota et al., 2001;
Thompson & Desrochers, 2009), namely, subjective judgments are arguably contaminated
by irrelevant information, which can spuriously strengthen the
relation between the subjective frequency rating and the behav-
ioral 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 by pitting subjective frequency
measures against lexical decision and eye-tracking data collected
from populations with highly variable reading experience. We
argue that subjective frequency (or similar introspective measures)
is crucial, and more advantageous than objective frequency mea-
sures, for an adequate characterization of individual variability in
word recognition.

Our present inquiry into the tools and theories of character-
izing individual lexical representations has the following struc-

1 We are indebted to Kevin Diependaele for this discussion.
ture. In the first part of the article, 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 article (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 × 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 vocabulary 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 U.S. 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 U.S. 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 $5 \times 10^5$ 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). We then took random samples of $10^6, 5 \times 10^5, 10^5, 5 \times 10^6, 10^7, 2 \times 10^7$ tokens from the vector, 1,000 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)

<table>
<thead>
<tr>
<th>SUBTLEX:</th>
<th>5.0e + 07</th>
<th>1e + 05</th>
<th>5e + 05</th>
<th>1e + 06</th>
<th>5e + 06</th>
<th>1e + 07</th>
<th>2e + 07</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Word types</td>
<td>74286</td>
<td>2076.2</td>
<td>18630.4</td>
<td>25064.5</td>
<td>43501.5</td>
<td>52451.0</td>
<td>61719.7</td>
</tr>
<tr>
<td>B. RelFreq</td>
<td>2.09</td>
<td>2.08</td>
<td>2.10</td>
<td>2.10</td>
<td>2.09</td>
<td>2.09</td>
<td>2.09</td>
</tr>
<tr>
<td>and</td>
<td>1.37</td>
<td>1.33</td>
<td>1.37</td>
<td>1.38</td>
<td>1.37</td>
<td>1.37</td>
<td>1.37</td>
</tr>
<tr>
<td>it</td>
<td>1.94</td>
<td>1.89</td>
<td>1.94</td>
<td>1.94</td>
<td>1.94</td>
<td>1.94</td>
<td>1.94</td>
</tr>
<tr>
<td>s</td>
<td>2.13</td>
<td>2.10</td>
<td>2.13</td>
<td>2.12</td>
<td>2.13</td>
<td>2.13</td>
<td>2.13</td>
</tr>
<tr>
<td>t</td>
<td>1.47</td>
<td>1.48</td>
<td>1.48</td>
<td>1.48</td>
<td>1.48</td>
<td>1.48</td>
<td>1.48</td>
</tr>
<tr>
<td>the</td>
<td>3.02</td>
<td>3.02</td>
<td>3.02</td>
<td>3.02</td>
<td>3.02</td>
<td>3.02</td>
<td>3.02</td>
</tr>
<tr>
<td>to</td>
<td>2.33</td>
<td>2.33</td>
<td>2.33</td>
<td>2.33</td>
<td>2.33</td>
<td>2.33</td>
<td>2.33</td>
</tr>
<tr>
<td>you</td>
<td>4.29</td>
<td>4.28</td>
<td>4.30</td>
<td>4.30</td>
<td>4.29</td>
<td>4.30</td>
<td>4.29</td>
</tr>
<tr>
<td>C. LowFreq</td>
<td>31156</td>
<td>0.04</td>
<td>2.13</td>
<td>4.34</td>
<td>21.77</td>
<td>43.41</td>
<td>86.73</td>
</tr>
</tbody>
</table>

*Note. (A) Number of word types. (B) Relative frequencies of 10 top frequency words. (C) Number of words with 1–5 occurrences in SUBTLEX, and the average percentage of those words occurring in the sample of the given size.*
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 occurrence (1 to 5 occurrences) 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 1-million-word sample, and is 10 times as large in a 10-million-word sample (43.41%). This pattern corroborates the observation of Brysbaert and New (2009, p. 980): “ 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 overestimate the likelihood of occurrence 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 overrepresent 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 1,000 samples) and a corpus of 50 million tokens, thus, a tenfold 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) frequencies attested in the corpus. (For full details of the pool creation and distributional statistics of frequency classes, see Study 2 in the next section. 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 onward) 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 through 4, reported as the ratio in Table 1, was about twice as large as the tenfold 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 compared with ones developed by individuals with limited reading experience. We further confirm this simulation-based result experimentally in Study 2.

In sum, this simulation study sheds light on one of the research questions outlined in this article: 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, because 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 compared with 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 occurred in all of the following corpora: SUBTLEX (Brysbaert & New, 2009), the British National Corpus (henceforth, BNC), the English component of the
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 1,000 Samples)

<table>
<thead>
<tr>
<th>Frequency class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-sample ratio</td>
<td>2.234</td>
<td>2.083</td>
<td>1.672</td>
<td>1.344</td>
<td>1.020</td>
<td>1.020</td>
<td>0.996</td>
<td>0.998</td>
<td>1.012</td>
<td>1.003</td>
</tr>
</tbody>
</table>

Note. 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.

CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995), 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 here. 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 of 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), and Snow, O’Connor, Jurafsky, and Ng (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 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,” “bachelor’s degree,” “some graduate school,” “completed graduate degree,” “not listed,” and “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)?” (e.g., 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 U.S. English.

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 hr. In the data cleaning phase, we removed all responses by individuals who took less than 1 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 U.S. 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 or 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 respondents to the 1989 U.S. General Social Survey (Davis & Smith, 1998). Although 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’s degree (3,155 responses), and “HiEd” the educational level equal to or above bachelor’s degree (7,043 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, M = 32.1 (SD = 9.6); LoEd, M = 35.0 (SD = 9.0); t(69.8) = 0.58, p = .56.

Figure 1 plots nonlinear functional relationships between subjective ratings and frequency categories defined according to objective (corpus) word counts. Distinct patterns were observed for the LoEd versus HiEd cohorts. The amount of variance that the nonlinear function of objective frequency class explained in the LoEd versus 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 three 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) 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 through 3 showed the maximum contrast in mean subjective frequency ratings, on the order of 0.8 points on a 7-point scale. In addition, the mean ratings that HiEd responders showed for frequency Classes 1 through 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, Classes 1 through 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 through 6 do not reliably (p > .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, Classes 6 (average SUBTLEX frequency of 0.37 per million) through 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 underutilize 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.

In sum, 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 frequen-

### Table 3

<table>
<thead>
<tr>
<th>Frequency class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of responses</td>
<td>1017</td>
<td>1014</td>
<td>1028</td>
<td>986</td>
<td>1030</td>
<td>1014</td>
<td>1026</td>
<td>1049</td>
<td>1021</td>
<td>1013</td>
</tr>
</tbody>
</table>
that we tested: 0.75 remained similarly strong for all pairs of sample sizes and corpora.

Completed graduate degree 980
Some graduate school 840
Bachelor’s degree 5223
Associate degree 838
Some college/no degree 1797
High school graduate 500
Some high school 0
No high school 20

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, \( p = 0.82, p = .006 \). (The correlation remained similarly strong for all pairs of sample sizes and corpora that we tested: 0.75 < \( p < 0.85 \), all \( ps < .01 \)). 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 \( \times \) 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.

We identified 500 words from our Study 1 sample—which were taken from the SUBTLEX corpus—that also occurred in four additional (sub)corpora: CELEX, BNC, COCA, and COCA-Spoken. These corpora were chosen to vary in size, language variety, and coverage of genres, in order to serve as a comprehensive test bed 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. Because 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 U.S. film and media subtitles; Panel A), the BNC (100 million word corpus of British English with spoken and written subcorpora; Panel B), CELEX (18 million word corpus of British English based on written sources; Panel C); COCA (400 million word written and spoken corpus; Panel D); and the spoken component of COCA (87 million words; Panel E).

### Table 4

Distribution of Educational Levels by the Number of Responses in the Subjective Frequency Norming Study

<table>
<thead>
<tr>
<th>Educational level</th>
<th>Number of responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>No high school</td>
<td>20</td>
</tr>
<tr>
<td>Some high school</td>
<td>0</td>
</tr>
<tr>
<td>High school graduate</td>
<td>0</td>
</tr>
<tr>
<td>Some college/no degree</td>
<td>500</td>
</tr>
<tr>
<td>Associate degree</td>
<td>838</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>5223</td>
</tr>
<tr>
<td>Some graduate school</td>
<td>840</td>
</tr>
<tr>
<td>Completed graduate degree</td>
<td>980</td>
</tr>
</tbody>
</table>

### Table 5

Values Are Reported Per Frequency Class for the Mean SUBTLEX Corpus Frequency (Per Million in Parentheses), Median SUBTLEX Frequency, and—Separately For LoEd and HiEd Cohorts—Mean Subjective Frequency Ratings (Standard Deviations in Parentheses), and Numbers of Observations

<table>
<thead>
<tr>
<th>FreqClass</th>
<th>MeanFreq (per million)</th>
<th>MedianFreq</th>
<th>MeanSubjFreq, LoEd</th>
<th>MeanSubjFreq, HiEd</th>
<th>N (LoEd)</th>
<th>N (HiEd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 (0.02)</td>
<td>1</td>
<td>1.86 (1.11)</td>
<td>2.68 (1.44)</td>
<td>312</td>
<td>705</td>
</tr>
<tr>
<td>2</td>
<td>2 (0.04)</td>
<td>2</td>
<td>1.91 (1.14)</td>
<td>2.60 (1.50)</td>
<td>328</td>
<td>686</td>
</tr>
<tr>
<td>3</td>
<td>3.46 (0.07)</td>
<td>3</td>
<td>1.92 (1.15)</td>
<td>2.74 (1.43)</td>
<td>309</td>
<td>719</td>
</tr>
<tr>
<td>4</td>
<td>5.95 (0.12)</td>
<td>6</td>
<td>2.33 (1.40)</td>
<td>3.17 (1.49)</td>
<td>312</td>
<td>674</td>
</tr>
<tr>
<td>5</td>
<td>9.84 (0.20)</td>
<td>10</td>
<td>2.29 (1.28)</td>
<td>3.15 (1.41)</td>
<td>323</td>
<td>707</td>
</tr>
<tr>
<td>6</td>
<td>18.21 (0.37)</td>
<td>18</td>
<td>2.57 (1.44)</td>
<td>3.13 (1.28)</td>
<td>320</td>
<td>694</td>
</tr>
<tr>
<td>7</td>
<td>34.45 (0.70)</td>
<td>34</td>
<td>2.72 (1.35)</td>
<td>3.38 (1.26)</td>
<td>313</td>
<td>713</td>
</tr>
<tr>
<td>8</td>
<td>70.65 (1.44)</td>
<td>68</td>
<td>3.22 (1.45)</td>
<td>3.83 (1.45)</td>
<td>312</td>
<td>737</td>
</tr>
<tr>
<td>9</td>
<td>222.43 (4.54)</td>
<td>207</td>
<td>3.64 (1.49)</td>
<td>4.17 (1.48)</td>
<td>316</td>
<td>708</td>
</tr>
<tr>
<td>10</td>
<td>3335.11 (680.64)</td>
<td>1795</td>
<td>5.39 (1.51)</td>
<td>5.56 (1.37)</td>
<td>310</td>
<td>703</td>
</tr>
</tbody>
</table>

**Note.** Corpus size of SUBTLEX is 50 million tokens.
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, whereas the LoEd cohort showed consistently and increasingly lower subjective frequency ratings in words with lower corpus frequencies. The reliability of this Frequency × Skill interaction was confirmed in statistical regression models fitted to subjective frequency as a dependent variable for each of the five corpora (all in statistical regression models fitted to subjective frequency as quency ratings in words with lower corpus frequencies. The showed consistently and increasingly lower subjective fre-

tually identical across corpora. Words with the highest corpus other.

there is no qualitative advantage in using one corpus over any

approximating those strengths via corpus-based frequencies—
of varying educational level. We conclude that—for the purpose of

considered corpora offered frequency counts that were differen-
tially in isolation by 71 participants. Participants (43 females;
2.6) and were not college-bound, that is, 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 81 sentences served as fillers in three eye-tracking experiments, all admin-

terased 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

in the eye-tracking experiment, the participants also undertook

an eye-tracking experiment based on objective frequencies of the 500 target words. All patterns reported here were confirmed in the models that incorporated predicted, rather than observed, subjective frequen-
cies. As Figure 3 illustrates, the predicted values faithfully replic-
cated the functional relation between the alternative measures of

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 fre-
cuency. 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 said
ently in isolation by 71 participants. Participants (43 females;
28 males) belonged to the age group of 16 to 24 years (M = 20.8, SD = 2.6) and were not college-bound, that is, 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 81 sentences served as fillers in three eye-tracking experiments, all admin-
istered 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 indi-

vidual 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, 2012) with

a comparable cohort of participants in which the Woodcock–

Johnson word identification task correlates with the Peabody


especially when individual differences in reading experience are

the subject of interest. We furthermore predict that (a) subjective

frequency will explain more variance in reading behavior than
corpus frequency, and (b) the interaction of skill and word fre-

quency would be attenuated if experience-specific subjective fre-
cency 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 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 sub-

jective frequency estimates in Study 2. Predictions were derived

using the statistical model in Table 6 to estimate subjective fre-
quencies for target words the model had not seen, based on a

nonlinear function of log objective frequency (from SUBTLEX) and

participant education level; the 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.

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 bachelor’s degree or higher (dot-dash line). Dashed lines are locally weight smoother (lowess) trend lines for respective cohorts. Error bars represent one standard error unit.

Study 4: Direct Comparisons of Subjective Versus Objective Frequency

We proposed that subjective frequency ratings are more accu-

rate estimators of the strength of the word’s representation in the

individual mental lexicon than objective corpus-based frequencies,
Picture Vocabulary Test at $r = .75$, and at $r = .54$ after partiaing out IQ.

The Woodcock–Johnson assessment followed standard administration procedures: Participants read aloud a list of individual words, divided into sets of 6 to 8 words, with 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, in order to establish their baseline, participants begin in the middle of the list and move backward if errors occur. Participants then advance through the list until they make six errors in a row. Difficulty ranges from words like “achieved,” “tremendous,” and “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 to 8 min.
Scores in the task ranged from 500 (grade equivalent of fifth 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 subpopulation was thus on par with the word identification skills expected for individuals with an associate degree (typically awarded after 2 years of college-level education) or higher. Based on this, we matched the HiEd cohort in the norming Study 2 (bachelor’s 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 through associate’s 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 codeterminers of word recognition. We discuss the implications of our matching procedure in the General Discussion.

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 data set 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 ($\text{PredSubjFreq}$) and the other with a nonlinear function of corpus frequency ($\text{CorpFreq}$), to each of the following dependent measures: single fixation duration, gaze duration, regression path time, and total fixation time. Nonlinearities in critical predictors were modeled using restricted cubic splines with three 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 five knots, the maximum number of knots that showed consistently reliable effects. The resulting models revealed strong nonlinear effects of both subjective and objective frequencies, and of word length (all $p < 0.001$; models not shown). In single fixation duration, the advantage of predicted subjective frequency over corpus frequency in

### Table 6

Linear Regression Model: Subjective Frequency Ratings are Predicted by the Interaction of Education Level by Log SUBTLEX Frequency ($\log_{\text{CorpFreq}}$)

| Estimate | Standard 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 |

Note. The nonlinear effect of log frequency was modeled as restricted cubic splines with three knots. The model reports the $n$th restricted component of predictor X as rcs(X). HiEd value of the factor education is the reference level. Adjusted $R^2 = 0.51$.

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).
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), whereas 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. 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 (cf. Baayen, 2008).

The Interaction of Word Frequency and Skill

We noted in the introduction of this article that an oft-reported interaction of word frequency and skill has had a significant impact on theorizing about the mechanisms of word recognition in skilled versus 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 nonparametric 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 × 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 × Skill interaction in our eye-tracking data and further report nonstandardized eye-movement latencies.

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 nonlinear function of subjective frequency, and the other an interaction of that same factor with the nonlinear function of corpus frequency. Nonlinearities were modeled with the help of restricted cubic splines with three knots. Both models had a restricted cubic splines function of word length (five knots) as an additional predictor.

Statistical models fitted to log gaze duration confirmed our hypothesis. The model with predicted subjective frequency (labeled as PredictSubFreq) as a critical predictor showed that it did not enter into a reliable interaction with experience level (p = .96; Table 7). The interaction did not reach statistical significance in the model with predicted subjective frequency as a linear predictor (p > .5; model not shown) either. At the same time, a reliable interaction was observed between log SUBTLEX frequency (labeled as CorpFreq) 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 model with subjective versus corpus-based frequency presented as Tables 7 and 8 comes out clearly in Figure 4. We plotted lowest 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 reduction (from 400 ms to 200 ms) in the model-estimated gaze duration between the least and most frequent words in the LoEd cohort, and a 115-ms reduction (from 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 × 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 and experience (ICorpFreq × LoEd: $\hat{\beta} = -0.039; SE = 0.009; p < .0001$), and a nonsignificant interaction of subjective frequency and experience (IPredictSubFreq × LoEd: $\hat{\beta} = -0.013; SE = 0.009; p = .125$). Nonlinearity of critical predictors was not supported by any model. Models fitted to regression path time revealed a marginally significant interaction of log corpus frequency and experience ($rcs1 \times LoEd: \hat{\beta} = -0.023; SE = 0.009; p = .008$; $rcs2 \times LoEd: \hat{\beta} = 0.023; SE = 0.012; p = .0520$), whereas the interaction of subjective frequency and experience was nowhere near the 0.05 threshold of statistical significance ($rcs1 \times LoEd: \hat{\beta} = -0.023; SE = 0.009; p = .008$; $rcs2 \times LoEd: \hat{\beta} = 0.005; SE = 0.011; p = .963$). In models fitted to total fixation time, the interaction of log corpus frequency and experience was significant ($rcs1 \times LoEd: \hat{\beta} = -0.046; SE = 0.007; p < .0001$; $rcs2 \times LoEd: \hat{\beta} = 0.064; SE = 0.010; p < .0001$), whereas the interaction of subjective frequency and experience was not ($rcs1 \times LoEd: \hat{\beta} = 0.050; SE = 0.011; p < .0001$; $rcs2 \times LoEd: \hat{\beta} = 0.016; SE = 0.012; p = .210$).
LoEd: $\hat{\beta} = -0.024; SE = 0.014; p = .098)$. In sum, subjective frequency did not enter into reliable interactions with experience, whereas log corpus frequency reliably interacted with experience throughout the eye-movement record. Taken together, these patterns suggest that the Frequency $\times$ 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, Lacey, Rastle, & Brysbaert, 2012) 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 the high school level to the PhD 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 in Study 2, 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 fitted 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, nonlinearities were modeled with restricted cubic splines with three knots. Once again, experience-matched subjective frequency showed an advantage in the amount of explained variance of about 0.57% compared with objective frequency (7.45 vs. 6.88% of explained variance; model not shown). The advantage was replicated in linear mixed effects models with subjective or

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)**

|                  | Estimate | Standard 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(LogLength)1  | 0.3114   | 0.0056         | 55.23   | 0.0000   |
| rcs(LogLength)2  | -3.2231  | 0.1112         | -28.98  | 0.0000   |
| rcs(LogLength)3  | 0.2688   | 0.0613         | 23.67   | 0.0000   |
| rcs(LogLength)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   |

Note. The critical interaction is presented in bold. The nonlinear effect of frequency was modeled as restricted cubic splines with three knots; the effect of word length was modeled as restricted cubic splines with five knots. The model reports the nth restricted components of Predictor X as rcs(X)_n.

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**

|                  | Estimate | Standard 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(LogLength)1  | 0.3133   | 0.0056         | 55.87   | 0.0000   |
| rcs(LogLength)2  | -3.1997  | 0.1108         | -28.88  | 0.0000   |
| rcs(LogLength)3  | 0.2506   | 0.0268         | 23.30   | 0.0000   |
| rcs(LogLength)4  | -2.9090  | 0.0181         | -16.07  | 0.0000   |
| LoEd:rcs(CorpFreq)1 | -0.0551 | 0.0089         | -9.31   | 0.0000   |
| LoEd:rcs(CorpFreq)2 | 0.0233  | 0.0078         | 2.99    | 0.0028   |

Note. The critical interaction is presented in bold. The nonlinear effect of frequency was modeled as restricted cubic splines with three knots; the effect of word length was modeled as restricted cubic splines with five knots. The model reports the nth restricted components of Predictor X as rcs(X)_n.
English Lexical Project (Balota et al., 2004). Nonlinearities were modeled with restricted cubic splines with three knots. The model with (a) and (b) as predictors explained 38% of variance, and 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 codeterminers of reading behavior, especially in investigations of less skilled reading.

Word frequency and 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 nonlinear word frequency by cohort is significant ($p < .01$, model not shown); the nonlinearity is modeled with restricted cubic splines with three 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, that is, 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 × 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 ($p = .048$; 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, Siakaluk, Chow, and Buchanan (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 because 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, 2012). 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 × 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, whereas 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 $r = 0.82$, $p = .006$. 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).

An additional 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; unfortunately, 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 (a) subjective measures explain less variance than objective ones, and especially so when good objective measures are selected (see a discussion of what constitutes “good” measures in Zevin & Seidenberg, 2002, and Brysbaert & Cortese, 2011); (b) experience-driven individual differences in subjective frequency are in fact differences in some other linguistic characteristics of words, and (c) 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 (a) is incorrect; Study 4 demonstrates that subjective frequency measures actually accounted for more variance across tasks. In what follows, we discuss points (b) and (c).

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-Gonzalez & Davis, 2006). This possibility is supported by the fact that subjective frequency correlates with a number of lexical variables, including AoA and imageability (Baayen et al., 2006; Brysbaert & Cortese, 2011). 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-Gonzalez, and Brysbaert (2012) 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 (c), Baayen et al. (2006) reported correlations between subjective frequency ratings and a number of objective lexical measures. Besides 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 Baayen et al., 2006, Figure 4). 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” (p. 305). 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 discussed previously) 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 & 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 following 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 data set 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 had 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 ($p < .05$) in both models. Moreover the polarity of regression coefficients 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 $r = .97$, $p < .0001$. We conclude that subjective frequency ratings are no more contaminated by correlations with other lexical properties than objective frequencies are. In fact, with a single exception of inflectional entropy, the correlational 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. Although we did not apply the same procedure to the data in Thompson and Desrochers (2009), Table 1 in their article 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 imageable 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 × Skill interaction, or even the variability of subject-specific adjustments to the word frequency effect (Diependaele et al., 2012). 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 × 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 skill that arguably comes closest to estimating the individual’s experience with words. This implies that neither experience, nor changes in vocabulary size and other dimensions of reading ability that experience gains via experience do not undermine the impact of the word’s representation on its recognition speed either; basic 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 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. Although 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 versus
poor readers cannot be explained away by any subjective measure of length.

We address this criticism by directly investigating word length effects in our data, and show that Word Length × 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). 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 to 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 nonparametric 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, and 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; nonlinearities were modeled with the restricted cubic splines with three 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 et al., 2012) 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

---

4 A need for a subjective measure of word length is not inconceivable. A recent study of Rayner, Slattery, and Belanger (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.
subject, removing the base-rate effect (see right panel of Figure 7). Statistical models confirm the statistical significance of the Word Length × 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? We argue that 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 occurrence and behavioral correlates of word recognition: 10 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 and 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 of 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 article is a step in this direction.

**References**


Received February 17, 2012
Revision received July 18, 2012
Accepted July 18, 2012