Behavior Research Methods
2009, 41 (2), 452-471
doi:10.3758/BRM.41.2.452

Corroborating biased indicators: Global and local agreement among objective and subjective estimates of printed word frequency

Glenn L. Thompson and Alain Desrochers
University of Ottawa, Ottawa, Ontario, Canada

The internal validity of several types of experiments in experimental psychology and neuroscience depends in part on the possibility of controlling or manipulating critical lexical variables such as word frequency of occurrence. Two ways of estimating this variable are (1) objective frequency counts and (2) subjective ratings of word frequency. Each method produces estimates that generally agree (i.e., they are highly correlated) but that disagree substantially concerning the relative frequency of a number of words. To investigate this issue more closely, the global and local agreement of subjective frequency estimates was examined in detail for a pool of 6,202 words drawn from the OMNILEX database of French words (Desrochers, 2006; www.omnilex.uottawa .ca). The results indicated that objective and subjective frequencies are strongly correlated, subjective frequencies share a significant amount of *bias* variance with other lexical characteristics (e.g., imageability), and the codeterminants of subjective frequency are in an antagonistic relationship with one another. The implications of these results for the selection of lexical stimuli are discussed, and multiple variables to aid in item selection are reported. Supplemental materials for this study may be downloaded from brm.psychonomic-journals.org/content/supplemental.

Perhaps the single most important determinant of performance in tasks that involve word encoding (e.g., for reading, recognition, or recall) is word frequency of occurrence (for reviews, see Ellis, 2002; Monsell, 1991; Norris, 2006). Words that occur frequently in written text, which are therefore likely to have been encountered often and recently by the average reader, are identified (Balota & Chumbley, 1984; Scarborough, Cortese, & Scarborough, 1977; Schilling, Rayner, & Chumbley, 1998) and read aloud (Balota & Chumbley, 1984; Monsell, Doyle, & Haggard, 1989; Schilling et al., 1998) more accurately and rapidly than are words that are relatively infrequent. Furthermore, judgments concerning their grammatical (Desrochers, Paivio, & Desrochers, 1989; Taft & Meunier, 1998) and semantic (Monsell et al., 1989) properties benefit similarly from higher word frequency. Lexical frequency is also a reliable predictor of recall and recognition performance using classic studied-list memory paradigms (for a review, see Nelson & McEvoy, 2000). In many instances, frequency as a predictor of performance on psycholinguistics tasks is known to interact antagonistically with other lexical attributes, such as orthographic regularity and imageability (de Groot, 1989; James, 1975; Lupker, Brown, & Colombo, 1997; Strain & Herdman, 1999), and even with participant attributes such as educational level (Tainturier, Tremblay, & Lecours, 1992) and exposure to print (Sears, Siakaluk, Chow, & Buchanan, 2008).

In addition to being among a number of important theoretical variables in psycholinguistics (for reviews in which a wide range of other important word variables are considered, see Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), lexical frequency has also proven to be a popular methodological consideration, since virtually all published research includes it either as a general descriptor of item lists or, more commonly, as a control variable. Why is it necessary to control for variables such as frequency? Linguistic stimuli are by their very nature multidimensional, which makes most published psycholinguistic research quasiexperimental (e.g., Coltheart et al., 2001, pp. 250–251, list several important theoretical effects that are simulated by their DRC model, all of which are based on quasiexperimental *between-item* variables). In sum, a limitation of most psycholinguistics research is that the effect of an item attribute on performance may actually be due to an uncontrolled item characteristic.[1]

Unfortunately, apart from experiments in which items can act as their own controls (see note 1), it is impossible to control for all item attributes, known and unknown. Nevertheless, it is desirable that such experiments meet certain minimum necessary conditions if their internal validity is to be provisionally accepted. Chief among these is that lexical characteristics that are known to influence performance, especially important determinants of performance such as frequency, should be adequately controlled. To the

G. L. Thompson, glennlthompson@gmail.com

extent that estimates of a lexical characteristic such as frequency are (1) available and (2) valid for the pool of words under study, it is possible to meet this minimum necessary condition through either experimental (e.g., stimulus list matching) or statistical (e.g., employing a variable as a covariate) means. Although estimates of lexical frequency are available for a large number of words and these estimates are commonly employed by psycholinguistics researchers in implementing experimental controls, we argue that the validity of these estimates is commonly overestimated and, therefore, so is the validity of whatever experimental manipulation is intended.

In general terms, a measure can be said to be valid if it reflects what it was intended to measure. In their critique of the correlation-based validation framework that currently prevails in psychology, Borsboom, Mellenbergh, and van Heerden (2004) evinced that the fundamental criteria for evaluating validity are (1) that the thing being measured must exist and (2) that the thing being measured must cause the value taken by the measurement. Our goal in this article is to address the validity of word frequency estimates in this inclusive sense of the term, as well as several related issues. First, we will consider what estimates of written frequency are intended to measure within the context of psychological research. Second, we will assess the degree to which objective and participant-generated estimates of lexical frequency are valid. Third, we will report the results of an empirical analysis of the global (i.e., across all words) and the local (i.e., for individual words) validity of extant frequency estimates for printed French words. Variables derived from this analysis are reported with the aim of guarding researchers against the use of words with frequency estimates that are likely to be invalid. Finally, we will discuss the implications of these results for the selection of lexical stimuli for research applications and hypothesis testing.

What Does Objective Frequency Estimate?

Objective frequency counts are obtained by recording the number of times that a word occurs in a sample of text, or *corpus*. The obtained value, expressed as a proportion of the total number of instances, or *tokens*, in the sample, can be conceived of as the probability of encountering a given word at any point within a corpus (for a discussion of this concept and its limitations, see Baayen, 2001). Typically, the corpus itself is not the true object of study, but rather the population from which it was sampled. Researchers are interested in statistics computed from the sampled corpus only insofar as these constitute estimates of parameters for a corresponding population of text. This statement, however, begs the question of what population is being described.

One possibility is that frequency counts obtained from samples are intended to estimate the objective frequency (e.g., frequency per million) that would be observed in the entire population of text (i.e., all printed work in a given language).[2] In this case, the accuracy of objective frequency counts depends on how the corpus was sampled from the population. In principle, all types of written text (e.g., literature, newspapers) should be represented in proportions that are equal to those found in the target population. The fact that major sources of objective frequency counts are based on samples that implicitly exclude various common types of texts (e.g., those presented on computer interfaces, newspapers, restaurant menus, street signs) points to a potential bias in the estimated parameters (for examples of typical objective frequency studies, see Baayen, Piepenbrock, & Gulikers, 1995; Imbs, 1971; Kučera & Francis, 1967; Leech, Rayson, & Wilson, 2001; New & Pallier, 2005; New, Pallier, Brysbaert, & Ferrand, 2004). Similarly, the population(s) of text to which the results should generalize is seldom stated, which makes it difficult to determine whether the sampling has been performed satisfactorily.

What impact could sampling too selectively from the target population of text have? If all words are evenly distributed throughout the population (i.e., *full dispersion* is present for all words; see Baayen, 2001, pp. 164–170), the impact will be minimal, since unrepresented segments of the population do not differ from the population as a whole. However, many relatively low-frequency content words tend to be *underdispersed*, which means that the probability of encountering them is heavily dependent on context and, therefore, factors such as corpus sample size and sampling strategy will affect the stability of parameter estimates (Baayen, 2001; for emerging evidence that degree of context dependency should be treated as an important variable in its own right, see Adelman, Brown, & Quesada, 2006). Indeed, empirical cross-validation of objective frequency counts with other lexical variables (i.e., subjective frequency, imageability) are consistent with the idea that the fidelity of objective frequency counts breaks down with rare items (as has been shown by Desrochers & Thompson, 2009; Gonthier, Desrochers, Thompson, & Landry, in press).

The challenge that underdispersion presents to the validity of objective frequency estimates, especially those for low-frequency content words, is twofold: (1) It causes substantial sample-to-sample variability in frequency estimates, regardless of sampling strategy (partially mitigated by large sample size; Burgess & Livesay, 1998; or by relying on multiple sources), and (2) it implies that estimates for many words that have been derived from a highly content-dependent source (e.g., newspapers) are unlikely to match those derived from other contexts or a more heterogeneous corpus (i.e., some items are likely to be specific to a *register*). The importance of these two problems is uneven. The first is essentially an exaggerated form of sampling error that causes estimates for some words (i.e., those that are underdispersed) to fluctuate more than others across corpora. This means that researchers who rely on a single corpus may be working unwittingly with unreliable frequency estimates for such words. In contrast, the second problem concerns a general validity issue arising from the decision to restrict the population of text from which a corpus can draw, which is a flaw that cannot be corrected by relying on multiple sources of objective frequency if these share the same sampling strategy. In this case, the result is a systematic underestimation or overestimation of words whose frequency of use varies across the two populations: that from which samples of text were taken and that from which they were not.

The problems that can arise with objective frequency estimates are magnified by the fact that the target population of interest to most experimental psychologists is not really the entire population of written text. Rather, within the context of psychological research, objective frequency counts are intended to capture a word's familiarity to the average reader (Nelson & McEvoy, 2000).[3] Globally, the problem of obtaining a corpus that is representative of normal experience is one of first determining how the average reader samples the entire population of written text and then mirroring the observed contribution of various types of content in assembling the corpus. Regrettably, it is not typical for such considerations to inform the sampling procedures of corpus studies, most of which were not designed solely with the needs of psycholinguistics researchers in mind.[4] As a consequence, frequency estimates for words whose use is largely context dependent (i.e., low-frequency content words) are especially vulnerable to validity problems. This source of bias is exacerbated by the fact that although the amount of text attributable to oft-ignored sources (e.g., restaurant menus, labels, street signs) may be trivial when compared with that produced for books and newspapers, it is nevertheless likely to represent an important part of the total exposure to written text of the average person, even an educated one. The importance of such encounters is evident when one considers that much of this material is stereotyped and read/encountered repeatedly (e.g., words found on a Web browser menu). In sum, despite the global reliability of frequency count data, there are reasons to expect the "objective" frequency of many words to be unreliable, invalid, or both.

Objective and Subjective Frequency: Links in a Causal Chain

According to the criteria proposed by Borsboom et al. (2004), objective frequency counts are valid for psychological research to the extent that (1) familiarity exists as a psychological state and, if we reverse the direction of the causal relationship they proposed, (2) changes in a word's objective frequency cause changes in its familiarity via shifts in the probability of its being encountered during reading. That readers encounter certain words more often than others is generally accepted. It is also clear that the frequency with which words occur in the population of text will cause readers to encounter them more often and more recently (for an analysis of temporal changes in objective frequency and its relationship to the familiarity ratings provided by old and young readers, see Balota, Pilotti, & Cortese, 2001). However, as was noted above, the causal relationship between objective frequency estimates and familiarity depends on how adequately a given corpus represents the average reader's experience. What is more, the underdispersion problem noted above implies that the validity of many estimates within an otherwise reliable set of norms is questionable, and relying on multiple sources of objective frequency is not a foolproof method for resolving the issue, given the systematic bias in sampling methods noted earlier.

Indeed, the problems associated with objective frequency estimates have led some researchers to seek alternative ways of estimating familiarity (e.g., Gernsbacher, 1984; Gilhooly & Logie, 1980). For example, Gernsbacher remarked on some problematic cases that are symptomatic of underdispersion, noting that words of vastly different familiarity were sometimes assigned identical objective frequency values (e.g., both *boxer* and *loire* are assigned a frequency count of 1 in Kučera & Francis, 1967). Instead of objective frequency, she proposed that the experiential familiarity or, perhaps more precisely (Balota et al., 2001), the subjective frequency of a word would provide a better estimate of familiarity. This way of estimating familiarity involves asking a group of participants to rate the subjective frequency of written words along a numeric (Likert-like) scale whereby high values denote high subjective frequency and low values denote low subjective frequency. The average rating that is obtained from the sample for each word is supposed to represent its relative familiarity to the average reader in the population.

Ostensibly, the subjective frequency judgments provided by the participants take on numerical values that are determined, at least in part, by the actual familiarity of the words (i.e., the experience of readers). According to the causality criterion adopted above, a word's subjective frequency rating is valid to the extent that it truly reflects the rater's experiential familiarity. Although it is not obvious that readers would have detailed access to their familiarity experience, the generally high correlation between objective and subjective frequency (Balota et al., 2001; Benjafield & Muckenheim, 1989; Desrochers & Thompson, 2009; Gonthier et al., in press; Stadthagen-Gonzalez & Davis, 2006) is consistent with the idea that objective frequency causes a psychological state corresponding to familiarity, which subjective frequency ratings are intended to capture. However, as with objective frequency, the causal link with familiarity may be disrupted for certain items. Words are multidimensional stimuli, and judgments made by participants on a specific characteristic such as subjective frequency may be systematically influenced by unrelated factors, such as semantics or orthographic redundancy (for demonstrations with English words, see Baayen, Feldman, & Schreuder, 2006; Balota et al., 2001). Thus, the various dimensions that characterize a word can be considered potential codeterminants of the value taken by subjective frequency judgments. To the extent that these codeterminants disrupt the causal link between familiarity and the value taken by subjective frequency judgments, the validity of subjective frequency estimates can be called into question (although Stadthagen-Gonzalez & Davis, 2006, presented evidence suggesting that this additional *bias* variance is not predictive of naming performance and is, therefore, of little practical importance).

Assessing the Validity of Word Frequency Estimates

So far, we have examined the issue of whether estimates of word frequency, both objective and subjective, can be considered valid. Globally, significant and positive correlations between objective and subjective frequency provide some necessary, although not necessarily sufficient or complete (correlations are imperfect), empirical support for validity within the framework sketched earlier. Locally, there

are ways in which the validity of familiarity estimates can fail, resulting in a breakdown of the link between familiarity and its indicators. For objective frequency, failures of validity may arise because of mismatch between the content of the text sample, or corpus, that is used to compute frequency estimates and the experience of the average reader. For subjective frequency, failures of validity may arise because participants implicitly rely on information other than their experiential familiarity when assigning ratings to words.

The importance of the validity issue raised here cannot be overemphasized. Although it is true that objective and subjective frequency estimates generally agree (Gonthier et al., in press), it is also true that the correlation between the two is not perfect and that the disagreement between them may be very large for some words. Even though it is tempting to minimize the importance of such residual variance, such impulses should be resisted. As methodological concepts, reliability and validity are attributes of a *sample* of scores, not of a test or of the population from which the sample was drawn (B. Thompson & Vacha-Haase, 2000). This consideration is important because researchers in experimental psychology and psycholinguistics seldom work with the entire population of words for which frequency ratings are available. Instead, they tend to use highly controlled and relatively short lists of 10–20 words (for a critique of this practice, see Balota et al., 2004). The small samples selected by researchers may easily comprise mostly words for which the familiarity estimate is invalid, notwithstanding the acceptable "validity" observed within the entire population of available words and their associated ratings.

That said, it may be argued, not without cause, that an explicit manipulation of frequency is unlikely to fail despite many local failures of validity, because it is such a powerful variable. We would counter that explicit manipulations of frequency are not as common as experimental control of frequency, whereby conditions are equated on this variable to isolate another effect that is usually much weaker. For example, a fair comparison of low-frequency orthographically regular words and low-frequency orthographically irregular words requires that their respective frequency of use be equivalent (see, e.g., Content, 1991; Hino & Lupker, 2000). Given the problems with objective frequency estimates, especially for low-frequency words, how much confidence can a researcher have that these two lists are actually equivalent? Even if frequency and regularity were uncorrelated in the lexicon (i.e., there is nothing systematic causing a confound), the items selected may be unequal in frequency by random chance (e.g., 20-item lists are hardly exempt from the hazards of sampling bias) or on account of the unconscious bias of the researcher performing the selection (Forster, 2000).[5] At the extreme, main effects or interactions that are due to undetected between-list differences in familiarity may be falsely attributed to other variables, causing researchers to waste their time and resources trying to replicate findings with different items or, perhaps more problematic, replicating the same effect with the same items and a different sample of participants and, again, misattributing its cause to another variable. Establishing the prevalence of such errors is beyond the scope of this article, but it is clear that the only way to reliably safeguard against them is to work exclusively with accurate frequency estimates, eschewing intuitive item-by-item judgments on this point for decisions based on objective descriptors.

Even experiments that rely on the regression analysis of a large pool of words (e.g., Balota et al., 2004) are not invulnerable to the sources of validity failure noted above. Subjective frequency, for example, may lead investigators to overestimate the association between the psychological construct familiarity and other lexical attributes, which, in practical terms, means that the effects of other variables in a regression equation may be masked as *shared variance* (Balota, Ferraro, & Connor, 1991). Furthermore, the problems with objective frequency estimates are likely to be more prevalent in words of relatively low frequency, which may pose problems in the detection and interpretation of interactions with frequency. For this reason, it is important to study the influence of nuisance factors in the production of subjective frequency ratings and to flag objective frequency estimates that are likely to be invalid.

Interestingly, the joint analysis of objective and subjective estimates of familiarity, each flawed in its own way, offers opportunities for identifying and explaining their respective failures of validity at both the global and local levels. In the case of subjective frequency, it is possible to estimate the amount of systematic bias caused by word properties unrelated to the construct *familiarity* by temporarily adopting objective frequency as a gold standard (as Balota et al., 2001, did). Here, the term *bias* denotes variance that subjective frequency shares with other lexical attributes but that is unshared with objective frequency. Such variance may be considered a potential source of bias away from the true score for relative familiarity that both objective and subjective frequency are intended to estimate. Viewed in this light, bias variance is disruptive to the validity of subjective frequency scores. Of course, the qualification *potential source of bias* must be emphasized, because bias has been defined operationally with reference to an imperfect gold standard (i.e., objective frequency). In any case, if we accept that bias variance is potentially disruptive, it is possible to take advantage of the fact that it is explainable in terms of the codeterminants of subjective frequency, which allows one to do more than simply evaluate the extent of bias; it can be quantified and then removed to produce an adjusted subjective frequency estimate with improved validity, assuming that the logic developed here holds. As a bonus, this adjusted subjective frequency estimate would not unduly mask the effect of important covariates (e.g., imageability) within the context of a regression analysis (Baayen et al., 2006).

The comparison of objective and subjective frequency also allows the computation of an index that quantifies the extent to which they agree for each word for which we have complete data. If both estimates of familiarity agree, the relative familiarity estimate for that particular word is probably valid for both indicators (for examples of such cross-checking in the literature, see Bertram & Hyönä, 2003; Monsell et al., 1989). In contrast, words for which objective and subjective frequency disagree are likely to suffer from a failure in validity within one or both indicators of

familiarity. An index of disagreement between the indicators would provide an empirical basis for the exclusion of items from stimulus lists and for explaining inconsistent results in terms of failures to properly control for familiarity. Furthermore, such problem words would also be flagged for future scrutiny by researchers interested in determining the source of the discrepancy. Plausible explanations for local disagreements between the indicators could provide reasons for trusting the familiarity estimate provided by one over the other for a particular word (for a related discussion, see New, Brysbaert, Véronis, & Pallier, 2007). The outcome of such an analysis, along with the results of analyses reported elsewhere (e.g., see the scatterplots in Desrochers & Thompson, 2009; Gonthier et al., in press) can serve to verify the claim (e.g., Stadthagen-Gonzalez & Davis, 2006) that subjective frequency is indeed redundant with objective frequency.

Purpose

The purpose of this article is to report the results derived from a detailed analysis of the relationship between objective and subjective frequency estimates for a large pool of French words. These estimates are conceived as links in a causal chain whereby objective frequency causes a psychological state called *familiarity*, which, in turn, although not necessarily by itself, determines the value taken by subjective frequency ratings. At the global level, we evaluate the extent to which spurious causes of subjective frequency disrupt this causal chain, causing bias in the ratings. Again, here the term *bias* is used to mean a failure in validity for subjective frequency, of which partial dissociation of objective and subjective frequency is a symptom and codetermination by other psycholinguistic variables is a cause. This aspect of the analysis replicates and extends that reported for English words by Balota et al. (2001) by using French lexical items, a different semantic variable (imageability), and, importantly, an objective estimate of spoken frequency so that the contribution of this variable could be examined (e.g., Baayen et al., 2006). At the local level, we quantify the disagreement among objective and subjective frequency estimates, both before and after controlling for bias in subjective frequency, in order to identify specific words for which the validity of available ratings is suspect. The latter analysis is similar in logic to that reported by New et al. (2007), examining local discrepancies between competing sources of spoken frequency estimates in French.[6]

METHOD

The present analysis was based on data extracted from the OMNILEX database (Desrochers, 2006). This database is a repository for various kinds of information about French words (e.g., structural, grammatical, distributional, relational, and semantic characteristics). For the present analysis, indicators for objective written frequency (henceforth, *written frequency*), objective spoken frequency (henceforth, *spoken frequency*), subjective written frequency (henceforth, *subjective frequency*), imageability, and orthographic neighborhood (N) were obtained for a total of 6,202 French words. This sample was made possible because OMNILEX contains the combined subjective frequency and imageability data of several different studies, greatly increasing the number of words for which ratings are typically available.

Written frequency is a composite of multiple sources for objective frequency (e.g., *Trésor de la langue française*, Imbs, 1971; *Lexique 2*, New et al., 2004; *Lexique 3*, New & Pallier, 2005), which is expressed as the number of occurrences out of a million. Spoken frequency is based on the *Lexique 3* spoken frequency norms, which reflect a word's frequency of use in transcribed movie and TV show dialogues (New & Pallier, 2005). These frequency count norms were log-transformed prior to analysis, due to positive skewness. Subjective frequency and imageability are based on the combined norms of several Canadian studies in the OMNILEX database that have collected subjective frequency and imageability data on French words (Desrochers & Bergeron, 2000; Desrochers & Thompson, 2009; Forget, 2005; Gonthier et al., in press).[7] Finally, N is an estimate of orthographic neighborhood based on the operational definition proposed by Coltheart, Davelaar, Jonasson, and Besner (1977): the number of words that can be formed by changing one letter of the target word to a different letter, with all shared letters in the same serial position. This variable was log-transformed as well, because of positive skew.

RESULTS

The statistical analyses reported in this section involved four incremental steps. In Step 1, we examined the global discrepancy between objective and subjective frequency and the extent to which the other variables considered here (i.e., imageability, spoken frequency, and N) may account for this discrepancy. The influence of each variable was analyzed one at a time. However, these *bias-causing* variables are correlated with one another to varying degrees, which is why the shared and unique contributions of these variables were evaluated subsequently. In Step 2, accordingly, we evaluated the unique contribution of each variable in disrupting the objective–subjective association. In Step 3, we tested whether the influence of these lexical properties on subjective frequency estimation varied depending on the level of objective frequency. This idea was evaluated using tests of statistical significance for the associated interaction terms in models predicting subjective frequency. Finally, Step 4 presents the results of a detailed analysis of the subjective–objective written frequency relationship, both before and after controlling for other factors. The comparison of the subjective and objective scores was achieved via a regression analysis, the indices of disagreement taking the form of the resulting standardized residuals. For each step of the analyses, the procedure used is described, and the results are reported.

Step 1: Using Mediation Tests to Estimate the Amount of Potential Bias Variance in Subjective Frequency

The potential bias in subjective frequency estimation caused by imageability, spoken frequency, and N can be conceived of as a mediation problem. Typically, mediation tests investigate the plausibility of a model of the causal

relationship between two or more predictors and a dependent variable (subjective frequency in this case) via a structured series of regression equations. The advantage of mediation tests over more conventional regression analyses is that the statistical significance of the variance shared by predictors is explicitly tested (i.e., as a potential mediation effect) within the context of a theoretically guided model of the data. It is important to note that mediation is used here merely as a tool for testing the significance of shared and unique variance. Thus, the mediation tests that follow are not intended as tests of causal models in the same way that such tests are normally employed.

The analyses reported below were conducted in part using the SPSS macros for nonparametric simple mediation tests proposed by Preacher and Hayes (2004). The advantage of this procedure is that, unlike simple Sobel tests of shared predictive variance, bootstrapping is used to avoid the restrictive assumption of normality. The logic of the mediation tests conducted here runs as follows. If written frequency *completely mediates* the association between some extraneous variable (say, imageability) and subjective frequency, the association is reducible to shared variance with written frequency, and therefore, there is no evidence to suggest that this extraneous variable causes bias in subjective frequency estimation. If *partial mediation* is found—which is to say, there is a statistically significant amount of both shared and unique variance—a portion of the overlap between subjective frequency and the potentially biasing variable may be legitimate (i.e., due to a real correlation with print exposure), and part of it may be bias (i.e., due to disruption of the process of estimating subjective frequency). Finally, if *no mediation* is observed, none of the variance shared by imageability and subjective frequency is reliably corroborated by written frequency, which means that there is a distinct lack of empirical evidence attesting to the validity of its interpretation as familiarity-based variance. Variance in subjective frequency that is shared with other lexical attributes may, therefore, be tentatively considered invalid, a source of bias. The bootstrapping procedure employed here consisted of a thousand iterations. Figure 1 schematically represents corroborated variance (region denoted by the letter A) and bias variance (region denoted by the letter B). The means, standard deviations, and zero-order correlations among all the variables involved in this analysis are reported in Table 1.

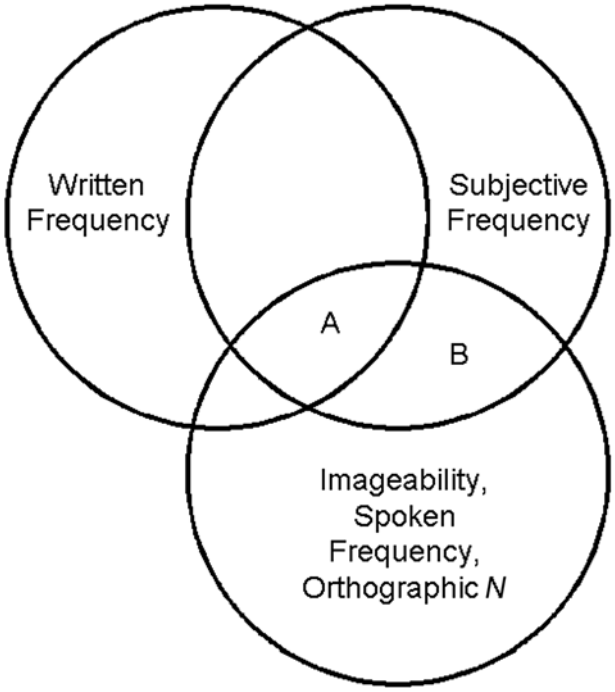

Figure 1. Schematic representation of the variance shared by subjective frequency and its codeterminants, which can be partitioned into two components: corroborated variance (A) and potential bias variance (B). *N* denotes neighborhood.

Imageability. Mediation tests revealed that the zero-order relationship between imageability and subjective frequency was significant [$b = .33$; $t(6201) = 27.63$, $p < .0001$, $r^2 = .11$] and that this effect was partially mediated by written frequency (Sobel test, $b = .06$; $Z = 6.92$, $p < .0001$, semipartial $r^2 = .03$). The direct effect of imageability (i.e., the variance component that is not corroborated by written frequency) was also significant [$b = .28$; $t(6200) = 30.65$, $p < .0001$, semipartial $r^2 = .08$]. Thus, the bulk of the relationship between imageability and subjective frequency is unrelated to written frequency and may, therefore, be due to bias in the process of estimating subjective frequency.

Spoken frequency. The zero-order relationship between spoken frequency and subjective frequency was significant [$b = 1.12$; $t(6201) = 64.15$, $p < .0001$, $r^2 = .41$]. The mediation tests revealed that written frequency

Table 1
Descriptive Statistics and Zero-Order Correlations Among Various Lexical Characteristics: Written Frequency, Spoken Frequency, Subjective Frequency, Imageability, and Orthographic Neighborhood (*N*)

	M	*SD*	Orthographic *N*	Imageability	Subjective Frequency	Spoken Frequency
Log *N*	0.44	0.38	–			
Imageability	4.00	1.10	.06	–		
Subjective Frequency	4.00	1.10	.06	.33	–	
Spoken Frequency	0.80	0.62	.15	.20	.64	–
Written Frequency	1.06	0.63	.17	.09	.65	.78

Note—All correlations are based on 6,202 items and are significant at the .001 level. Written and spoken frequency refer to the log-transformed frequency per million occurrences for each word in the written and oral domains.

partially mediates this relationship (Sobel $b = .57$; $Z = 26.46$, $p < .0001$, semipartial $r^2 = .36$). Thus, the bulk of the association between spoken frequency and subjective frequency is also shared with written frequency (i.e., 36%). However, the direct effect was also statistically significant, albeit less important [$b = .55$; $t(6200) = 20.75$, $p < .0001$, semipartial $r^2 = .04$], suggesting that spoken frequency may bias the process of estimating subjective frequency to some extent. Of note is the fact that the direct effect of imageability was comparatively twice as large.

Orthographic neighborhood. The zero-order relationship between *N* and subjective frequency was statistically significant [$b = .17$; $t(6201) = 4.72$, $p < .0001$, $r^2 = .004$]. The mediation tests revealed that written frequency partially mediated this relationship (Sobel $b = .32$; $Z = 13.41$, $p < .0001$, semipartial $r^2 = .003$). Thus, the bulk of the association between *N* and subjective frequency is also shared with written frequency. However, the direct relationship between *N* and subjective frequency was also statistically significant, but in the opposite direction [$b = -.15$; $t(6200) = -5.51$, $p < .0001$, semipartial $r^2 = .001$]. The magnitude of the effect sizes suggests that the practical importance of *N* is quite small in this context at 0.1%, despite the fact that these tests have achieved statistical significance. This issue was examined further by computing a likelihood ratio comparing the latter direct test with a null model. The resulting value indicated that a model including this parameter is preferable to one without it, as evidenced by the value greater than 1 ($\lambda = 22.25$). However, two alternative adjustments to the likelihood ratios that compensate for model complexity disagreed on the utility of *N*: Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), respectively ($\lambda_{AIC} = 8.17$, $\lambda_{BIC} = 0.28$; for details, see Glover & Dixon, 2004). The former compensates for the complexity of additional model parameters, whereas the second compensates for sample size as well. At this point, we reserve judgment about the utility of *N* and will return to this issue in Step 2.

In summary, the results of the Step 1 analysis support the idea that imageability, spoken frequency, and *N* influence the value taken by subjective frequency ratings. The potential for bias was suggested by the significant zero-order correlations and was further supported by the fact that these associations remained significant after written frequency was controlled for. Of course, the uncorroborated variance that subjective frequency shares with these other variables is actually an overestimate of the potential bias in subjective frequency ratings. The reason for this is that imageability, spoken frequency, and orthographic *N* are correlated with one another, and so the effect sizes for the direct effects reported above are not necessarily additive. To address this ambiguity, we now will turn to the question of whether each of these variables is a unique source of bias in subjective frequency estimation.

Step 2: Assessing the Unique Contribution of Imageability, Spoken Frequency, and *N* to Bias Variance

In the preceding analysis, we examined the influence of each potential confounding variable independently. The purpose of the following analysis was to examine jointly the shared and unique contribution of each variable in explaining variance in subjective frequency. To begin, the variance that the imageability, spoken frequency, and *N* variables shared with written frequency was removed through a series of regressions, using written frequency as a predictor. The resulting unstandardized residuals for each variable effectively represent the variance that is uncorrelated with written frequency (i.e., the potential bias component). These *bias-only* variables were then entered into a regression equation predicting subjective frequency.

The results of the standard regression predicting subjective frequency from the bias-only variables for imageability, spoken frequency, and *N* confirmed the results of the mediation tests reported in Step 1. Overall, the predictors were significantly related to subjective frequency [$F(3,6198) = 226.34$, $MS_e = 1.09$, $p < .001$, $\eta^2 = .10$ (.02 shared)]. Furthermore, each of the predictors made an independent contribution to the equation. Bias-only imageability alone explained 5.8% of the variance in subjective frequency ($t = 20.02$, $SE = 0.012$, $p < .001$). Bias-only spoken frequency added another 1.9% of explained variance ($t = 11.49$, $SE = 0.035$, $p < .001$). Finally, bias-only *N* added 0.04% variance explained to the overall equation over and above what was already explained by the other variables ($t = -5.49$, $SE = 0.035$, $p < .001$). A likelihood ratio was performed on the basis of the performance of this model to test again whether *N* was a useful predictor. The results confirm that a model including the *N* variable is far more likely than one including only imageability and spoken frequency ($\lambda = 938{,}169$, $\lambda_{AIC} = 344{,}687$, $\lambda_{BIC} = 11{,}912$). Thus, it would seem that, on balance, the contribution of *N* is not trivial. Note that in this data set, the bulk of the association between bias-only imageability and subjective frequency is not shared with spoken frequency, which is consistent with the idea that we are dealing with bias variance. Conversely, we can reject an alternative explanation: It is not the case that the tendency of professional writers to overrepresent abstract words in their productions has led to an underestimation of the real association between imageability and frequency (as suggested by Gonthier et al., in press). We will return to this issue in the Discussion section.

Step 3: Does the Potentially Biasing Influence of Other Lexical Properties Interact With Objective Written Frequency?

Balota et al. (2001) found that the association between subjective frequency and both meaningfulness and orthographic *N* varied across levels of objective frequency in English, although they did not explicitly test the interaction—proceeding, instead, straight to post hoc decomposition. In other words, objective frequency was deemed to moderate the relationship between potential biasing variables and subjective frequency. In anticipation of this result, Balota et al. (2001) evoked a connectionist model that posits deceleration in the accrual of activation for a particular pattern as the overall activation in a network nears its asymptote (for evidence that frequency effects

are activation based, see Allen, Smith, Lien, Grabbe, & Murphy, 2005). Such a view predicts that there is more opportunity for variables to exert an effect on judgments involving words when overall network activation is low (e.g., low frequency). This suggests that no matter what variable is examined, its explanatory power will decrease as a function of the total amount of information contributing to the activation of a particular word. The observed pattern of interaction should then be antagonistic rather than synergistic. In fact, they observed the predicted pattern with meaningfulness but the opposite pattern with orthographic *N*, which would seem to provide mixed support for their hypothesis.

The same issue was examined here using French words and imageability, instead of meaningfulness, as the semantic variable. Spoken frequency was included as an additional variable. Because the main effects were examined in detail above, we will report only the results of the interaction tests. The reliability of key moderation tests involving objective frequency was tested and will be reported first. Then all reliable interactions were evaluated within the same analysis in order to examine the contribution of shared variance and to test higher order interactions. The techniques employed for modeling and testing interactions between continuous variables are detailed in Cohen, Cohen, West, and Aiken (2003, chap. 7) and Bauer and Curran (2005).

A moderation test revealed that the bias caused by imageability varies in magnitude depending on the written frequency of a word [$b = -.191$, $SE = .013$, semipartial $r^2 = .014$; $F(1,6198) = 182.33$, $p < .001$]. As was expected, a simple-slope (simple-effect) analysis of this interaction indicated that imageability was a consistent source of bias, regardless of written frequency, but that it became a more important determinant of subjective frequency rating as objective frequency decreased. When written frequency was high (+1 *SD*), the effect of imageability was small but significant [$b = .149$; $t(6198) = 11.58$, $p < .001$]. When written frequency was average, the effect of imageability increased and was still significant [$b = .269$; $t(6198) = 30.32$, $p < .001$]. Finally, when written frequency was low (−1 *SD*), the effect of imageability was largest [$b = .388$; $t(6198) = 31.86$, $p < .001$]. To put these effects into perspective, the standard deviation of subjective frequency is 1.10. Thus, for every one-unit increase in imageability, there is a corresponding .388 (or .35 *SD*) increase in subjective frequency when words are low in written frequency.

Note that a greater range in the simple-slope effects reported above would be observed if more extreme points on the moderator (written frequency) had been used (e.g., ±2 *SD*s). Because the choice of value at which to evaluate simple slopes is arbitrary, we will report a region-of-significance test as an additional description of the observed results (Bauer & Curran, 2005). An advantage of this strategy is that the reader can verify whether the substantive conclusions would have been different had different values on the independent variable been chosen. The upward bias in subjective frequency rating that was caused by imageability, for instance, was significant at the .05 alpha level for all values of written frequency up to 2.54 ($Z = .98$), at which point the effect ceased to be significant. Because conventional interactions are linear, the effect of imageability eventually reversed and became significant again at values of 3.12 ($Z = +1.35$) and above, which actually describes a very low percentage of words in the positively skewed objective frequency distribution (<1%). Taken together, the results indicate that imageability significantly biased subjective frequency ratings upward for all but the most frequent words.

Similarly, written frequency was found to significantly moderate the relationship between spoken frequency and subjective frequency [$b = -.315$, $SE = .022$, semipartial $r^2 = .018$; $F(1,6198) = 213.34$, $p < .001$]. A simple slopes analysis revealed that this effect was significant at all levels of written frequency but increased from high written frequency [$b = .506$; $t(6198) = 19.29$, $p < .001$], to average written frequency [$b = .703$; $t(6198) = 25.03$, $p < .001$], and finally to low written frequency [$b = .900$; $t(6198) = 25.42$, $p < .001$]. As can be seen, the final unstandardized beta coefficient nearly reached a full standard deviation on subjective frequency. The region-of-significance test indicated that the biasing effect of spoken frequency was significant for values on the written frequency moderator of 3.52 ($Z = +1.62$) and below and became a statistically significant effect in the opposite direction at values of 4.16 ($Z = +2.04$) and above. Like imageability, we can conclude that spoken frequency significantly biases subjective frequency rating upward for most of the meaningful range of written frequency, becoming unreliable and reversing at the extreme upper tail of its distribution.

In contrast, a similar moderation test involving orthographic *N* did not detect an interaction with written frequency [$b = -.055$, $SE = .04$, semipartial $r^2 = .0001$; $F(1,6198) = 1.69$, $p = .19$]. This result fails to confirm the interaction between these variables that was assumed to exist by Balota et al. (2001). This difference is attributable to one of two things: (1) The interaction that they interpreted was not significant (they did not test for it), or (2) the relationship between orthographic *N* and subjective frequency differs for French and English.

Evaluating all these interaction effects within the same analysis did not change the observed pattern of results. All previously significant effects remained statistically significant after controlling for the variance shared among the interaction terms. The moderation of the association between imageability and subjective frequency by written frequency was significant, if reduced ($b = -.124$, $SE = .023$, semipartial $r^2 = .002$; $t = -5.47$, $p < .001$). This drastic reduction is attributable to the high correlation between the two interaction terms involving imageability and objective frequency estimates (written, spoken; $r = .79$). This shared variance is interpretable in terms of either written frequency or spoken frequency. However, the unique moderating effects of written frequency (see above) and spoken frequency (see below) are not.

In a similar manner, written frequency remained a significant moderator of the relationship between spoken frequency and subjective frequency after controlling for shared variance ($b = -.187$, $SE = .021$, semipartial $r^2 =$

.005; $t = -8.92$, $p < .001$). The explained variance associated with the interaction between written frequency and spoken frequency (1.8%) was roughly cut in half after controlling for variance shared with other predictors. Weak but significant ($p < .001$) correlations with imageability ($r = -.063$) and the imageability × written frequency interaction ($r = .07$) are largely responsible for creating the shared variance in this analysis that was not present in the moderation tests reported above, thereby reducing the magnitude of the unique effect.

The analysis also revealed a new effect, since imageability and spoken frequency were found to interact independently of the other effects. The sign of the interaction coefficient indicates that as the level of one increased, the effect of the other decreased ($b = -.071$, $SE = .023$, semipartial $r^2 = .0007$; $t = -3.17$, $p = .002$). This effect, however, must be considered trivial, given the high power of the test, its negligible unique contribution to explained variance, and the presence of a three-way interaction.

Entering the three-way interaction into the equation contributed a significant amount of additional variance [$b = .121$, $SE = .019$, semipartial $r^2 = .003$; $F(1,6194) = 39.56$, $p < .001$]. A simple-slope analysis revealed that imageability influenced subjective frequency rating at any combination of spoken frequency and written frequency (high vs. low) but that its effect was largest when words were low in frequency across the board. When words were low in written and spoken frequency, the effect of imageability was greatest [$b = .39$; $t(6194) = 27.17$, $p < .001$]. When words were low in written frequency but high in spoken frequency, the imageability coefficient was cut by half [$b = .131$; $t(6194) = 3.81$, $p < .001$]. In the reverse situation (high written and low spoken), the effect of imageability was similar [$b = .147$; $t(6194) = 4.97$, $p < .001$]. Finally, when both moderators were high, the effect of imageability was weakest [$b = .079$; $t(6194) = 5.89$, $p < .001$].

In summary, with the exception of *N*, which did not interact with any other factor, the results are consistent with the hypothesis advanced by Balota et al. (2001), whereby sources of information interact with each other competitively (i.e., antagonistically) rather than synergistically in determining the value of subjective frequency ratings. Having examined the global correlation-based issues related to the validity of subjective frequency estimates, we now will turn to the issue of local failures in subjective frequency estimation.

Step 4: Detecting Local Discrepancies in Subjective Frequency Estimation

In producing estimates of disagreement, objective written frequency estimates were compared with subjective frequency estimates via a series of ordinary regression models of their relationship. First, the zero-order relationship was examined by predicting written frequency from subjective frequency. This analysis was conducted as a way to quantify the discrepancy between familiarity estimates, using a standard metric before controlling for potential confounds (see Figure 2). Then this relationship was examined again, after having removed the variance in subjective frequency, which we have argued distorts the purity of subjective frequency as a measure of the psychological state familiarity (see Figure 3). The sources of bias variance presented above, including interactions, were used to adjust subjective frequency values.[8] Finally, a rescaled subjective frequency variable that represents a relatively unbiased estimate of subjective frequency rating was generated on the basis of residuals left over after subjective frequency was predicted from the known bias-causing variables considered here.

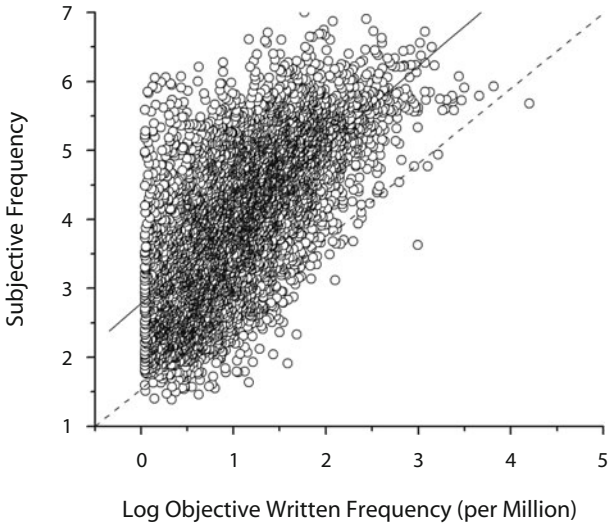

Figure 2. Scatterplot of subjective frequency (ordinate) and log written frequency (abscissa). The solid line denotes the regression line.

Zero-order local discrepancies. A bivariate regression analysis was performed predicting subjective frequency from objective frequency, for the sole purpose of estimating the relative *discrepancy* between these sources of information for each word. Comparing the two variables is otherwise problematic, since they do not share the same metric, although other mathematical solutions might be envisioned. This regression analysis yielded a standardized residual (deleted) for each word, which can be interpreted like *z* scores, evaluating how well individual scores fit within a linear least-squares model relating the two variables (e.g., it is possible to apply a critical standardized residual value such as 1.96; for a list of extreme values, see the Appendix; the full listing is available for download). A scatterplot of the regression analysis that generated these residuals is reported in Figure 2.

Predicting zero-order discrepancy. The relationship between the potential bias-causing factors considered above and the disagreement between subjective frequency and written frequency may be more evident if the issue is examined directly. A regression analysis predicting the discrepancy between written frequency and subjective frequency (i.e., unstandardized residuals from the preceding analysis; henceforth, *discrepancy index*) from the main effects of the potentially bias-causing components of imageability, spoken frequency, and *N* explained 17.2% of the variance in the initial discrepancy index [$F(3,6198) =$

430.25, $p < .001$]. The fit of this model improved significantly with the inclusion of the statistically significant interaction terms reported above, including the interactions among the bias-causing components and written frequency. The addition of the four interaction terms contributed an additional 3.7% of variance explained [$F_{change}(4,6194) = 71.47, p < .001$]. It is important to note that the predictors representing the interactions were adjusted for their lower order effects (e.g., written frequency was partialed out of all associated interactions) prior to analysis so that they would represent only interaction variance. The full model explained 20.9% of the discrepancy index variance [$F(7,6194) = 233.62$, $p < .001$]. These results provide direct support for the idea that some of the discrepancy between objective and subjective frequency is more than random noise; it is systematically related to other variables.

Bias-free local discrepancies. In addition to looking at the simple bivariate discrepancy between written frequency and subjective frequency, we generated a second index that was corrected as much as possible—which is to say, globally—for the potential bias-causing variance identified in the preceding analysis. This goal was achieved by partialing out this variance from subjective frequency. Combined, all sources of bias, including the interactions, explained 12.8% of the variance in subjective frequency [$F(7, 136.91) = 129.63$, $MS_e = 1.06$, $p < .001$]. The residuals from this analysis constitute a relatively bias-free estimate of subjective frequency. A rescaled version of these residuals (to reflect the original 7-point scale) was then predicted by written frequency in a standard regression analysis. The results reveal that adjusted subjective frequency is significantly predicted by written frequency [$r = .69$, $r^2 = .48$; $F(1,6200) = 5{,}613.52$, $MS_e = 0.554$, $p < .001$]. This association presents an additional 6% of overlap between subjective frequency and written frequency over that observed prior to adjustment. The relationship between adjusted subjective frequency and written frequency is represented graphically in Figure 3.

Note that the relationship is much more linear now, since many of the discrepancies associated with low written frequency words in Figure 2 (preadjustment) have been corrected or minimized. The fact that scores at the lower end of objective frequency remain relatively more dispersed is not unexpected, since it was suggested in the introduction that low-frequency words were more likely to be associated with validity problems within the written frequency variable (see also Desrochers & Thompson, 2009; Gonthier et al., in press), and these were not addressed by the adjustments made here.

Extreme local discrepancies. In the preceding, discrepancy between the two familiarity indicators was operationalized in two ways: before and after adjusting for potential bias variance in subjective frequency. We will report the most extreme cases (i.e., those exceeding 1.96 standard deviations from the regression line), along with the associated values presented in the Appendix. As an additional source of information, we performed a recursive outlier deletion procedure whereby successive regression models generated standardized residuals. Between each step, highly discrepant items (exceeding a 1.96 z score value) were discarded, and the procedure was stopped when no new outliers were detected. A total of 13 steps were required to reach the stop rule when this outlier screening procedure was employed using the preadjustment subjective frequency, by which time a total of 1,100 words were flagged as outliers (i.e., potentially invalid). Similarly, a total of 12 steps were required to reach the stop rule with the bias-free version of subjective frequency, resulting in the detection of 1,070 extremely discrepant words. The overlap in discarded items was not perfect, since only 722 words were flagged by both outlier screen procedures. For this reason, we will report the results of both procedures as two dichotomous variables, where the value 1 denotes words whose subjective frequency data should be treated as suspect and the value 0 denotes an item that was not flagged. A complete listing of the 6,202 words, along with their French-Canadian subjective frequency, composite objective frequency value, adjusted bias-free subjective frequency, two indices of disagreement (pre- and postadjustment), and the two fields (pre- and postadjustment) indicating whether or not the word was flagged as an outlier can be downloaded (see the Supplemental Materials section).

We will close with a note regarding the interpretation of the standardized residual values presented here as indices of disagreement. Positive values mean that written frequency underestimates the word's familiarity, subjective frequency overestimates it, or both. Conversely, negative values are consistent with an underestimation of familiarity by subjective frequency, an overestimation of familiarity by written frequency, or both. Regardless of sign, extreme standardized residuals, especially after confounding variables are controlled for, indicate a serious problem in either the written frequency estimate or the subjective frequency estimate (i.e., disagreement), because each is

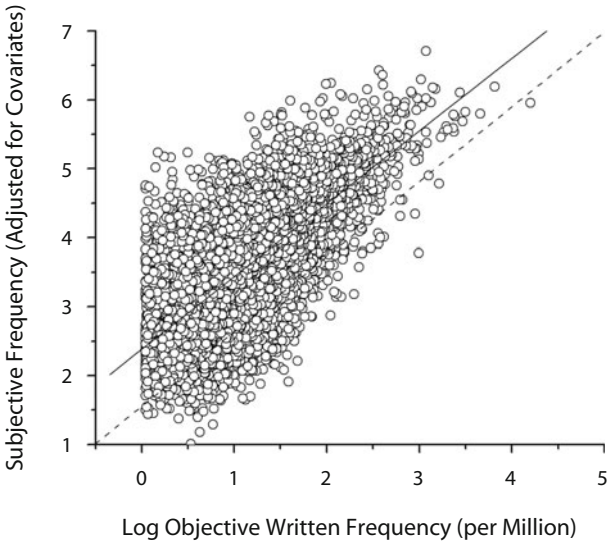

Figure 3. Scatterplot of adjusted subjective frequency (ordinate) and log written frequency (abscissa). The solid line denotes the regression line.

taken as an estimate of the same underlying trait: the familiarity of written words. The discrepancy values for all words are made available so that researchers may decide for themselves how to use this information. We will explore their implications further in the Discussion section.

DISCUSSION

The results reported here support a number of broad conclusions about familiarity estimates and their relationship to each other. In the first place, the strong overall relationship between objective and subjective word frequency was confirmed, which is consistent with the results of many other studies carried out on the French lexicon (e.g., Bonin et al., 2003; Desrochers & Bergeron, 2000; Desrochers & Thompson, 2009; Gonthier et al., in press) and the English lexicon (e.g., Balota et al., 2001; Benjafield & Muckenheim, 1989; Stadthagen-Gonzalez & Davis, 2006; Toglia & Battig, 1978) and with the idea, implicit in the literature, that the two variables are valid indicators of the same psychological construct. Second, the analyses reported here revealed that subjective frequency shares a significant amount of *bias* variance with other lexical characteristics—namely, imageability, spoken frequency, and orthographic *N*. This result suggests that subjective frequency is an impure measure of familiarity (see also Baayen et al., 2006; Balota et al., 2001). Third, moderation tests provided support for the idea that the codeterminants of subjective frequency have an antagonistic relationship with one another, a finding that is broadly consistent with the hypotheses and data presented by Balota et al. (2001). Finally, written frequency and subjective frequency were found to disagree markedly for a large number of words, especially those of low objective written frequency. This last finding is consistent with the observation that low-frequency items tend to be underdispersed (Baayen, 2001), an attribute that is associated with both measurement instability and bias in objective estimates (see also Desrochers & Thompson, 2009; Gonthier et al., in press). We now will turn to a discussion of specific issues of interest.

Bias in Subjective Frequency

The variance that subjective frequency shares with variables other than objective written frequency was divided into two components: legitimate variance and bias variance. Legitimate shared variance was also shared with objective written frequency, thereby providing a sort of corroboration. In contrast, bias shared variance was uncorroborated by objective written frequency, which leaves its relationship to the key construct, familiarity, open to question. Indeed, there are aspects of the data that support the *bias* interpretation. The bias variance associated with imageability did not overlap with spoken frequency, a finding that tends to support the idea that this uncorroborated shared variance is unrelated to familiarity and, therefore, exerts a biasing influence on ratings. Spoken frequency, for instance, would not be expected to suffer from the bias toward low-frequency abstract words that is sometimes listed among the weaknesses of objective written frequency counts (e.g., Gonthier et al., in press). A limitation of the statistical correction applied to the discrepancy ratings is, of course, that it was necessarily applied globally, which means that some residual bias due to imageability may still be present for some words and yet undetected. Another potential limitation is that objective frequency estimates are relatively unreliable at the low end of the frequency continuum, which means that their use as a gold standard in the identification of bias variance in subjective frequency may be especially flawed in this range. Specifically, the amount of bias in subjective frequency caused by variables such as imageability may have been overestimated here, because objective written frequency was unable to capture legitimate associations among the familiarity construct and these other variables for words in the low-frequency range. This methodological limitation may even partially account for the interactions reported here. Thus, the convenient term *bias variance* should be interpreted with caution. In any case, the discrepancy estimates are included in the archived database both before and after controlling for codeterminants, so that researchers may decide for themselves which estimates most accurately reflect disagreement among the two indicators of orthographic familiarity considered here.

Objective Frequency and Underdispersion

Figures 1 and 2 confirm that words with a low objective frequency count are far more likely to be mismatched with subjective frequency. This result confirms those reported by Gonthier et al. (in press) and Desrochers and Thompson (2009), which is not surprising given that the data from these studies contributed to the French-Canadian subjective frequency norms used in the present analysis. Nevertheless, the analysis reported here offers additional value, in that a larger sample of words was involved and multiple sources of objective frequency and subjective frequency were used. Taken together, these results suggest that researchers should be cautious in selecting items on the basis of objective frequency alone, especially when these are drawn from the low-to-moderate range of objective frequency.

Antagonism and Subjective Frequency Estimation

Despite the importance of subjectively estimated variables such as imageability and frequency, questions related to the process of forming such estimates have not attracted much interest from researchers beyond the generally acknowledged fact that such variables are not pure measures of their underlying concept. Balota et al. (2001) explicitly addressed this issue when they posited an antagonistic relationship among various sources of information associated with a given word based on connectionist principles. From this perspective, words that are rated high on the variable being examined are less likely to be codetermined, whereas words that are rated low are more susceptible to contamination by other lexical properties. The results reported here are consistent with this idea and, by extension, the joint use of both objective frequency and subjective frequency in item selection, especially for low-frequency words.

Major Discrepancies: A Qualitative Analysis

The purpose of reporting the estimates of disagreement for each entry of a word pool is to provide researchers with some additional criteria for lexical item selection. The logic behind using this information for item screening is that marked disagreement between objective and subjective estimates of familiarity/frequency suggests a substantive local failure in validity. Without knowing for sure where the problem lies, such items are best avoided in research. In support of this recommendation, we report whether items were classified as outliers in the bivariate model relating objective and subjective written frequency. This information may prove useful for constraining the pool of words from which experimental stimuli can be drawn. By avoiding the extreme and, therefore, most suspect items, researchers may have greater confidence in the validity of the relative frequency ratings they use, whether they be subjective or objective in nature. The pool of suspect items reported here comprises approximately 20% of an already sizable sample. This observation suggests that local failures in validity are a fairly common occurrence. This being the case, it is preferable that objectively defined criteria constrain the freedom of researchers to pick and choose among words with equivalent objective frequencies but vastly different real familiarity (e.g., Forster, 2000).

Some insight into the causes underlying these discrepancies can be gained via a qualitative analysis of the outlier cases identified in the first round of screening, which constitute the discrepancy indices reported here (for a partial list, see the Appendix). At the extreme high end, among other items, we found such words as *imprimante* (printer), *disquette* (floppy disk), *vidéo* (video), *hockey*, and *météo* (weather report, "the weather"), which are computer-related or media terms that are unlikely to be well represented in printed literary sources. At the extreme low end, we have a number of function words whose objective frequency tends to overestimate subjective frequency values. Verbs and function words seem to be especially vulnerable to such discrepancies. One possibility is that readers become relatively insensitive to highly frequent function words (e.g., determiners, conjunctions, and prepositions) over time. Among the content words in this range are *luire*, *coup*, *esprit*, *deux*, *lieu*, *temps*, *moment*, *saint*, *siècle*, *maréchal*, *homme*, *âme*, *air*, *coureur*, *monde*, *abbé*, *oeil*, *dame*, *jour*, *foutre*, and *guerre. Luire* (glisten), in particular, is an example of a word that might be deemed more literary than most, and whose frequency would therefore be overestimated on the basis of traditional corpuses. Other words, such as *siècle*, *coup*, *temps*, *moment*, *jour*, and *homme*, may be used so frequently in print (e.g., in various stereotyped idiomatic expressions) that participants treat them like function words when they rate them for subjective frequency. These interpretations are speculative, but they suggest that many of these discrepancies are more than just random fluctuations about the regression line. Specifically, for a word of any given relative frequency, there may be principled reasons for favoring objective over subjective frequency in the item selection process, and vice versa. In the absence of such principled motivations, however, it is preferable that words be selected for which estimates of both subjective and objective frequency agree.

CONCLUSION

The aim of this study was to explore the possibility that both subjective frequency and objective written frequency are flawed indicators of their common psychological construct: familiarity. The sources of bias attributed to these variables were examined in detail, and indices of disagreement between them were reported. This information provides an additional criterion for experimental stimulus selection, reducing the risk of selecting items with invalid estimates. Researchers may choose to examine the uncorrected disagreement index or the index that is based on subjective frequency estimates that have been adjusted for three codeterminants (imageability, spoken frequency, and orthographic *N*). The bias-free estimate of subjective frequency is reported for all the words used in the analysis. Use of this indicator as a predictor reduces the risk of masking the contribution of other important variables. Overall, the statistical analyses conducted on the database of 6,202 words support the use of both subjective frequency and objective frequency ratings when one attempts to control or manipulate familiarity, especially with words of moderate to low objective written frequency.

AUTHOR NOTE

This research was funded in part by a grant from the Canadian Language and Literacy Research Network (to Jean Saint-Aubin, Raymond Klein, and A.D.) and a grant from the Social Sciences and Humanities Research Council of Canada (to A.D. and Stanislaw Szpakowicz). We thank two anonymous reviewers for their feedback on the manuscript. Correspondence concerning this article should be addressed to G. L. Thompson, School of Psychology, University of Ottawa, 145 Jean-Jacques Lussier, P.O. Box 450, Station A, Ottawa, ON, KIN 6N6 Canada (e-mail: glennlthompson@gmail.com).

REFERENCES

ADELMAN, J. S., BROWN, G. D. A., & QUESADA, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, **17**, 814-823. doi:10.1037/a0012887

ALLEN, P. A., SMITH, A. F., LIEN, M.-C., GRABBE, J., & MURPHY, M. D. (2005). Evidence for an activation locus of the word-frequency effect in lexical decision. *Journal of Experimental Psychology: Human Perception & Performance*, **31**, 713-721. doi:10.1037/0096-1523.31.4.713

BAAYEN, R. H. (2001). *Word frequency distributions*. Boston: Kluwer.

BAAYEN, R. H., FELDMAN, L. B., & SCHREUDER, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory & Language*, **55**, 290-313. doi:10.1016/j.jml.2006.03.008

BAAYEN, R. H., PIEPENBROCK, R., & GULIKERS, L. (1995). *The Celex lexical database* [CD-ROM–Release 2]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

BALOTA, D. A., & CHUMBLEY, J. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance*, **10**, 340-357. doi:10.1037/0096-1523.10.3.340

BALOTA, D. A., CORTESE, M. J., SERGENT-MARSHALL, S. D., SPIELER, D. H., & YAP, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, **133**, 283-316. doi:10.1037/0096-3445.133.2.283

BALOTA, D. A., FERRARO, F. R., & CONNOR, L. T. (1991). On the early

influence of meaning in word recognition: A review of the literature. In P. J. Schwanenflugel (Ed.), *The psychology of word meaning* (pp. 187-222). Hillsdale, NJ: Erlbaum.

BALOTA, D. A., PILOTTI, M., & CORTESE, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, **29**, 639-647.

BAUER, D. J., & CURRAN, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, **40**, 373-400. doi:10.1207/s15327906mbr4003_5

BENJAFIELD, J., & MUCKENHEIM, R. (1989). Dates of entry and measures of imagery, concreteness, goodness, and familiarity for 1,046 words sampled from the Oxford English Dictionary. *Behavior Research Methods, Instruments, & Computers*, **21**, 31-52.

BERTRAM, R., & HYÖNÄ, J. (2003). The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long Finnish compounds. *Journal of Memory & Language*, **48**, 615-634. doi:10.1016/S0749-596X%2802%2900539-9

BONIN, P., MÉOT, A., AUBERT, L., MALARDIER, N., NIEDENTHAL, P., & CAPPELL-TOCZEK, M.-C. (2003). Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots. *Année Psychologique*, **104**, 655-694.

BORSBOOM, D., MELLENBERGH, G. J., & VAN HEERDEN, J. (2004). The concept of validity. *Psychological Review*, **111**, 1061-1071. doi:10.1037/0033-295X.111.4.1061

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

COHEN, P., COHEN, J., WEST, S. G., & AIKEN, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

COLTHEART, M., DAVELAAR, E., JONASSON, J. T., & BESNER, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). Hillsdale, NJ: Erlbaum.

COLTHEART, M., RASTLE, K., PERRY, C., LANGDON, R., & ZIEGLER, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, **108**, 204-256.

CONTENT, A. (1991). The effect of spelling-to-sound regularity on naming in French. *Psychological Research*, **53**, 3-12. doi:10.1007/BF00867327

DE GROOT, A. M. B. (1989). Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 824-845. doi:10.1037/0278-7393.15.5.824

DESROCHERS, A. (2006). OMNILEX: Une base de données sur le lexique du français contemporain. *Cahiers Linguistiques d'Ottawa*, **34**, 25-34.

DESROCHERS, A., & BERGERON, M. (2000). Valeurs de fréquence subjective et d'imagerie pour un échantillon de 1,916 substantifs de la langue française. *Canadian Journal of Experimental Psychology*, **56**, 274-325. doi:10.1037/h0087347

DESROCHERS, A., PAIVIO, A., & DESROCHERS, S. (1989). L'effet de la fréquence d'usage des noms inanimés et de la valeur prédictive de leur terminaison sur l'identification du genre grammatical. *Canadian Journal of Psychology*, **43**, 62-73. doi:10.1037/h0084253

DESROCHERS, A., & THOMPSON, G. L. (2009). Subjective frequency and imageability ratings for 3,600 French nouns. *Behavior Research Methods*, **41**, 546-557.

ELLIS, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, **24**, 143-188. doi:10.1017/S0272263102002024

FORGET, H. (2005). Valeurs d'imagerie et de fréquence subjective de 354 mots du vocabulaire sexuel de la langue française. *Revue canadienne des sciences du comportement*, **37**, 49-69. doi:10.1037/h0087245

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

GERNSBACHER, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, **113**, 256-281. doi:10.1037/0096-3445.113.2.256

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

GLOVER, S., & DIXON, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, **11**, 791-806.

GONTHIER, I., DESROCHERS, A., THOMPSON, G. L., & LANDRY, D. (in press). Normes d'imagerie et de fréquence subjective pour 1,760 mots monosyllabiques de la langue française. *Canadian Journal of Experimental Psychology*.

HINO, Y., & LUPKER, S. J. (2000). Effects of word frequency and spelling-to-sound regularity in naming with and without preceding lexical decision. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 166-183. doi:10.1037/0096-1523.26.1.166

IMBS, P. (ED.) (1971). *Études statistiques sur le vocabulaire français: Dictionnaire des fréquences. I: Tables alphabétiques*. Paris: Didier.

JAMES, C. T. (1975). The role of semantic information in lexical decisions. *Journal of Experimental Psychology: Human Perception & Performance*, **1**, 130-136. doi:10.1037/0096-1523.1.2.130

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

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

LÉTÉ, B., SPRENGER-CHAROLLES, L., & COLÉ, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavioral Research Methods, Instruments, & Computers*, **36**, 156-166.

LUPKER, S. J., BROWN, P., & COLOMBO, L. (1997). Strategic control in a naming task: Changing route or changing deadlines? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 570-590. doi:10.1037/0278-7393.23.3.570

LYNE, A. A. (1985). *The vocabulary of French business correspondence: Word frequencies, collocations and problems of lexicometric method*. Geneva: Slatkine-Champion.

MONSELL, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 148-197). Hillsdale, NJ: Erlbaum.

MONSELL, S., DOYLE, M., & HAGGARD, P. (1989). The effect of frequency on visual word recognition: Where are they? *Journal of Experimental Psychology: General*, **118**, 43-71. doi:10.1037/0096-3445.118.1.43

NELSON, D. L., & MCEVOY, C. L. (2000). What is this thing called frequency? *Memory & Cognition*, **28**, 509-522.

NEW, B., BRYSBAERT, M., VÉRONIS, J., & PALLIER, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, **28**, 661-677. doi:10.1017/S014271640707035X

NEW, B., & PALLIER, C. (2005). *Manuel de Lexique 3* (Unpublished technical report, version 3.03). Boulogne-Billancourt, France: Université Paris Descartes, Laboratoire de psychologie expérimentale.

NEW, B., PALLIER, C., BRYSBAERT, M., & FERRAND, L. (2004). *Lexique 2*: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, **36**, 516-524.

NORRIS, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, **113**, 327-357. doi:10.1037/0033-295X.113.2.327

PREACHER, K. J., & HAYES, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, **36**, 717-731.

SAILOR, K., BROOKS, P. J., BRUENING, P. R., SEIGER-GARDNER, L., & GUTERMAN, M. (in press). Exploring the time course of semantic interference and associative priming in the picture–word interference task. *Quarterly Journal of Experimental Psychology*. doi:10.1080/17470210802254383

SCARBOROUGH, D. L., CORTESE, C., & SCARBOROUGH, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception & Performance*, **3**, 1-17. doi:10.1037/0096-1523.3.1.1

SCHARLAU, I., & NEUMANN, O. (2003). Temporal parameters and time course of perceptual latency priming. *Acta Psychologica*, **113**, 185-203. doi:10.1016/S0001-6918%2802%2900157-9

SCHILLING, H. E. H., RAYNER, K., & CHUMBLEY, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, **26**, 1270-1281.

SEARS, C. R., SIAKALUK, P. D., CHOW, V. C., & BUCHANAN, L. (2008).

Is there an effect of print exposure on the word frequency effect and the neighborhood size effect? *Journal of Psycholinguistic Research*, **37**, 269-291. doi:10.1007/s10936-008-9071-5

STADTHAGEN-GONZALEZ, H., & DAVIS, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, **38**, 598-605.

STRAIN, E., & HERDMAN, C. M. (1999). Imageability effects in word naming: An individual differences analysis. *Canadian Journal of Experimental Psychology*, **53**, 347-359. doi:10.1037/h0087322

TAFT, M., & MEUNIER, F. (1998). Lexical representation of gender: A quasiregular domain. *Journal of Psycholinguistic Research*, **27**, 23-45. doi:10.1023/A:1023270723066

TAINTURIER, M. J., TREMBLAY, M., & LECOURS, A. R. (1992). Educational level and the word frequency effect: A lexical decision investigation. *Brain & Language*, **43**, 460-474. doi:10.1016/0093-934X%2892%2990112-R

THOMPSON, B., & VACHA-HAASE, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational & Psychological Measurement*, **60**, 174-195. doi:10.1177/00131640021970448

THOMPSON, G. L. (2008). Eliminating aggregation bias in experimental research: Random coefficient analysis as an alternative to performing a "by-subjects" and/or "by-items" ANOVA. *Tutorials in Quantitative Methods for Psychology*, **4**, 21-34.

TOGLIA, M. P., & BATIG, W. F. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Erlbaum.

NOTES

1. We thank an anonymous reviewer for pointing out that in some cases, extraneous variables, even unmeasured variables, can be neutralized by employing an appropriate counterbalancing scheme. The example given was that of a priming paradigm whereby the same items would serve as both *related* and *unrelated* primes for target items that would be presented twice, once in a related context and once preceded by an unrelated prime. The reviewer argued that in this case, the words would serve as their own controls. We concede that in cases in which such designs are possible (e.g., where priming can be used; where experimental conditions are varied), they provide an interesting alternative to stimulus matching and statistical control. Such an approach does have limitations, however, in that (1) it virtually commits researchers to using the ANOVA approach and, therefore, a relatively restricted number of items (for a discussion of the limitations of this approach, see Balota et al., 2004); (2) priming is a relatively complex paradigm whose properties are still the object of active research (e.g., the direction and magnitude of effects can depend on timing; see Sailor, Brooks, Bruening, Seiger-Gardner, & Guterman, in press; Scharlau & Neumann, 2003); and (3) repetitions of items, even when *neutralized* by counterbalancing within participants, create a special experimental context from which results may not generalize (for a general discussion of context effects, especially those created by short stimulus lists, see Balota et al., 2004). In any case, experiments that employ stimulus list matching are currently much more common.

2. Some corpora are specifically intended to capture the lexical frequency of particular registers, styles, or domains (e.g., business correspondence; Lyne, 1985). We are concerned here with frequency data that are intended or used as an overall estimate of frequency of use in print.

3. Why the "average" reader? Why not estimate custom frequency estimates for each participant? In the first place, to collect individualized frequency estimates would be costly. Second, it is not necessary to do so in any case. The dominant method of analyzing data in the psycholinguistic literature is the analysis of mean accuracy or mean reaction times, whereby each participant responds to multiple stimuli in every condition, which creates a set of participant $\times$ condition cells. The data points in these cells are normally summarized by a mean prior to analysis (G. L. Thompson, 2008). The ensuing ANOVA tests the null hypothesis that there is no difference between conditions in mean reaction time for the group of participants. The hypothesis being tested is at the level of the group, as opposed to the level of the individual, which is why the "average reader" is what matters when it comes to estimating lexical frequency or familiarity. Individual differences related to idiosyncratic representations of frequency are either extracted as individual differences or represented in the ANOVA error term. The same is true within a multilevel random coefficient framework (see G. L. Thompson, 2008), although in this case, it is possible, in principle, to reduce such error by estimating frequency for each participant individually (or for subgroups of participants, as in the use of separate estimates for men and women), a sort of customized frequency estimate. To date, the multilevel strategy has been seldom employed, and therefore, little use has been made of ratings that are specific to subgroups (for a split by gender, see Desrochers & Bergeron, 2000; for a split by age, see Balota, Pilotti, & Cortese, 2001).

4. An exception is Lété, Sprenger-Charolles, and Colé (2004), who used book sale data to inform the way they sampled books for their corpus. More typically, an arbitrary (with respect to reader experience) sampling procedure is selected and then described in detail.

5. Forster (2000) demonstrated that researchers can reliably guess which items will be responded to most quickly and accurately. Without safeguards in place, it would be possible for researchers to unwittingly select items for which the objective frequency estimates are invalid, creating lists that are equivalent in frequency "on paper," but not in psychological impact. The erroneous frequency estimates that exist in a database allow more room for item selection bias than might be possible otherwise, such as when an impartial estimate of a frequency value's validity is made available (such estimates for French words are available for download as supplemental materials with this article).

6. New et al. (2007) expressed the disagreement between two sources of spoken frequency on individual words as ratios. This information was used to identify sets of highly discrepant words for the purpose of describing local failures of validity within both sources (e.g., inflated estimates due to methodology of one source or the other). Some of the analyses presented here employ a similar reasoning but rely on standardized estimates of discrepancy (i.e., standardized residuals) that are more meaningful expressions of disagreement within a quantitative model of the data. Furthermore, unlike New et al. (2007), who were concerned mostly with establishing the validity of frequency counts based on movie subtitles in an ad hoc manner, we are explicitly casting estimates of disagreement as an index of the validity of individual ratings to be used by researchers in the selection of stimuli.

7. The French-Canadian subjective frequency and imageability norms were computed for two reasons: (1) to improve the accuracy of item scores by averaging over multiple sources, where available, and (2) to provide a single source of ratings for a large pool of words. The methodological features of the studies allowed them to be ranked by stability (i.e., total number of items, number or ratings per item, psychometric properties). On the basis of this evaluation, the Desrochers and Thompson (2009) study was taken as a gold standard. The ratings from the Desrochers and Bergeron (2000) study, correlated with the Desrochers and Thompson study at .95, were then combined with the gold standard either by accepting new words and their ratings or by computing an average rating in the case of overlap with the gold standard. Importantly, Desrochers and Bergeron ratings were discarded, instead of computing an average rating, if they were over three units of the standard error away from the gold standard rating [(3)(.15) = .45; Desrochers & Thompson, 2009]. In a similar manner, the Forget (2005) and Gonthier et al. (in press) norms were each integrated into this running average in turn, with pre-merge correlations of .92 and .90, respectively. In total, the outlier screening procedure resulted in the exclusion of only 11 data points. The merging procedure yielded a total of 6,202 mean ratings, each of which was based on at least 30 participants, for both subjective frequency and imageability.

8. Variance in subjective frequency associated with the following variables was partialed out: bias-only imageability, bias-only spoken frequency, and bias-only *N*; the two-way interactions of written frequency and imageability, written frequency and spoken frequency, imageability and spoken frequency (minus variance correlated with the main effects); and the three-way interaction between imageability, written frequency, and spoken frequency (minus variance correlated with the main effects).

SUPPLEMENTAL MATERIALS

The full set of word norms discussed in this article may be downloaded from brm.psychonomic-journals.org/content/supplemental.

(*Continued on next page*)

APPENDIX
List of Items With Extreme Local Discrepancy Values

Extreme Low End		Extreme High End	
Item	Standardized Residual	Item	Standardized Residual
Discrepancy Index Prior to Statistical Adjustment of Subjective Frequency			
maréchal	−3.23	stress	3.72
luire	−3.10	hockey	3.61
taillis	−2.98	imprimante	3.59
métaphysique	−2.86	sexy	3.58
abbé	−2.86	vidéo	3.55
sultan	−2.74	prof	3.55
talus	−2.69	week-end(s)	3.51
ire	−2.64	météo	3.39
képi	−2.62	pollution	3.36
monceau	−2.58	pizza	3.35
baron	−2.57	dance	3.32
boche	−2.56	ordinateur	3.26
ruer	−2.48	hamburger	3.24
rabbin	−2.46	disquette	3.18
hêtre	−2.46	spaghetti	3.17
mansarde	−2.45	jalouse	3.08
môme	−2.44	lunch	3.08
tumulte	−2.44	diète	3.04
tréteau	−2.44	lavage	3.02
platane	−2.43	frigidaire	3.00
écu	−2.39	mercredi	2.99
pèlerin	−2.36	cuillère	2.98
godasse	−2.35	seule	2.95
décret	−2.34	laveuse	2.94
être	−2.31	shampooing	2.94
gaillard	−2.30	sécheuse	2.88
obus	−2.28	technologie	2.87
faubourg	−2.26	condom	2.87
cloison	−2.24	campus	2.87
étoffe	−2.24	margarine	2.85
ignominie	−2.21	job	2.84
laquais	−2.21	brocoli	2.83
exaltation	−2.19	tuque	2.83
cité	−2.16	lundi	2.83
voûte	−2.16	basket-ball	2.83
fragment	−2.16	hot(-)dog	2.83
puritaine	−2.15	épicerie	2.82
abîme	−2.15	barbecue	2.81
échine	−2.15	pleine	2.77
labeur	−2.14	pénis	2.73
goulot	−2.13	céleri	2.73
villa	−2.13	jeudi	2.73
gibecière	−2.13	menstruation	2.71
margelle	−2.13	céréale	2.71
flanc	−2.12	réservation	2.69
duperie	−2.12	vendredi	2.67
fiacre	−2.12	patate	2.65
façade	−2.11	calorie	2.64
truand	−2.11	merci	2.63
vague	−2.11	sandwich	2.62
antichambre	−2.10	macaroni	2.62
lande	−2.09	steak	2.61
cohue	−2.09	dollar	2.61
larron	−2.09	orgasme	2.60
dédicace	−2.09	ciels	2.59
nef	−2.09	agenda	2.59
las	−2.08	math	2.57

APPENDIX (Continued)

Extreme Low End		Extreme High End	
Item	Standardized Residual	Item	Standardized Residual
jante	−2.08	pure	2.56
maître	−2.08	mardi	2.54
cortège	−2.08	pourcentage	2.54
lucarne	−2.07	ingrédient	2.52
jatte	−2.06	caissière	2.52
buffle	−2.05	réfrigérateur	2.51
marquise	−2.05	chandail	2.50
anarchiste	−2.05	température	2.50
prunelle	−2.04	douche	2.49
cavalier	−2.04	croche	2.47
paillasson	−2.03	mitaine	2.47
lassitude	−2.02	logiciel	2.44
tyran	−2.01	dure	2.44
préfet	−2.00	séchoir	2.41
roseau	−1.99	rabais	2.40
torpilleur	−1.99	facture	2.40
comble	−1.99	droitier	2.40
âtre	−1.98	skier	2.38
lubie	−1.98	chicane	2.38
clebs	−1.97	jaloux	2.38
rempart	−1.97	écoute	2.37
aïeul	−1.97	banane	2.37
		coke	2.36
		calculatrice	2.35
		molle	2.35
		massage	2.34
		souper	2.33
		grille-pain	2.32
		plagiat	2.32
		participant	2.31
		gardienne	2.31
		pilule	2.31
		fumeur	2.29
		dessert	2.29
		toast	2.27
		toutou	2.27
		lunettes	2.27
		garderie	2.26
		beigne	2.26
		télévision	2.26
		minuit	2.24
		patinage	2.24
		fourchette	2.23
		biscuit	2.22
		environnement	2.22
		pluriel	2.21
		gang	2.19
		triche	2.19
		pale	2.18
		carotte	2.18
		négatif	2.18
		taxe	2.18
		légume	2.18
		grosseur	2.17
		cartable	2.16
		maquillage	2.15
		gomme	2.15
		gagnant	2.14
		vitamine	2.13
		téléphone	2.13

APPENDIX (Continued)

Extreme Low End		Extreme High End	
Item	Standardized Residual	Item	Standardized Residual
		pouding	2.13
		pince(s)	2.13
		université	2.13
		politicien	2.13
		résumé	2.13
		canette	2.13
		mayonnaise	2.12
		magazine	2.12
		bisou	2.11
		amoureuse	2.11
		orteil	2.10
		tomate	2.10
		jus	2.10
		score	2.10
		impact	2.09
		cadenas	2.09
		contraception	2.08
		patin	2.08
		congélateur	2.08
		aspirateur	2.08
		horaire	2.08
		papa	2.07
		test	2.07
		infirmière	2.07
		masturbation	2.07
		coiffeuse	2.06
		maman	2.06
		coupon	2.05
		finale	2.05
		piment	2.05
		chèque	2.04
		participe	2.04
		cuisine	2.04
		dimanche	2.04
		sexiste	2.04
		poivre	2.03
		salut	2.02
		jeep	2.02
		plastique	2.02
		paragraphe	2.02
		signature	2.02
		fichier	2.01
		téléviseur	2.01
		clitoris	2.01
		cancer	2.01
		étudiant	2.01
		toilette	2.01
		frite	2.00
		plie	2.00
		terrorisme	2.00
		midi	2.00
		nageuse	2.00
		skieur	2.00
		samedi	1.99
		lesbienne	1.99
		février	1.99
		sofa	1.98
		tatou	1.98
		nouille	1.98
		masturber	1.97

APPENDIX (Continued)

Extreme Low End		Extreme High End	
Item	Standardized Residual	Item	Standardized Residual
Discrepancy Index Following the Statistical Adjustment of Subjective Frequency			
rabbin	−3.00	week-end(s)	3.83
maréchal	−2.92	jalouse	3.67
pote	−2.91	diète	3.55
môme	−2.88	calorie	3.36
abbé	−2.78	pure	3.34
sultan	−2.77	stress	3.34
foutre	−2.75	dance	3.25
buffle	−2.75	pollution	3.22
fric	−2.74	participe	3.12
baron	−2.73	prof	3.09
pharaon	−2.72	seule	3.04
truand	−2.62	étude	3.03
bagnole	−2.58	hockey	3.01
colonel	−2.56	chicane	2.99
cavalier	−2.56	miens	2.99
pèlerin	−2.54	hier	2.95
villa	−2.53	météo	2.95
salope	−2.52	lunch	2.94
luire	−2.52	mercredi	2.94
godasse	−2.49	due	2.93
baïonnette	−2.44	vidéo	2.93
donjon	−2.43	après-midi	2.90
cité	−2.41	lavage	2.89
maître	−2.40	imprimante	2.85
paquebot	−2.40	pizza	2.84
bouclier	−2.38	ordinateur	2.81
paillasson	−2.38	disponibilité	2.80
boche	−2.37	sexisme	2.77
autel	−2.37	pluriel	2.76
tyran	−2.37	sexy	2.74
bonnet	−2.35	pleine	2.73
artillerie	−2.34	libérale	2.73
roseau	−2.33	amoureuse	2.73
slip	−2.33	plagiat	2.70
télégramme	−2.32	lundi	2.68
étoffe	−2.30	nouvelle	2.68
fragment	−2.29	menteuse	2.68
mammouth	−2.27	molle	2.66
commandant	−2.27	croche	2.65
dragon	−2.27	jeudi	2.62
marchand	−2.27	participant	2.60
caillou	−2.25	synonyme	2.57
façade	−2.25	pourcentage	2.57
pantin	−2.23	température	2.56
obus	−2.23	souper	2.56
vautour	−2.23	cuillère	2.56
flanc	−2.22	épicerie	2.56
nonne	−2.22	jaloux	2.56
képi	−2.21	fiabilité	2.55
cuvette	−2.20	droitier	2.54
brigand	−2.20	math	2.53
empereur	−2.19	spaghetti	2.52
machin	−2.18	hamburger	2.52
sabot	−2.18	coucher	2.51
vague	−2.18	rabais	2.49
rivage	−2.18	menstruation	2.49
marquis	−2.17	mien	2.49
roi	−2.17	planification	2.48

APPENDIX (Continued)

Extreme Low End		Extreme High End	
Item	Standardized Residual	Item	Standardized Residual
cloison	−2.17	vendredi	2.48
lieutenant	−2.17	mardi	2.47
salaud	−2.16	laveuse	2.46
cavalerie	−2.16	pale	2.45
gaillard	−2.15	caissière	2.44
parchemin	−2.15	disquette	2.44
cortège	−2.14	réservation	2.44
chevalier	−2.13	écoute	2.43
gare	−2.13	grosseur	2.43
gamin	−2.13	gardienne	2.43
monastère	−2.12	pronom	2.43
canon	−2.12	prévention	2.42
duc	−2.12	campus	2.42
ravin	−2.11	dure	2.42
harpon	−2.11	préposition	2.39
flingue	−2.11	complément	2.38
berger	−2.10	résumé	2.38
montagnard	−2.09	février	2.36
dame	−2.09	adjectif	2.36
sentinelle	−2.08	sexiste	2.36
valet	−2.07	dollar	2.36
antiquaire	−2.07	plie	2.34
yacht	−2.06	relaxation	2.34
baraque	−2.06	douche	2.33
clairon	−2.04	assistanat	2.32
oursin	−2.04	patate	2.30
taillis	−2.03	triche	2.29
clebs	−2.03	négatif	2.28
laquais	−2.03	ohm	2.28
blaireau	−2.02	taxe	2.27
talus	−2.02	shampooing	2.27
scène	−2.02	midi	2.26
hêtre	−2.02	adverbe	2.26
sorcier	−2.02	sécheuse	2.25
chevalet	−2.02	décembre	2.25
palais	−2.01	actualité	2.25
tonneau	−2.01	total	2.25
galère	−2.01	tiède	2.25
parquet	−2.01	excitant	2.24
brancard	−2.00	ajout	2.24
mulet	−2.00	paresseux	2.24
moine	−2.00	verbe	2.24
pasteur	−2.00	céréale	2.23
cafard	−2.00	dimanche	2.22
tronche	−1.99	nôtre	2.22
abri	−1.99	frigidaire	2.22
guillotine	−1.99	ingrédient	2.22
locomotive	−1.99	dissertation	2.22
garce	−1.99	moyenne	2.21
tourelle	−1.99	coordinatrice	2.21
tartine	−1.98	neutre	2.20
cale	−1.97	révision	2.20
vipère	−1.97	finale	2.19
cocotier	−1.97	introduction	2.19
		tôt	2.19
		orgasme	2.18
		job	2.17
		condom	2.16
		sandwich	2.16
		compatible	2.16

APPENDIX (Continued)

Extreme Low End		Extreme High End	
Item	Standardized Residual	Item	Standardized Residual
		technologie	2.15
		propreté	2.15
		pitre	2.15
		agenda	2.14
		stimulus	2.14
		chère	2.13
		pénis	2.13
		format	2.12
		conjugaison	2.12
		sacre	2.12
		bien-être	2.12
		avant	2.12
		senteur	2.12
		caractéristique	2.11
		jouable	2.11
		janvier	2.11
		chaude	2.11
		téléphone	2.10
		contraception	2.10
		logiciel	2.10
		ciels	2.10
		clarification	2.10
		participation	2.09
		papa	2.09
		résidente	2.09
		margarine	2.08
		maman	2.07
		souvenirs	2.06
		cuisine	2.05
		tuque	2.04
		évaluation	2.04
		minuit	2.04
		paiement	2.03
		barbecue	2.03
		définition	2.02
		oui	2.02
		admission	2.02
		spécification	2.02
		basket-ball	2.02
		blâme	2.01
		grammaire	2.01
		sexologie	2.01
		visibilité	2.00
		rires	2.00
		chandail	1.99
		lente	1.99
		banane	1.99
		terrorisme	1.99
		facture	1.98
		chauffe	1.98
		steak	1.98
		horaire	1.98
		partage	1.97

Note—For a complete listing, see the online supplemental materials.

(Manuscript received September 19, 2008;
revision accepted for publication February 24, 2009.)