Behavior Research Methods
2006, 38 (4), 598-605

The Bristol norms for age of acquisition, imageability, and familiarity

HANS STADTHAGEN-GONZALEZ and COLIN J. DAVIS
University of Bristol, Bristol, England

Ratings of age of acquisition (AoA), imageability, and familiarity were collected for 1,526 words. The methodology made use of a modular approach, in which the full sample of words was divided into five separate blocks. Within each block, each word was rated on each of the three variables by 20 participants (undergraduate students from the University of Bristol). Analyses comparing these ratings to existing norm databases demonstrated that this methodology resulted in high reliability (assessed by Cronbach's α) and validity. The ratings were also transformed to be compatible with the Gilhooly and Logie (1980) norms. This transformation resulted in a set of norms for 3,394 words, which is by far the largest database of ratings for AoA, imageability, and familiarity to date. The resulting database should be useful for researchers interested in manipulating or controlling these factors in word recognition, neuropsychological, or memory studies. These norms can be downloaded from language.psy.bris.ac.uk/bristol_norms.html.

Previous research has identified a number of variables that affect the speed and accuracy with which words can be recognized, recalled, named, and/or classified (see, e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Hulme, Stuart, Brown, & Morin, 2003; Roodenrys, Hulme, Alban, Ellis, & Brown, 1994; Whaley, 1978). Some of these variables, such as word length and word onset, are intrinsic to each word and can be determined directly from their surface structure, without reference to any other materials. Other variables, such as neighborhood size, word frequency, bigram frequency, or regularity depend on the relationship of the target item to a larger corpus of words; these values are estimated by placing the word within a certain linguistic context. Obtaining data for these two types of variables is, in general, not problematic: Given the existence of a reasonably comprehensive word frequency corpus, such variables can be obtained for virtually all of the words in a language, and various software tools exist to facilitate the retrieval of these statistics (see, e.g., Davis, 2005; Davis & Perea, 2005; Duyck, Desmet, Verbeke, & Brysbaert, 2004). However, there are also other variables that affect the speed and accuracy of word recognition and recall that are a reflection of the personal experiences of language users. Such variables include age of acquisition (AoA), imageability, and familiarity. Measurements for this type of variable are usually estimated by asking people to make subjective ratings (e.g., "How old do you estimate you were when you learned this word?").

Collecting norms for these subjective variables is a relatively time-consuming process, and most norming studies have been limited to the number of items that a participant can rate in one session of reasonable length. Most of the published norming studies have concentrated on only one or two of the three variables that we focus on here, and typically they have included less than 1,000 items.

The limited availability of ratings for variables like AoA, imageability, and/or familiarity poses a particular problem for researchers. When designing experimental stimuli, experimenters who wish to match stimuli on subjective variables like AoA must typically run their own norming study on candidate stimuli. More frequently, experimenters simply forgo the possibility of matching stimuli on these subjective variables (and hope for the best). The limited availability of norms for subjective variables also hampers the research strategy of performing multiple regressions on large-scale databases. For example, Balota et al.'s (2004) recent regression analyses of speeded naming and lexical decision latencies for 2,428 monosyllabic English words omitted AoA as a predictor variable because norms were only available for a quarter of the items in their set.

The present article aims to improve this situation by increasing the number of words for which subjective norms are available. The *Bristol norms* that we present here consist of ratings of AoA, imageability, and familiarity for 1,526 words. We chose to obtain norms for these particular variables because of the interest that each of them has generated in the language and memory literature in recent years, either as the explicit focus of study or as extraneous variables to control.

Age of acquisition refers to the age at which a word was learned and has been proposed as a significant contributor to language and memory processes (see, e.g., Carroll

This work was partially supported by ESRC Postdoctoral Fellowship PTA-026-27-0769 to H.S.-G., an Australian Research Council APD grant to C.J.D., and Biotechnology and Biological Sciences Research Council Grant 7/S17491 to Jeff Bowers and Markus Damian. We thank Jeff Bowers and Mike Cortese for their assistance. Correspondence relating to this article may be sent to H. Stadthagen-Gonzalez, Department of Experimental Psychology, University of Bristol, 12A Priory Road, Clifton, Bristol BS8 1TU, England (e-mail: hs0443@bris.ac.uk).

& White, 1973; Hirsh & Funnell, 1995; Juhasz & Rayner, 2003; Morrison, Ellis, & Quinlan, 1992; Roodenrys et al., 1994). Although some studies have used "objective" AoA norms (e.g., Morrison, Chappell, & Ellis, 1997), most experimenters choose to use subjective measures of AOA. These AoA estimates have been shown to be reliable (Carroll & White, 1973; Gilhooly & Logie, 1980) and to provide a valid estimate for the objective age at which a word was acquired (Gilhooly & Gilhooly, 1980; Morrison et al., 1997). *Imageability* is a semantic variable that measures how easy it is for a word to elicit mental images. It has been used to evaluate the effects of meaning on memory and word recognition (see, e.g., Balota, 1990; Balota et al., 2004; Paivio, 1971; Paivio, Yuille, & Madigan, 1968). Imageability is closely related to concreteness; for most words, the two measures are quite similar, although there are some exceptions (Bird, Franklin, & Howard, 2001, give the example of the word *armadillo*, which generates high concreteness but low imageability ratings, presumably because of a lack of personal exposure to armadillos). Imageability has been shown to be a better predictor of performance than concreteness in several studies (e.g., Marcel & Patterson, 1978; Richardson, 1975).

It is not entirely clear what processes are involved when readers make ratings of the *familiarity* of a word. Familiarity ratings have often been interpreted as a measure of the frequency of exposure to a word, and Gernsbacher (1984) suggested that familiarity is a better predictor of word performance than printed word frequency, particularly for low-frequency words. However, it has also been argued that familiarity ratings include a semantic component (see, e.g., Balota, Pilotti, & Cortese, 2001). The present approach of obtaining ratings of both subjective familiarity and imageability for a large sample of words (for which printed word frequency estimates are also available) allows us to investigate this issue further.

The relevance of each of these three variables has been challenged at one point or another. On the one hand, Zevin and Seidenberg (2002) have argued that effects purportedly due to AoA actually reflect the cumulative frequency of exposure to a word (though see Stadthagen-Gonzalez, Bowers, & Damian, 2004, for recent experimental evidence undermining this cumulative frequency hypothesis). On the other hand, after controlling for AoA, Morrison and Ellis (2000) failed to find effects of either imageability or familiarity on word naming latencies. Finally, although subjective familiarity has been used widely, its usefulness has been strongly challenged: Brown and Watson (1987) questioned Gernsbacher's (1984) claim that familiarity is a more suitable frequency measure than objective frequency, and Balota et al. (2001) proposed that subjective frequency is a more useful estimate than subjective familiarity. We have chosen these three variavvvbles precisely because of the many lines of debate surrounding them, which generate a need for normed stimuli that can be used to explore each position. Indeed, we think that the norms presented here would be equally useful for studies trying to prove or to disprove the relative importance of each of these variables.

To date, the largest database of ratings for AoA, imageability, and familiarity in English is the set compiled by Gilhooly and Logie (1980; henceforth G&L), which includes norms for these variables for 1,944 words, as well as measurements of concreteness and ambiguity. This database has proved to be a valuable resource for experimenters: The G&L norms have been referenced by at least 200 articles since their publication (and are also incorporated into the MRC psycholinguistic database; Coltheart, 1981a). In order to provide an even wider choice of words available to experimenters, the Bristol norms were designed to be compatible with the G&L norms with respect to the definition of the variables as well as the composition of the sample of participants (i.e., mostly young undergraduate students). Below, we report analyses that show that it is legitimate to merge the two databases. The combined Bristol/G&L norms provide normative data for 3,394 words, the largest set of norms for those variables so far in English. We consider that a wider availability of items will be of great help in the implementation of factorial designs or regression studies in the fields of language, memory, and neuropsychology. We also propose a modular approach to collecting norms, which enables incremental increases in the size of the database.

METHOD

Participants

Ratings from 100 participants were obtained for each of the variables under study (AoA, imageability, and familiarity). All participants were undergraduate students enrolled in level 1 or 2 psychology courses at the University of Bristol. They were all native British English speakers, average age 19.7 years, range 18–40.

The Word Corpus

A set of 1,526 words were rated in this study. Our selection of words was governed by two main criteria, both of which aimed to maximize the usefulness of the norms. The first was that the words should be fairly representative of the types of words typically used in psycholinguistic experiments. Consequently, we chose words that were relatively short (between four and seven letters, with one or two syllables) and excluded regular past tenses and plurals; some irregular past tense and plural forms were included (e.g., *bought*, *teeth*). Furthermore, although the full sample covered a range of frequencies from 0.34 to 1,642 counts per million in the CELEX database (Baayen, Piepenbrock, & van Rijn, 1995), we focused on words in the frequency range most often sampled by psycholinguistic experiments; thus, 93% of the words had CELEX written frequencies between 1 and 100 counts per million. The second criterion for inclusion in the set of words to be rated was that the word was not already included in the G&L norms. This resulted in a set of 1,450 words. In addition, we included a set of 76 words that also had AoA, imageability, and familiarity ratings in the G&L norms; these items were included as reliability controls.

Procedure

The 1,450 new words were randomly divided into five sets of equal size, and the 76 control words were added to each set. Each block of 366 words was presented in four columns on a computer spreadsheet and rated on one of the three variables (AoA, imageability, or familiarity) by 20 volunteers in sessions that lasted no more than 30 min. The words were presented in a different random order for each group of 4 participants. This procedure was repeated for AoA, imageability, and familiarity, with pertinent changes in the instructions for each variable as outlined below.

Raters of AoA were asked to type next to each word an estimate (in years) of when they learned the word. In order to make the Bristol norms compatible with the G&L norms, responses were then converted to the same 1–7 scale used by G&L (from 1 for ages 0–2 years to 7 for ages 13 years or more, with intervening bands spanning 2 years).

Raters of imageability were asked to indicate how easily each word elicited mental images. They indicated their answers on a scale from 1 to 7, where 1 corresponded to *low imageability* and 7 to *high imageability*. The instructions given to participants were almost identical to those used by G&L, which were based on the instructions devised by Paivio et al. (1968). The only differences in wording related to the method by which raters made their responses. Instead of circling the chosen number on the scale, participants typed it on a space given next to each word; the 7-point scale was visible at all times at the top of the screen.

Raters of familiarity were also asked to provide ratings using a 7-point scale, with 1 assigned to words that they never had seen and 7 to words that they had seen very often (nearly every day). The wording of the instructions for this task was also very similar to the wording provided by G&L. As in the case of the imageability ratings, participants typed the appropriate number next to each word, with the 7-point scale visible at all times.

In all cases, ratings on the 1–7 scale were subsequently multiplied by 100 and rounded to the nearest integer so that the ratings could be presented as integers on a scale from 100 to 700.

RESULTS AND DISCUSSION

The first step in the analysis involved combining the five separate blocks into a single data set. We used the 76 control words (which were rated by all participants) to verify the validity of this approach and to homogenize the ratings across blocks. The ratings for control words had interblock correlations of at least .91 across all three variables, and for all three variables the interrater reliability coefficient (Cronbach's α) was at least .98. We followed the linear transformation procedure outlined by Coltheart (1981b; see the Appendix) in order to homogenize the means and standard deviations across the five blocks. (The same procedure was used by Bird et al., 2001, to transform their norms to the same scale used for the G&L norms.) This transformation involved three steps. First, we computed the overall mean rating (across all 100 participants) for each of the 76 control words. We then generated a separate regression equation for each of the five blocks, using the overall means to predict the mean rating (across 20 participants). For each block of 366 words, we then transformed the raw mean rating of each item by subtracting the intercept of the regression equation for that block and dividing by the regression coefficient. The five blocks were then combined into a single data set. For the control words, the value entered into the database was the average of the transformed ratings for all five blocks.

We assessed the reliability of the resulting norms by examining the correlations between the ratings for the 76 control words in the Bristol norms and the corresponding ratings for the same words in the G&L norms. These correlations were very high (AoA, $r = .89$; IMG, $r = .93$; FAM, $r = .86$), and the interrater reliability between the Bristol and the G&L norms was also very high (AoA, $\alpha = .93$; IMG, $\alpha = .96$; FAM, $\alpha = .86$). There was also a very high correlation between our imageability norms and those recently reported by Cortese and Fugett (2004; $r = .84$, $N = 680$). We also examined the correlations with the standardized reaction times for lexical decision and word naming taken from the English Lexicon Project (ELP; Balota et al., 2002). For lexical decision latencies, the correlations were as follows: AoA, $r = -.27$; IMG, $r = -.13$; FAM, $r = -.39$. For naming latencies, the correlations were AoA, $r = -.08$; IMG, $r = -.16$; FAM, $r = -.10$. These correlations were very similar to those between the latency variables and the G&L norms across the same set of words (e.g., the correlation in G&L between AoA and lexical decision latency was $-.26$). There were no significant differences between the correlations for the two sets of norms (all ps $> .24$ using the procedure outlined by Meng, Rosenthal, & Rubin, 1992).

Across the entire sample, the correlations between each of the three subjective variables and standardized reaction times for lexical decision and word naming taken from the ELP (Balota et al., 2002) were highly significant (see Table 1).[1] This result offers some support for the validity of the collected ratings. Note that the correlations with the reaction time measures were particularly strong in the case of the AoA and familiarity ratings.

Correlations Between the Subjective Norms and Other Lexical Variables

Table 2 shows the correlations between AoA, imageability, and familiarity and a selection of lexical variables: word length (in letters, syllables, and phonemes), mean log bigram frequency (MLBF), neighborhood size (*N*), and a variety of measures of written and spoken frequency (CELEX written frequency [Baayen et al., 1995]; British National Corpus [BNC] written frequency [Burnard, 1995]; ratings from the *Educator's Word Frequency Guide* [WFG; Zeno, Ivens, Millard, & Duvvuri, 1995]; and BNC spoken frequency, based on the BNC demographic database). As can be seen, AoA is significantly correlated with each of these variables. The direction of these correlations is in accord with expectations regarding the age at which words are acquired. Thus, words that are acquired later tend to be longer, less frequent, less imageable, and less subjectively familiar and to have fewer neighbors and lower bigram frequencies than words that are acquired earlier. To investigate the independent contribution of each of

Table 1
Correlations Between Subjective Ratings From the Bristol Norms and Reaction Time Measures (Lexical Decision and Naming Latencies From the English Lexicon Project [ELP] Corpus; *N* = 1,503)

	Lexical Decision	Naming
AoA	.51**	.33**
IMG	−.22**	−.13**
FAM	−.53**	−.33**

Note—AoA, age of acquisition; IMG, imageability; FAM, familiarity. Lexical decision and naming reaction times correspond to the standardized latencies reported in the ELP corpus (Balota et al., 2002). **$p < .001$.

Table 2
Correlations Between Subjective Ratings From the Bristol Norms and Other Lexical Variables (Length, Bigram Frequency, *N*, and Frequency)

	Variable	AoA	IMG	FAM
1	AoA	+1.00	–	–
2	IMG	−0.53	+1.00	–
3	FAM	−0.61	+0.12	+1.00
4	LEN_L	+0.30	−0.21	−0.13
5	LEN_S	+0.35	−0.22	−0.15
6	LEN_P	+0.34	−0.22	−0.15
7	MLBF	−0.21	+0.05$^{n.s.}$	+0.14
8	*N*	−0.29	+0.18	+0.11
9	$\log_{10}$ (CELEX written + 1)	−0.48	+0.02$^{n.s.}$	+0.60
10	$\log_{10}$ (BNC written + 1)	−0.38	−0.05$^{n.s.}$	+0.57
11	$\log_{10}$ (WFG + 1)	−0.54	+0.09	+0.57
12	$\log_{10}$ (BNC spoken + 1)	−0.66	+0.25	+0.72

Note—AoA, age of acquisition; IMG, imageability; FAM, familiarity; LEN_L, word length in letters; LEN_S, word length in syllables; LEN_P, word length in phonemes; MLBF, mean log bigram frequency; *N*, neighborhood size; BNC, ratings from British National Corpus; WFG, ratings from *Educator's Word Frequency Guide*; see the text for further details. n.s., not significant at the .05 level; all other correlations are significant at the $p < .001$ level.

these variables, a simultaneous multiple regression was conducted with AoA as the dependent variable and six independent variables (see Table 3). To avoid problems of excessive multicollinearity among the independent variables, we used only a single measure of written word frequency (the WFG count) and a single measure of length (number of phonemes). As can be seen in Table 3, all six variables included in the regression made independent contributions to predicting rated AoA, with the best predictors being imageability and familiarity, followed by spoken and written frequency. This finding agrees with the conclusions drawn in previous investigations of AoA based on different sets of words (e.g., Bird et al., 2001; Gilhooly & Logie, 1980; Morrison et al., 1997).

Imageability is significantly correlated with AoA, familiarity, length, and *N*, but less clearly with word frequency. Thus, more imageable words tend to be acquired earlier, are more familiar, tend to be shorter, and tend to have more orthographic neighbors than less imageable words. It is unlikely that the correlation with *N* has any causal component; rather, it probably reflects the fact that shorter, earlier-acquired words have more neighbors than longer, later-acquired words. When AoA and length are partialed out, the partial correlation between *N* and imageability is −.01 (i.e., nonsignificant, negligible, and in the opposite direction from the raw correlation). With respect to the correlations between imageability and frequency, one of the three measures of written frequency (the CELEX written frequency count) shows a negligible correlation with imageability, another (the BNC written frequency count) shows a trend toward a negative correlation with imageability ($p = .051$), and the third (the WFG) shows a significant positive correlation with imageability; the latter correlation may reflect AoA, given that the WFG count is based exclusively on a corpora of school texts. It may be concluded that any correlation between imageability and written frequency is, at best, very weak. Likewise, the positive correlation between imageability and spoken frequency probably depends on the strong correlations with the third variable, AoA. When AoA is partialed out, the partial correlation between imageability and spoken frequency is negative (−.15; i.e., high-imageability words tend to be lower in spoken frequency).

Some authors (e.g., Zevin & Seidenberg, 2002, 2004) have proposed that AoA effects in visual word recognition actually reflect the effect of cumulative frequency—that is, words that are acquired earlier in life will have been encountered more often overall, when printed word frequency has been equated. Zevin and Seidenberg (2002, 2004) calculated cumulative frequency as the sum of frequency estimates for all grades included in the Zeno et al. (1995) norms. We computed cumulative frequency in the same way for the 1,307 words in our sample that are listed in the Zeno et al. database. The correlation between cumulative frequency and rated AoA was relatively high ($r = -.24$), which is at least consistent with the possibility that previously observed effects attributed to AoA could have been the result of a confound with cumulative frequency. However, the correlations between cumulative frequency and lexical decision ($r = -.21$) and word naming ($r = -.14$) latencies were much weaker than those for the AoA subjective estimates included in the Bristol norms, where rated AoA explained about 25% of the variance for lexical decision latencies, whereas cumulative frequency explained only about 4% of this variance. This difference leads us to conclude that the effects of rated AoA on visual word recognition performance are not simply due to cumulative frequency. Experimental investigations of the cumulative frequency hypothesis have reached the same conclusion (Ghyselinck, Lewis, & Brysbaert, 2004; Stadthagen-Gonzalez et al., 2004).

What Does Subjective Familiarity Measure?

One question that the norms may help to address is exactly what is being measured in subjective familiarity ratings. As noted already, it has been suggested that these ratings may provide a better measure of the relative frequency of exposure to a word than do objective measures of printed word frequency (see, e.g., Gernsbacher, 1984; Gilhooly & Logie, 1980). The Bristol norms show relatively strong correlations between familiarity and both

Table 3
Multiple-Regression Analysis With Rated Age of Acquisition as the Dependent Variable and Six Independent Variables

Independent Variable	*B*	*SE*	β	*t*
Imageability	−.373	.015	−.409	25.17
Familiarity	−.675	.044	−.342	15.17
Spoken frequency	−44.672	5.491	−.210	8.14
Written frequency	−22.897	5.654	−.087	4.05
Number of phonemes	8.531	2.084	.077	4.09
N	−1.915	.569	−.062	−3.37

Note—Spoken frequency, $\log_{10}$ (BNC spoken frequency + 1); Written frequency, $\log_{10}$ (WFG frequency + 1). All *t* values are significant at $p < .005$.

written and spoken frequency (see Table 2), which supports the idea that subjective familiarity ratings reflect frequency of exposure. One way to assess the claim that familiarity ratings provide a better measure of the relative frequency of exposure to a word than do objective measures of printed word frequency is to examine how the correlations between familiarity and objective frequency measures compare with the intercorrelations between different objective frequency measures. Theoretically, variance in measures of word frequency can be partitioned into two separate components: a systematic component, reflecting "true" word frequency, and a random component, reflecting measurement error. The measurement error component varies across metrics, with some frequency measures containing greater error variance than others. The correlation between different frequency measures will decrease as a function of the magnitude of the error variance in the two measures (e.g., there will be a correlation of 1 for two measures with no error variance, and a correlation approaching 0 if one or both of the measures has extremely large error variance). Thus, if subjective familiarity offers a better (e.g., less noisy) measure of frequency of exposure than do objective measures, subjective frequency should correlate more highly with these objective measures than the objective measures correlate with each other.[2] As Table 2 shows, the maximum correlation between subjective familiarity and any of the objective written frequency measures was .60. By contrast, the minimum correlation between the objective written frequency measures was .83; in the case of the two objective measures based on British English (the CELEX and BNC counts), the correlation was .91. This agrees with findings derived from entirely distinct sets of words (Brown & Watson, 1987; Gordon, 1985) and implies that objective printed frequency measures are more reliable measures of the frequency with which readers encounter a word in print. Furthermore, the correlation between our subjective familiarity norms and those from the MRC database (.65) is greater than any of the correlations of our familiarity norms with objective written frequency measures.

There is some support for the possibility that subjective familiarity is a measure of the spoken frequency of the word. The correlation between our familiarity norms and the BNC spoken frequency measure is .72, which is slightly greater than the .67 correlation between the BNC and CELEX (objective) spoken frequency counts. However, the correlation between our familiarity norms and the CELEX spoken frequency measure is only .58; the weaker correlation in this case probably reflects the poorer reliability of the CELEX spoken frequency count, which is based on a much smaller sample of speech than the BNC count. The possibility that familiarity is tapping into spoken rather than written frequency is also supported by multiple-regression analyses in which familiarity is the criterion variable. Log BNC written frequency explains 33% of the variance in the familiarity norms, but adding log BNC spoken frequency to the equation accounts for an additional 20% of variance ($R^2 = .53$). By contrast, when the order of entry of these variables is reversed, log BNC spoken frequency accounts for 51% of variance in familiarity by itself, and the addition of log BNC written frequency explains only an extra 2% of variance. It should be noted that (following Gilhooly & Logie, 1980) the instructions for rating familiarity specified that raters should take into account the frequencies with which they had both seen and heard the word in question and should rate the word on the basis of the higher of the two measures. The majority of individuals probably hear many more words than they read, so it is not surprising that subjective estimates of word familiarity would be biased toward frequency in the spoken rather than the written modality.

A related question concerns whether subjective familiarity ratings tap into anything beyond frequency information. As Balota et al. (2001) noted, familiarity ratings may also reflect variables unrelated to word frequency, such as the meaningfulness of the word (see Toglia & Battig, 1978) or the familiarity of the sublexical spelling–sound correspondence. Brown and Watson (1987) noted that subjective familiarity was strongly correlated with AoA. The correlations in Table 2 show that the Bristol familiarity norms correlate most strongly with spoken frequency, written frequency, and AoA, but correlate relatively weakly with imageability. Together, the two BNC log frequency measures and our AoA norms account for 57.3% of the variance in familiarity ratings ($N = 1{,}526$). The addition of imageability increases this to 59.3%. However, the partial correlation with imageability is $-.22$; that is, once frequency and AoA have been partialed out, there is a negative correlation between imageability and familiarity that is opposite in direction from what might be expected.

A final question regarding familiarity concerns how useful it is as a predictor of the speed of word identification. By itself, familiarity accounts for 28% of the variance in standardized lexical decision latencies from the ELP corpus (Balota et al., 2002) and 10% of the variance in standardized naming latencies. This is not especially surprising, given the very high correlation of familiarity with written frequency, spoken frequency, and AoA. Once these three variables are partialed out (using the BNC frequency counts), the residual variance in subjective familiarity explains only 1.4% of the variance in standardized lexical decision latencies from the ELP corpus, and only 0.4% of the variance in standardized naming latencies. In summary, subjective familiarity appears to be inferior to objective frequency counts as a measure of the frequency with which words are encountered in print, although it may offer a reasonably good measure of the frequency with which words are encountered in speech. To the extent that subjective familiarity taps into something beyond frequency, it appears to offer no advantage over more clearly defined measures such as objective frequency or AoA. On this basis, we are inclined to question the usefulness of this variable in psycholinguistic research. A similar conclusion has recently been drawn by Baayen (2005). Nevertheless, subjective familiarity ratings may be appropriate for use in research in other areas (such as memory

experiments and neuropsychological case studies) when experimenters are seeking to control their stimuli on a very limited set of variables.

Merging the Bristol Norms With the Gilhooly and Logie (1980) Norms

In view of the evidence that the Bristol and G&L databases are quite compatible, and given the fact that we used instructions very similar to those used by G&L, we decided to form a "megadatabase" that merged the Bristol and G&L norms, resulting in norms for a total of 3,394 words.[3] The two databases were merged after applying the linear transformation procedure already described above (i.e., transforming the Bristol norms on the basis of the coefficients of a regression equation predicting these norms on the basis of the G&L norms). Some ratings that were already near the extreme values of the scales (100 and 700) were pushed outside the scale by the linear transformation, so the values for those items were set to the extreme values of 100 or 700, according to the particular case. There were 27 such items for AoA, 6 for imageability, and 1 for familiarity. The mean transformed ratings for AoA (in the 100–700 scale, as well as in years), imageability, and familiarity for all of the Bristol norms can be downloaded from language.psy.bris .ac.uk/bristol_norms.html.

A methodologically interesting aspect of this study is the use of a control set of words to enable a modular approach to the collection of subjective norms. The high reliability and validity of the Bristol norms demonstrate the feasibility of assembling large sets of norms from modules of words, provided that both the instructions for each rating and the pool of participants are similar, and that an adequate quality control process is applied by assessing the interrater reliability of each block for a set of common words, as if each block was a judge evaluating the same items. Future application of this methodology will allow further increases in the size of the database, with the potential goal of providing subjective norms for the entire set of words that are likely to be used in psycholinguistic or memory experiments.

In summary, the present study provides a very large set of norms for variables that are currently relevant to diverse lines of research in the fields of language and memory. We expect that this large norming study, integrated with the G&L norms, will prove a valuable resource in facilitating experimental research in those fields.

REFERENCES

Baayen, R. H. (2005). Data mining at the intersection of psychology and linguistics. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 69-83). Mahwah, NJ: Erlbaum.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1995). The CELEX lexical database, release 2 [CD-ROM]. Philadelphia: Linguistic Data Consortium & University of Pennsylvania.

Balota, D. A. (1990). The role of meaning in word recognition. In D. A. Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 9-32). Hillsdale, NJ: Erlbaum.

Balota, D. A., Cortese, M. J., Hutchison, K. A., Neely, J. H., Nelson, D., Simpson, G. B., & Treiman, R. (2002). *The English Lexicon Project: A Web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords*. Accessed September 30, 2004, at the Washington University Web site, elexicon.wustl.edu.

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

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

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers*, **33**, 73-79.

Brown, G. D. A., & Watson, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition*, **15**, 208-216.

Burnard, L. (1995). *The user's reference guide for the British National Corpus*. Oxford: Oxford University Computing Services.

Carroll, J. B., & White, M. N. (1973). Age-of-acquisition norms for 220 picturable nouns. *Journal of Verbal Learning & Verbal Behavior*, **12**, 563-576.

Coltheart, M. (1981a). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33A**, 497-505.

Coltheart, M. (1981b). *MRC psycholinguistic database user manual: Version 1*. Accessed September 30, 2004, at the Web site of the Council for the Central Laboratory of the Research Councils, www.psych .rl.ac.uk/User_Manual_v1_0.html.

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

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, **37**, 65-70.

Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, **37**, 665-671.

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, **36**, 488-499.

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

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

Gilhooly, K. J., & Gilhooly, M. L. (1980). The validity of age-of-acquisition ratings. *British Journal of Psychology*, **71**, 105-110.

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.

Gordon, B. (1985). Subjective frequency and the lexical decision latency function: Implications for mechanisms of lexical access. *Journal of Memory & Language*, **24**, 631-645.

Hirsh, K. W., & Funnell, E. (1995). Those old, familiar things: Age of acquisition, familiarity and lexical access in progressive aphasia. *Journal of Neurolinguistics*, **9**, 23-32.

Hulme, C., Stuart, G., Brown, G. D. A., & Morin, C. (2003). High- and low-frequency words are recalled equally well in alternating lists: Evidence for associative effects in serial recall. *Journal of Memory & Language*, **49**, 500-518.

Juhasz, B., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 1312-1318.

Marcel, A. J., & Patterson, K. (1978). Word recognition and production: Reciprocity in clinical and normal studies. In J. Requin (Ed.), *Attention and performance VII* (pp. 209-226). Hillsdale, NJ: Erlbaum.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, **111**, 172-175.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal of Experimental Psychology*, **50A**, 528-559.

Morrison, C. M., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology*, **91**, 167-180.

Morrison, C. M., Ellis, A. W., & Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition*, **20**, 705-714.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart & Winston.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monographs*, **76**(1, Pt. 2), 1-25.

Richardson, J. T. E. (1975). The effect of word imageability in acquired dyslexia. *Neuropsychologia*, **13**, 281-288.

Roodenrys, S., Hulme, C., Alban, J., Ellis, A. W., & Brown, G. D. A. (1994). Effects of word frequency and age of acquisition on short-term memory span. *Memory & Cognition*, **22**, 695-701.

Stadthagen-Gonzalez, H., Bowers, J. S., & Damian, M. F. (2004). Age-of-acquisition effects in visual word recognition: Evidence from expert vocabularies. *Cognition*, **93**, B11-B26.

Toglia, M. P., & Battig, W. R. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Erlbaum.

Whaley, C. P. (1978). Word–nonword classification time. *Journal of Verbal Learning & Verbal Behavior*, **17**, 143-154.

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

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

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

NOTES

1. Twenty-three items included in the Bristol norms did not have entries in the Balota et al. (2002) database, mainly because of differences in spelling between the British and American dialects (e.g., *favour* vs. *favor*).

2. As an anonymous reviewer pointed out, though, the correlations between objective and subjective measures of word frequency could be slightly diluted because of the difference in ranges for these two types of measures.

3. Other large norming studies were also considered as candidates to be merged into this megadatabase, but for a variety of reasons only G&L was deemed compatible enough with the Bristol norms. For example, the control words in the Bristol norms were highly correlated with the norms for AoA and imageability collected by Bird et al. (2001). However, the correlations between the Bird et al. ratings and the ELP latencies for lexical decision and naming (Balota et al., 2002) were relatively poor (for lexical decision, AoA $r = .12$, IMG $r = .00$; for naming, AoA $r = .03$, IMG $r = .06$). This result may be attributable to two characteristics of the Bird et al. norms. First, raters in that study were considerably older than those who participated for either the Bristol or the G&L norms, as well as than the participants in the lexical decision and naming tasks reported by Balota et al. (2002). Second, words in the Bird et al. norming study were unambiguously presented as verbs or nouns, a disambiguation that is not possible in single-word recognition experiments.

APPENDIX

Instructions for Age of Acquisition Ratings

Please indicate (in years) the age at which you learned each of the words on the list. An approximate age is good enough for this rating. If you do not know the meaning of a word, just write an *X* on that space.

By "learning a word" we mean the age at which you would have understood that word if somebody had used it in front of you, EVEN IF YOU DID NOT use, read or write it at the time.

Instructions for Imageability Ratings (adapted from Paivio, Yuille, & Madigan, 1968, p. 4)

Words differ in their capacity to arouse mental images of things and events. Some words arouse a sensory experience, such as a mental picture or sound, very quickly and easily, whereas others may do so only with difficulty (i.e., after a long delay) or not at all. The purpose of this experiment is to rate a list of words as to the ease or difficulty with which they arouse mental images. Any word which, in your estimation, arouses a mental image (i.e., a mental picture, or sound, or other sensory experience) very quickly and easily should be given a *high imagery* rating; any word that arouses a mental image with difficulty or not at all should be given a *low imagery* rating. Think of the words "apple" or "fact." "Apple" would probably arouse an image relatively easily and would be rated as high imagery; "fact" would probably do so with difficulty and would be rated as low imagery. Since words tend to make you think of other words as associates, e.g., knife–fork, it is important that you note only the ease of getting a mental image of an object or an event to the word.

Your ratings will be made on a seven-point scale, where *one* is the low imagery end of the scale and *seven* is the high imagery end of the scale. Make your rating by typing a number from 1 to 7 that best indicates your judgment of the ease or difficulty with which the word arouses imagery. The words that arouse mental images most readily for you should be given a rating of 7; words that arouse images with the greatest difficulty or not at all should be rated 1; words that are intermediate in ease or difficulty of imagery, of course, should be rated appropriately between the two extremes. *Feel free to use the entire range of numbers*, from 1 to 7; at the same time, don't be concerned about how often you use a particular number *as long as it is your true judgment.* Work fairly quickly but do not be careless in your ratings.

If necessary, refer back to these instructions when rating the words on the following pages. If there are any questions, ask them now. Otherwise, turn the page and begin.

Instructions for Familiarity Ratings (adapted from Gilhooly & Logie, 1980)

This is an experiment to find out how often you have come in contact with certain words. You will be given a list of words and you are to rate each one as to the number of times that you experienced it by simply writing down a number according to a 1 to 7 scale. In this scale, 1 represents "NEVER," that is, you have *never* seen or heard or used the word in your life; the number 2 represents "RARELY," that is you have seen or heard or used the word at least once before, but only *rarely*; and so on until 7, which represents "VERY OFTEN," that is, you have seen or heard or used the word *nearly every day* of your life.

Do not be bothered if you are unable to give a definition of some of the words. Simply rate each one as to the number of times you have come in contact with it regardless of its meaning.

There may be some words which you have *used* or *heard* more often than you have *seen* them. Or there may be other words which you have *seen* more often than you have *used* or *heard* them. In such cases, always give the word in the *highest* rating of the three. For example, you probably use or hear the word "cheers" often, but you may never have seen it in print. In this case, you would rate "cheers" as "OFTEN" and write down the number 6.

When the experimenter tells you to start, go to the list of words and begin rating them at your own speed. This is not a "speed" experiment, each participant will be given plenty of time to finish. On the other hand, do not spend too much time on each word. The important thing is for you to be as accurate as possible.

Be as honest in your ratings as you can. Many of the words in this experiment are very rare, so you are not expected to have come in contact with all of them. Just make the best estimates you are capable of.

(Manuscript received December 9, 2004;
revision accepted for publication July 28, 2005.)