Behav Res (2012) 44:978–990
DOI 10.3758/s13428-012-0210-4

Age-of-acquisition ratings for 30,000 English words

Victor Kuperman · Hans Stadthagen-Gonzalez · Marc Brysbaert

Published online: 12 May 2012

Abstract We present age-of-acquisition (AoA) ratings for 30,121 English content words (nouns, verbs, and adjectives). For data collection, this megastudy used the Web-based crowdsourcing technology offered by the Amazon Mechanical Turk. Our data indicate that the ratings collected in this way are as valid and reliable as those collected in laboratory conditions (the correlation between our ratings and those collected in the lab from U.S. students reached .93 for a subsample of 2,500 monosyllabic words). We also show that our AoA ratings explain a substantial percentage of the variance in the lexical-decision data of the English Lexicon Project, over and above the effects of log frequency, word length, and similarity to other words. This is true not only for the lemmas used in our rating study, but also for their inflected forms. We further discuss the relationships of AoA with other predictors of word recognition and illustrate the utility of AoA ratings for research on vocabulary growth.

Keywords Word recognition · Age of acquisition · Ratings · Amazon Mechanical Turk

Researchers using words as stimulus materials typically control or manipulate their stimuli on a number of variables. The four that are most commonly used are word frequency, word length, similarity to other words, and word onset. In this article, we will argue that age of acquisition (AoA) should be part of this list, and we provide ratings for a substantial number of words in order to do so. First, however, we will discuss the evidence in favor of the big four.

Word frequency is the most influential variable to take into account, especially when lexical decision is the task in question (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Ferrand et al., 2011). If the frequency measure comes from an adequate corpus, the percentage of variance explained by this variable in lexical-decision times can easily exceed 30 % (Brysbaert & New, 2009; Ferrand et al., 2010; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2012).

Word length—measured either in characters or in syllables—is an important variable in word naming and progressive demasking (Ferrand et al., 2011), and also in lexical decision. In general, word processing time increases the more letters that a word contains, although in lexical decision the effect seems to be curvilinear rather than linear, as it is not observed for short words (Ferrand et al., 2010; New, Ferrand, Pallier, & Brysbaert, 2006). Additional syllables induce a processing cost as well (Ferrand et al., 2011; Fitzsimmons & Drieghe, 2011; New et al., 2006).

The similarity of a word to other words has traditionally been measured with bigram frequency or Coltheart's *N*. *Bigram frequency* refers to the average frequency of the letter pairs in the word. Coltheart's *N* refers to the number of words that can be formed by changing one letter in the word. These are so-called *word neighbors* (e.g., "dark," "lurk," and "lard" are neighbors of the word "lark"). Yarkoni, Balota, and Yap (2008), however, introduced a measure, OLD20, that captures more variance in both lexical-decision times (Ferrand et al., 2011; Ferrand et al., 2010) and naming latencies (Yarkoni et

V. Kuperman (✉)
Department of Linguistics and Languages, McMaster University,
Togo Salmon Hall 626, 1280 Main Street West,
Hamilton, Ontario, Canada L8S 4 M2
e-mail: vickup@mcmaster.ca

H. Stadthagen-Gonzalez
Bangor University,
Bangor, North Wales, UK

M. Brysbaert
Ghent University,
Ghent, Belgium

al., 2008). OLD20 is a measure of orthographic similarity calculated as the minimum number of letter changes needed to transform the target word into 20 other words. For instance, an OLD20 value of 1 means that 20 words can be formed from the target word by either adding, deleting, or changing one of the word's letters.

Finally, the quality of the first phoneme of a word, or its place or manner of articulation, is the most influential variable in word naming (Balota et al., 2004; Yap & Balota, 2009) and auditory lexical decision (Yap & Brysbaert, 2009). The first letter(s) also play an important role in progressive demasking (Ferrand et al., 2011).

Brysbaert et al. (2011) ran a stepwise regression analysis on the lexical-decision times of the English Lexicon Project (Balota et al., 2007). In this project, lexical-decision and word-naming times were collected for over 40,000 English words. In addition, they made available information about 20 word variables, including:

- Frequency
- Orthographic length of the word (number of letters)
- Number of orthographic, phonological, and phonographic neighbors (i.e., the number of words that differ in one letter or phoneme from the target word, either with or without the exclusion of homophones), both unweighted and weighted for word frequency
- Orthographic and phonological distance to the 20 closest words (OLD20 and PLD20)
- The mean and sum of the bigram frequencies (i.e., the number of words containing the letter pairs within the target word), either based on the total number of words or limited to the syntactic class of the target word
- Number of phonemes and syllables of the word
- Number of morphemes in the word

When all variables were entered in Brysbaert et al.'s (2011) stepwise multiple regression analysis, the most important variable for predicting lexical-decision time was word frequency, accounting for 40.5 % of the variance. The second most important variable was OLD20, which accounted for an additional 12.9 % of the variance. The unique contribution of the third variable, the number of syllables, dropped to 1.2 %, and the summed contributions of the remaining variables amounted to a mere 2.0 % (Brysbaert et al., 2011). Other authors have also reported that the percentage of variance explained by new variables is usually less than 1 % once the big four are partialed out (e.g., Baayen, Feldman and Schreuder 2006; Juhasz, Yap, Dicke, Taylor, & Gullick, 2011).

A promising variable to add to the big four is AoA, or the age at which a word was learned (for reviews, see Brysbaert & Ghyselinck, 2006; Ghyselinck, Lewis, & Brysbaert, 2004; Johnston & Barry, 2006; Juhasz, 2005). Several studies have attested to the importance of this variable. For instance, Brysbaert and Cortese (2011) reported that it explained up to 5 % more variance in the lexical-decision times of English monosyllabic words, in addition to the best word frequency measure available (see also Juhasz et al., 2011). A similar conclusion was reached by Ferrand et al. (2011) for monosyllabic words in French.

Two reasons have been proposed for the importance of AoA in word recognition. The first is that word frequency measures as currently collected do not fully match the cumulative frequency with which participants have been exposed to words (Bonin, Barry, Méot, & Chalard, 2004; Zevin & Seidenberg, 2002; but see Pérez, 2007). Because word frequency estimates are mostly based on materials produced for adult readers, they underestimate the frequency of words typically used in childhood. The second reason for an important contribution of AoA is that the order in which words are learned influences the speed with which their representations can be activated, independently of the total number of times that they have been encountered. Words learned first are easier to access than are words learned later (Izura et al., 2011; Monaghan & Ellis, 2010; Stadthagen-Gonzalez, Bowers, & Damian, 2004), possibly because their meaning is more accessible (Brysbaert, Van Wijnendaele, & De Deyne, 2000; Sailor, Zimmerman, & Sanders, 2011; Steyvers & Tenenbaum, 2005).

Unfortunately, in many experiments AoA cannot be controlled, because the measure only exists for a small percentage of words. AoA estimates are typically obtained by asking a group of participants to indicate the age at which they learned various words. Because gathering such ratings is time-consuming, they are limited in number relative to the total possible range of stimuli. A major step forward was realized in English when Cortese and Khanna (2008) published AoA ratings for 3,000 monosyllabic words, making it possible to include the variable in most subsequent analyses of these words (e.g., Brysbaert & Cortese, 2011; Juhasz et al., 2011). A similar investment was made in French (see Ferrand et al., 2011).

Still, three thousand words is a limited number if one aims to analyze the data of megastudies such as the English Lexicon Project (40 thousand words; Balota et al., 2007) or the British Lexicon Project (28 thousand mono- and disyllabic words; Keuleers et al., 2012). The number of available AoA ratings in English can be doubled to 6,000 if the ratings of Cortese and Khanna (2008) are combined with those of Gilhooly and Logie (1980a, 1980b), Bird, Franklin, and Howard (2001), and Stadthagen-Gonzalez and Davis (2006). However, this still imposes serious constraints on stimulus selection for typical experiments.

Recent developments in the techniques of linguistic data collection may alleviate the situation, however. In particular, the crowdsourcing technology of the Amazon Mechanical Turk as an Internet marketplace has provided language researchers with an attractive new tool. Amazon Mechanical

Turk (https://www.mturk.com/mturk/welcome) is a Web-based service through which a pool of anonymous Web surfers can earn money by completing tasks supplied by researchers. One of these types of task is a questionnaire, which enables the fast and cheap collection of subjective ratings, including norms of the properties of words. Basic demographics, statistics, and best practices for the use of the Amazon Mechanical Turk have been recently reviewed by Mason and Suri (2012). Also, the last few years have seen a proliferation of studies addressing the validity of Amazon Mechanical Turk data as compared to laboratory data, as well as the procedures that need to be followed for ensuring good data quality (Gibson, Piantadosi, & Fedorenko, 2011; Mason & Suri, 2012; Munro et al., 2010; Schnoebelen & Kuperman, 2010; Snow, O'Connor, Jurafsky, & Ng, 2008; Sprouse, 2011). In the vast majority of studies and across tasks, Web-collected data were judged to be indistinguishable in quality from lab-collected data, and were preferable in practical terms (but see Barenboym, Wurm & Cano 2010; Wurm & Cano, 2010, for significant differences between data collected via other Internet services and lab studies). Below, we investigate whether the same is true for the large-scale collection of AoA ratings.

Method

Stimuli

From a list of English words that one of the authors (M.B.) is currently compiling, we selected all of the base words (lemmas) that are used most frequently as nouns, verbs, or adjectives. This became possible after we parsed the SUBTLEX-US corpus (Brysbaert, New, & Keuleers, in press), so that for all words we had information about the frequencies of the different syntactic roles taken by the words. For instance, the word "appalled" was included in the list because it occurred 49 times as an adjective in the corpus, versus 10 times as a verb form. In contrast, the word "played" was not included, because it was used much more often as an inflected verb form than as an adjective (2,843 times vs. 26). The selection resulted in a total of 30,121 words. No further restrictions (e.g., number of letters or syllables, or frequency thresholds) were placed on the words.

Data collection

The stimuli were distributed over lists of 300 target words each, roughly matched on word frequency (using the SUBTLEX-US frequency norms of Brysbaert & New, 2009). The matching was achieved by dividing the total word list into 10 equally sized frequency bins and selecting 30 words from each bin per stimulus list. In order to further improve the validity of the ratings, we introduced "calibrator" and "control" words to each of the stimulus lists. Each list was preceded by ten calibrator words representing the entire range of the AoA scale, based on the Bristol ratings.[1] In this way, the participants were exposed to the diversity of words that they were likely to encounter. A further 52 control words covering the entire AoA range were randomly distributed over the word lists. The AoA distribution of these control words was roughly normal, and it reflected the distribution of ratings in the Bristol norms, with fewer very early and very late words and more words toward the middle of the scale.

We used the same instructions used in the collection of the Bristol norms (Stadthagen-Gonzalez & Davis, 2006). For each word, the participants were asked to enter the age (in years) at which they thought they had learned the word. It was specified that 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." Unlike many other studies, we did not ask participants to use a 7-point Likert rating scale, because this artificially restricts the response range and is also more difficult for participants to use (see Ghyselinck, De Moor, & Brysbaert, 2000, for a comparison of both methods; also see Fig. 3 below). When participants did not know a word, they were asked to enter the letter "x." This prevented us from collecting wrong AoA ratings and also provided us with an estimate of how familiar responders were with the words. A complete list of 362 words (300 test words, 10 calibrator words, and 52 control words) took about 20 min to complete. Participants were paid half a U.S. cent per rated word (i.e., $1.81 for a validly completed list).

Responders were limited to those residing in the U.S., but no further restrictions were imposed (e.g., no requirement of English being the first language or only language of the responder). The participants were asked also to report their age, their gender, their first language or languages, which country/state they had lived in the longest between birth and the age of 7, and which educational level describes them best: some high school; high school graduate; some college, no degree; associate degree; bachelors degree; or masters degree or doctorate.

The lists were initially presented to 20 participants each. Because of values missing as a result of the exclusion criteria and data trimming discussed below, some of the words had less than 18 valid observations after this phase.

[1] The calibrator words were, along with their AoA ratings (in years) according to the Bristol norms: "shoe" 3.3, "knife" 4.5, "honest" 5.5, "arch" 6.5, "insane" 7.6, "feline" 8.5, "obscure" 9.5, "nucleus" 10.5, "deluge" 11.4, and "hernia" 12.6.

These words were recombined in new, comparable lists at the end of the data collection and were presented to new participants until the required number of observations was reached for all words.

All in all, a total of 842,438 ratings were collected from 1,960 responders over a period of 6 weeks (153 of the responders contributed responses to more than one list). The total cost of using the Amazon Mechanical Turk for this megastudy was slightly less than $4,000.

Results

Data trimming

About 7 % of the responses were empty cells, which were removed. Valid responses were defined as either a numeric AoA rating that was less than the responder's age or an "x" response, which signified a "Don't know" answer. AoA ratings that were equal to the responder's age were relabeled as "Don't know" responses (less than 0.5 % of all responses). About 1 % of the nonempty responses were removed as they did not match our definition of a valid response or exceeded the responder's age. The participants were instructed that a lower-boundary correlation with control words in the list was required in order for them to earn payment for the completed list. This discouraged participants from simply entering random numbers in order to receive easy payment (a similar precaution is taken in laboratory studies, where participants are excluded if their ratings do not correlate with the ratings from the other participants; e.g., Ghyselinck et al., 2000). Participants were paid if they provided valid numeric ratings to 30 or more of the 52 control words and if those ratings correlated at least .2 with the Bristol norms.

In the data analysis, we removed all target lists with a correlation of less than .4 with the Bristol norms for the set of control words. This led to the removal of 350 lists or 126,700 ratings (15 % of the collected ratings). Finally, the distribution of AoA ratings had a positive skew. Therefore, we removed another 1 % of extremely large values of AoA ratings (ratings exceeding 25 years of age) to attenuate the disproportionate influence of outliers on statistical models. The resulting data set comprised 696,048 valid ratings, accounting for 83 % of the original data set. Of these, 615,967 were numerical (89 % of the valid ratings), and 76,211 (11 % of the valid ratings) were "Don't knows." The resulting set of responders included 1,729 responders, or 88 % of the original participant pool. Of the words that we included in our study, 2,300 (7.7 %) were not known to half of the respondents. For completeness, this article and the supplementary materials provide mean numeric ratings for *all* of the words; we also base our correlational and regression analyses on the full word list. For experiments with a small number of items, it is advisable, however, to use the mean numeric ratings only if they are reported to be based on at least five numeric responses.

All but eight of the words received 18 or more valid ratings. The correlation between the mean numeric ratings for the control words and the Bristol norms was r = .93 (N = 50, p < .0001). The correlation between the odd-numbered and even-numbered participants for the items with 10 or more numeric ratings (N = 26,532) was r = .843, which gives a very high split-half reliability estimate of 2 × .843/(1 + .843) = .915.

Some previous studies collecting AoA norms in blocks of words (e.g., Bird et al., 2001; Stadthagen-Gonzalez & Davis, 2006) used a linear transformation procedure to homogenize the means and standard deviations of the blocks (for details, see p. 600 of Stadthagen-Gonzalez & Davis, 2006). We applied this procedure to a random sample of five of our lists and found that the differences between the raw and the corrected ratings were negligible (usually less than 0.2). Therefore, we decided not to apply this transformation to our data.

Demographics

Of the valid responders, 1,136 were female and 593 were male. Their ages ranged from 15 to 82 years, with 8 % of the responders being younger than 20 years, 47 % from 20 to 29 years old. 22 % from 30 to 39, 12 % from 40 to 49, and 11 % older than 49. Twelve of the participants (0.7 %) reported a single language other than English as their first language; another 31 responders (1.8 %) reported more than one language as their first languages, including English. As their responses did not differ from the rest, they were included.

Education levels were labeled as follows: 1, *Declined to answer or No high school*; 2, *High school graduate*; 3, *Some college, no degree*; 4, *Associate degree*; 5, *Bachelors degree*; and 6, *Master or higher degree*. Table 1 shows the distribution of ratings and responders over the various categories. Most of the participants came from categories 3 (some college) and 5 (bachelor's degree)

Table 1 Education levels of the responders

Education Level	Percentage of Ratings
Declined to answer or No high school	6
High school graduate	12
Some college, no degree	35
Associate degree	10
Bachelors degree	27
Master or higher degree	10

Does demography affect the numeric ratings?

Women gave slightly but significantly higher AoA numeric ratings ($M = 10.2$, $SD = 4.4$) than did men ($M = 10.1$, $SD = 4.2$) ($t = -10.27$, $df = 440{,}410$, p value $< .0001$). The numeric AoA ratings did not vary by the education levels of the responders, as is shown in the box plots of the AoA ratings in Fig. 1. This null effect in subjective judgments of AoA is surprising, given the wealth of developmental literature showing that early advantages in vocabulary size (e.g., larger numbers of word types learned earlier) are excellent predictors of future educational achievements (e.g., Biemiller & Slonim, 2001).

AoA correlated strongly with word frequency, and the relationship was log-linear (see below). To test whether this association was affected by education level, we divided education into low (Levels 1–3, up to and excluding the associate college degree) and high (4–6). Figure 2 shows the functional relationships between the AoA ratings and $\log_{10}$ SUBTLEX frequency for both groups. There is a hint of an interaction (which is significant at $p < .05$, due to the very high number of observations), but the size of the effect is very small. Higher-educated individuals tended to give earlier AoAs for high-frequency words and later AoAs for low-frequency words than did lower-educated individuals; both differences were well within 0.2 year.

Finally, there was a weak positive correlation between AoA ratings and the age of the participants ($r = .07$; $t = 61.00$, $df = 615{,}965$, $p < .0001$). On average, older participants gave higher AoA ratings than did younger participants, presumably because they had a broader age range to choose from.

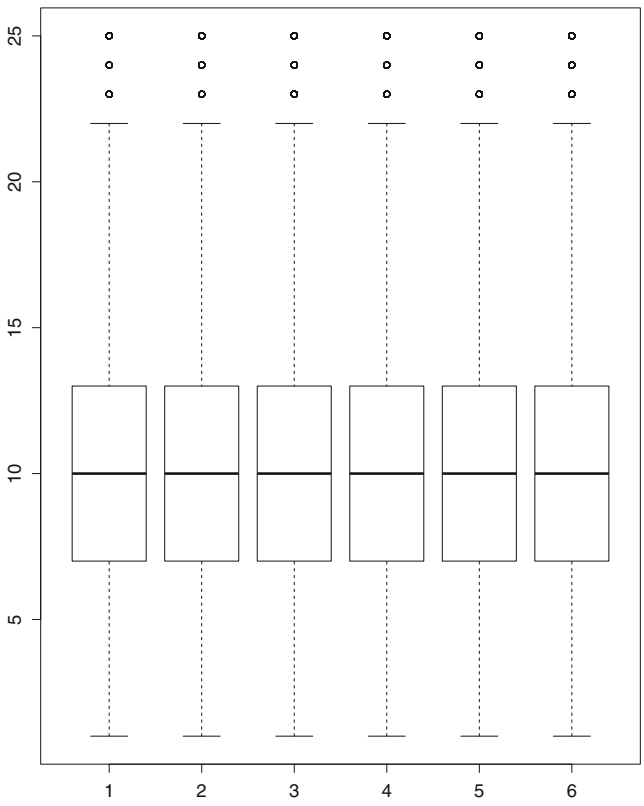

Fig. 1 Age-of-acquisition (AoA) ratings as a function of education level

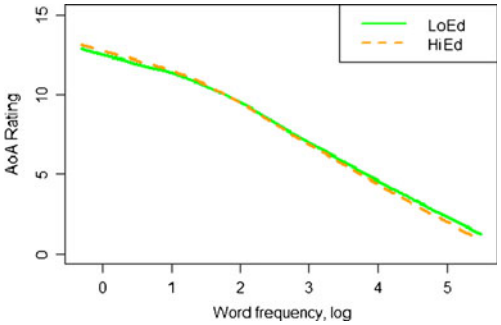

Fig. 2 Association between age of acquisition (AoA) and log word frequency as a function of education level. LoEd comprises Education Levels 1–3 (808 responders), and HiEd comprises Education Levels 4–6 (686 responders)

Does demography affect the number of "don't knows"?

For each word, we computed the ratio of numerical responses to total responses as an index of the responders' familiarity with the word. The ratio correlated strongly with the log frequency of the word ($r = .56$; $t = 509.9$, $df = 565{,}587$, $p < .0001$), but no demographic variable was a significant predictor of the ratio. Perhaps most surprisingly, the average percentages of unknown words did not vary by education level, ranging from 12 % for the "no high school" level to 11 % for the "masters or higher" level.

Correlations with other AoA norms

Of course, the most important question is how strongly our Web-collected ratings correlate with those of typical laboratory studies, and whether we jeopardized the quality of the data by using less controlled sources. We can compare our mean ratings with those from three large-scale studies: Cortese and Khanna (2008) collected AoA ratings for 3,000 monosyllabic words from 32 psychology undergraduates from the College of Charleston. Bird et al. (2001) collected ratings for 2,700 words from 45 participants in the U.K.; most of their participants were between 50 and 80 years of age (mean age 61 years). Finally, Stadthagen-Gonzalez and Davis (2006) collected norms for 1,500 words from 100 undergraduate psychology students from Bristol and combined them with the Gilhooly and Logie (1980a, 1980b) ratings (collected in Aberdeen) for another 1,900 words.

Our data set had 2,544 words in common with that of Cortese and Khanna (2008). The correlation between our ratings and theirs is $r = .93$ (Fig. 3). A total of 1,787 words were shared with Bird et al. (2001), and these ratings

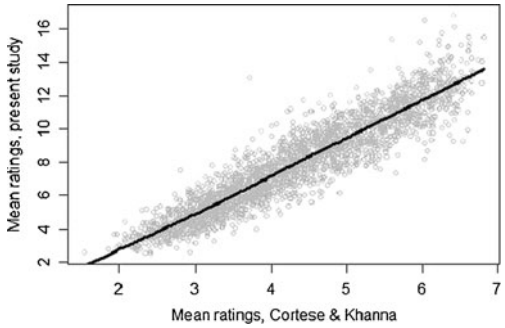

Fig. 3 Age-of-acquisition (AoA) ratings of Cortese and Khanna (2008; collected on the 1–7 Likert scale) plotted against the present AoA ratings, with a solid black loess trend line. $r = .93$, $p < .0001$, based on 2,544 monosyllabic words

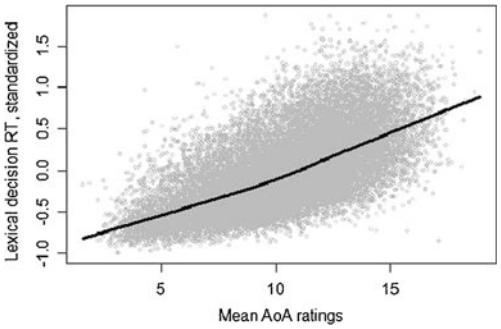

Fig. 4 Standardized English Lexicon Project (ELP) lexical-decision response times plotted against the present AoA ratings, with a solid black trend line. $r = .64$, $p < .0001$, based on 20,302 words

correlated $r = .83$. Finally, 3,117 words were shared with the Bristol norms, which correlated $r = .86$ with our ratings.

On the basis of these correlations, we can safely conclude that our ratings are as valid as those previously collected under more controlled circumstances. Some small differences may be present in the AoA ratings between the U.S. and the U.K., given the higher correlation with the Cortese and Khanna (2008) ratings than with the Bird et al. (2001) and Stadthagen-Gonzalez and Davis (2006) ratings.

Correlation with the lexical decision data of the English Lexicon Project

Further validation of our AoA ratings was obtained by correlating them with the lexical-decision data of the English Lexicon Project (ELP). Our lists have 20,302 words in common with the ELP. For these words, we calculated the correlations with AoA, log frequency, word length in number of letters and syllables, Coltheart's *N*, and OLD20 (values were from the ELP website). Because the correlations are higher with standardized response times (*z*RTs) than with raw response times (Brysbaert & New, 2009), we used the former behavioral measure. Table 2 summarizes the results.

As can be seen in Table 2, AoA has the second highest correlation with *z*RT (after log frequency) and the highest correlation with the percentage of correct responses. Surprisingly, the relationship of the mean AoA ratings with lexical-decision times was completely linear, with an estimated 27-ms increase in response time per increase by 1 year of AoA (see Fig. 4).

Table 2 Correlations between word characteristics and the standardized response times (*z*RTs) and accuracy levels in the lexical-decision task of the English Lexicon Project ($N = 20{,}302$ lemmas)

	*z*RT	Accuracy
AoA	.637	–.507
Log frequency (SUBTLEX)	–.685	.464
Nletters	.554	.041
Nsyllables	.537	.021
Coltheart's *N*	–.347	.069
OLD20	.600	–.082

The importance of the AoA variable becomes yet more clear in stepwise multiple regression analyses. In these analyses, we took into account the finding that the effects of log frequency and word length on lexical-decision outcome variables are nonlinear, by using restricted cubic splines for these variables. Of the many analyses that we ran (and which can easily be replicated by any interested reader, as all of the values are freely available), we list below the ones that highlight the predictive power of AoA. For their interpretation, it is important to realize that R^2 differences of even .01 typically (and in the present analyses) come with *p* values below the conventional thresholds of significance (because of the large numbers of observations).

R^2 values for regressions on *z*RT:

$$\begin{aligned} &\text{Freq} + \text{AoA}: && R^2 = .549 \\ &\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20}: && R^2 = .615 \\ &\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20} + \text{AoA}: && R^2 = .653 \end{aligned}$$

R^2 values for regressions on accuracy:

$$\begin{aligned} &\text{Freq} + \text{AoA}: && R^2 = .318 \\ &\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20}: && R^2 = .335 \\ &\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20} + \text{AoA}: && R^2 = .433 \end{aligned}$$

AoA explains an extra 4 % of variance in *z*RTs after log word frequency (Freq), word length (in letters [Nlett], and syllables [Nsyl]), and similarity to other words (OLD20) are controlled for. For the accuracy data, the extra variance explained by AoA approaches 10 %. Relative to the influence of other variables (which usually explain less than 1 % additional variance; see the introduction), these are substantial effects.

Are AoA ratings also predictive of inflected word forms?

Having access to AoA ratings of 30,000 lemmas is beneficial in itself, as this is a tenfold increase in the existing pool

of AoA ratings. However, it would be even more beneficial if the ratings we collected for lemmas could also be used for the lemmas' inflected forms. Given that each base noun has one inflected form (the plural) and that a regular base verb has three inflected forms (3rd person as well as present and past participles), the number of words to which our ratings apply would be considerably higher if the ratings also explained differences in lexical-decision performance to inflected word forms. A total of 10,011 inflected word forms in the ELP are associated with one of the lemmas rated in our study. For the correct interpretation of this finding, it is important to realize that the inflected forms do not include verb forms used more frequently as adjectives (such as "appalled"). These were included in our list of lemmas presented to the participants of the AoA study (see above). Table 3 shows the results for the inflected words.

As Table 3 suggests, there were strong correlations between lexical-decision performance on the inflected forms and the AoAs of the base words. The same was true for the frequencies of the base words (e.g., for the inflected form "played," this would be the frequency of the word "play"). However, because the correlation between the frequency of the inflected form and the frequency of the lemma was higher than the correlation between the frequency of the inflected form and the AoA of the lemma, AoA came out as a better predictor in multiple regression analyses, as can be seen below:

R^2 values for regressions on *z*RT:

$$\begin{aligned}
&\text{Freq} + \text{AoA}: && R^2 = .488\\
&\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20}: && R^2 = .558\\
&\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20} + \text{AoA}: && R^2 = .583\\
&\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20} + \text{Freq_lemma}: && R^2 = .571\\
&\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20} + \text{Freq_lemma} + \text{AoA}: && R^2 = .585
\end{aligned}$$

R^2 values for regressions on accuracy:

$$\begin{aligned}
&\text{Freq} + \text{AoA}: && R^2 = .243\\
&\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20}: && R^2 = .271\\
&\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20} + \text{AoA}: && R^2 = .318\\
&\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20} + \text{Freq_lemma}: && R^2 = .297\\
&\text{Freq} + \text{Nlett} + \text{Nsyl} + \text{OLD20} + \text{Freq_lemma} + \text{AoA}: && R^2 = .322
\end{aligned}$$

Table 3 Correlations between word characteristics and the standardized response times (*z*RTs) and accuracy levels in the lexical-decision task of the English Lexicon Project for inflected word forms ($N = 10{,}011$)

	zRT	Accuracy
AoA lemma	.588	–.369
Log frequency, inflected form	–.629	.421
Log frequency, lemma	–.587	.373
Nletters, inflected form	.524	.053
Nsyllables, inflected form	.505	.003
Coltheart's *N*, inflected form	–.334	.039
OLD20, inflected form	.549	–.035

By controlling inflected word forms on lemma AoA in addition to word frequency, word length, and similarity to other words, one gains 2.5 % explained variance in standardized response times and more than 4.5 % in the percentage-accurate value.

How does AoA relate to other ratings?

Our data also allow us to examine the relationship of AoA to other word variables. Clark and Paivio (2004) ran an analysis of 925 nouns for which they had information about many rated values, in addition to the usual objective measures (frequency, length, and similarity to other words). More specifically, they looked at the impact of 32 variables, including:

- word frequency (Kučera & Francis, Thorndike & Lorge)
- estimated word familiarity (two ratings from different studies)
- word length (in letters and syllables)
- word availability (the number of times a word is given as an associate to another word or is used in dictionary definitions)
- number of meanings a word has
- estimated context availability (how easy participants find it to think of a context in which the word can be used)
- estimated concreteness and imageability (two ratings from different studies)
- estimated AoA and number of childhood dictionaries in which the word is explained
- emotionality, pleasantness, and goodness ratings of the words, and the degree of deviation from the means
- how gender-laden the word is (two ratings from different studies)
- number of high-frequency words starting with the same letters
- subjective estimates of the number of words that begin with the same letters and sounds, rhyme with the words, sound similar, and look similar
- pronounceability ratings of the words
- estimated ease of giving a definition, and estimate of whether a word has different meanings

Factor analysis suggested that the 32 variables formed nine factors: frequency, length, familiarity, imageability, emotionality, word onset, gender-ladenness, pleasantness, and word ambiguity. The last factor was the weakest and on the edge of significance.

To see how the new AoA measure related to the variables investigated by Clark and Paivio (2004), we added three extra variables (log SUBTLEX frequency, our new AoA rating, and OLD20) to the list and looked at the correlations with *z*RT in the ELP lexical-decision task. Values were

present for 896 of the original 925 words. Table 4 lists the correlations in decreasing order of absolute values. This shows that the correlation with *z*RT was strongest for word frequency, followed by the estimated pronounceability of the word, familiarity, word availability, and context availability. The lowest correlations were observed for the estimated similarity of the word to other words, the emotionality, and the gender-ladenness of the words. Also interesting is that our AoA ratings correlated .90 with those of Clark and Paivio, and correlated slightly higher with *z*RTs than did the Clarke and Paivio AoA ratings.

To examine the relationship between our AoA ratings and the many ratings mentioned by Clark and Paivio (2004), we repeated their factor analysis (using the factanal procedure of R with the default varimax rotation). As we had slightly fewer data (896 instead of 925), we failed to observe a significant contribution of the final factor (meaning ambiguity). Therefore, we worked with an eight-factor model instead of the original nine-factor model. We also included the additional variables log SUBTLEX-US frequency, OLD20, and *z*RT in the ELP lexical-decision task. The latter variable allowed us to see on which factors lexical-decision times load and to what extent these differ from those on which the other variables load.

The outcomes of the factor analysis are shown in Table 5. This analysis indicates that lexical-decision times only loaded on the first four factors (word frequency, length, familiarity, and imageability). They were

Table 4 Correlations between word characteristics and the standardized response times in the lexical-decision task of the English Lexicon Project for the words listed in Clark and Paivio (2004; $N = 896$, ordered from high to low)

Log SUBTLEX-US frequency	–.757**
Estimated ease of pronunciation	–.735**
Familiarity Rating 1	–.727**
Familiarity Rating 2	–.724**
Log Thorndike–Lorge frequency	–.714**
Word availability (number of times word was produced as associate)	–.711**
Estimated ease to produce context	–.691**
AoA rating (present study)	.690**
AoA rating (Paivio)	.657**
Log Kučera–Francis frequency	–.640**
Word availability (times word is used in dictionary definitions)	–.625**
Estimated ease of defining the word	–.615**
Log number of childhood dictionaries in which word occurs	–.595**
Imageability Rating 1	–.582**
OLD20	.577**
Length in letters	.549**
Length in syllables	.528**
Estimated number of similar-sounding words	–.515**
Estimated number of associates of the word	–.465**
Estimated number of similar-looking words	–.442**
Estimated number of rhyming words	–.427**
Meaningfulness (number of associates produced in 30 s)	–.424**
Imageability rating	–.328**
Estimated number of meanings of the word (ambiguity)	–.287**
Pleasantness rating	–.266**
Emotionality rating	–.217**
Estimated number of words that start with the same sounds	–.201**
Estimated goodness/badness of the word's meaning	–.176**
Concreteness rating	–.166**
Deviation of emotionality rating from the mean rating	–.122**
Deviation of goodness rating from the mean rating	–.071**
Estimated number of words starting with the same letters	–.064*
Gender-ladenness Rating 1	–.027
Gender-ladenness Rating 2	–.017
Log number of high-frequency words starting with the same two letters	.008

*$p < .05$, **$p < .01$

Table 5 Factor loadings of the different variables in Clark and Paivio's (2004) study and of four new variables on the words for which we had all of the data (N = 896)

	Freq.	Len.	Fam.	Ima.	EmoDev.	Gender	Onset	Pleasant
*z*RT ELP lexical-decision task	**–.522**	**–.428**	**–.526**	–.138				
SUBTLEX-US frequency	**.739**	.284	**.394**	.127	.178			
Estimated ease of pronunciation	**.388**	**.361**	**.623**	.138				.107
Familiarity Rating 1	**.615**	.131	**.627**	.140		.125		.104
Familiarity Rating 2	**.371**		**.876**	.112	.111		.117	
Thorndike–Lorge frequency	**.795**	.257	.285	.171				.129
Word availability (produced as associate)	**.706**	**.381**	.266	.293	.118			
Estimated ease to produce context	.298	.104	**.842**	.285	.141			
AoA rating (present study)	**–.432**	**–.315**	**–.496**	**–.467**				
AoA rating (Paivio)	**–.421**	**–.326**	**–.445**	**–.513**		–.108		–.117
Kučera–Francis frequency	**.824**	.112	**.305**				.113	.121
Word availability (used in dictionary)	**.778**	**.312**	.143					
Estimated ease of defining the word	.267		**.729**	**.424**				
Number of childhood dictionaries	**.593**	.283	.238	**.489**				.106
Imageability Rating 1	.197	.184	**.543**	**.715**	.119			
OLD20	–.259	**–.851**		–.104				
Length in letters	–.256	**–.793**		–.186			.273	
Length in syllables	–.189	**–.755**		–.251			.103	
Similar-sounding words (estimation)	.185	**.846**	.145	.102			.154	
Associates to the word (estimation)	**.419**		**.386**		**.381**			.127
Similar-looking words (estimation)	.155	**.700**	.199				.251	
Rhyming words (estimation)	.120	**.762**	.144				.233	
Meaningfulness (number of associates)	.200	.155	.295	**.651**				
Imageability Rating 2		.174	.187	**.908**				
Meanings of the word (estimation)	.249	.197	.183	**–.306**	.228			.101
Pleasantness rating	.205		.151		.125	.229		**.928**
Emotionality rating	.143		.204	–.150	**.799**			.108
Start with the same sounds (estimate)	.104	.200	.160				**.726**	
Goodness/badness of meaning	.174					.240		**.864**
Concreteness rating		.149		**.863**	–.287			
Deviation of emotionality from mean					**.838**			
Deviation of goodness from mean					**.900**			
Start with same letters (estimation)							**.785**	
Gender-ladenness Rating 1						**.964**		.184
Gender-ladenness Rating 2						**.940**		.231
High-frequency words starting with same letters							**.658**	
SS loadings	5.443	4.956	4.782	3.962	2.582	1.966	1.894	1.870
Proportion of variance	.151	.138	.133	.110	.072	.055	.053	.052
Cumulative variance	.151	.289	.422	.532	.603	.658	.711	.763

Lexical-decision times load on four factors only, and word frequency and AoA load on the same variables. In factor analysis, loadings higher than .3 are considered important, and these are given in bold. The variables are ordered as in Table 4

not significantly related to the emotionality, word onset, gender-ladenness, or pleasantness of the words. Interestingly, AoA loaded on exactly the same factors, just as word frequency did. This is further evidence that AoA and word frequency are strongly related to lexical-decision times. For the Clark and Paivio (2004) set of nouns, we also see a strong influence of familiarity that is surprising, given that in two previous analyses on monosyllabic words, familiarity no longer seemed to have a strong influence, if a good frequency measure and an AoA measure were used (Brysbaert & Cortese, 2011; Ferrand et al., 2011).

AoA ratings and vocabulary growth

The availability of AoA ratings for a large number of content words also makes it possible to estimate the number of words thought to be learned at various ages—that is, the guesstimated vocabulary growth curve. We divided the mean AoA ratings into yearly bins, from 1 to 17, and computed the cumulative sum of the word types falling in each bin. This subjective estimate of vocabulary growth is compared in Fig. 5 to the estimates obtained via experimental testing of children's vocabulary in Biemiller and Slonim (2001). Biemiller and Slonim presented both a representative sample and a sample with an advantaged socioeconomic status with multiple-choice questions requiring definitions of words from a broad frequency range. They tested children from Grades 1, 2, 4, and 5 and estimated the number of words acquired from infancy to Grade 5 (see Tables 10 and 11 in Biemiller & Slonim, 2001). We relabeled Grades 1–5 to ages 6–10, respectively.

Figure 5 shows the subjectively estimated vocabulary growth curve on the basis of the AoA ratings (solid line). As can be seen, this is a sigmoid curve typical of learning tasks. Figure 5 further includes the estimates of vocabulary size both for the representative (or normed) sample (thick dashed line) and the group with an advantaged socioeconomic status (thick dotted line), as reported by Biemiller and Slonim (2001). For each group, we also included confidence intervals (based on the estimated numbers of lemmas known to the 0–25th and 75th–100th percentiles of the groups).

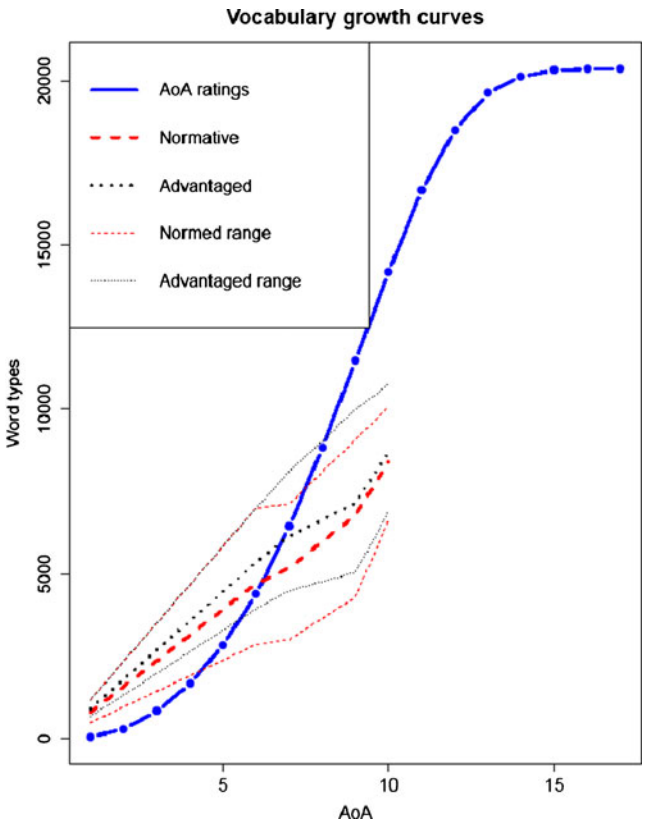

Fig. 5 Number of lemma types estimated from the age-of-acquisition (AoA) ratings (solid line), and reported for the normative and advantaged samples of elementary school students (Biemiller & Slonim, 2001)

Several aspects of the comparison between the estimated and measured vocabulary growth are noteworthy. First, our responders put the main weight of word learning in the elementary school years, from ages 6 to 12. This underestimates growth in the years 2–5 (the AoA estimates are lower than those in Biemiller & Slonim, 2001) and overestimates the growth after the age of 9 (AoA estimates are higher than those in Biemiller & Slonim, 2001). Also, responders reported that hardly any words entered their vocabulary before the age of 3 and after the age of 14. Only a small percentage (1.2 %) of the mean AoA ratings were below 4 years of age, even though the receptive vocabulary is not negligible in these age cohorts. This result is in line with the well-described phenomenon of infantile amnesia, the inability of adults to retrieve episodic memory (including lexical memory) before a certain age (de Boysson-Bardies & Vihman, 1991). Reporting only a small percentage of words acquired after the age of 15 (3 %–5 %) was true even for a more educated population (bachelors, masters, or PhD degree) that is likely to have substantially broadened their vocabulary throughout the higher education years.

Discussion

In this article, we have described the collection of AoA ratings for 30,000 English content words (nouns, verbs, and adjectives) with the Amazon Mechanical Turk. Several interesting findings were observed. First, the Web-based ratings correlated highly with previous ratings collected under more controlled circumstances. For various samples, the correlations varied between $r = .83$ and .93. In particular, the correlations with previously collected American ratings were high (Clark & Paivio, 2004; Cortese & Khanna, 2008; see Fig. 3). This means that the Internet crowdsourcing technology forms a useful tool for the rapid gathering of large numbers of word characteristics (nearly 2,000 participants in 6 weeks) if some elementary precautions are taken. In particular, we found it necessary to limit the respondents to those living in the U.S. (or to English-speaking countries more in general) and to have some online control of the quality of the data. This was done by inserting a limited number of stimuli with known values across the entire range and checking whether the ratings provided by the respondents for these stimuli correlated with the values already available. In this way, the quality of the data was controlled. With these checks in place, we were able to collect a large amount of useful data in a short period of time and at a low

price. This opens perspectives for research on other variables.

Second, we confirmed that AoA is an important variable to control in word recognition experiments. In the various analyses that we ran, AoA always had a high correlation with the dependent variable (in particular, with lexical-decision time), and it explained 2 %–10 % of variance in addition to word frequency, word length (both number of letters and number of syllables), and similarity to other words (operationalized as OLD20). AoA also came out well in a comparison with the 32 word features collected by Clark and Paivio (2004), as is shown in Tables 4 and 5. The effect of AoA was found not only for the lemmas included in the rating study (Table 2), but also for inflected forms based on them (Table 3).

The robust additional effect of AoA was expected on the basis of theories of word learning (Izura et al., 2011; Monaghan & Ellis, 2010) and theories of the organization of the semantic system (Brysbaert et al., 2000; Sailor et al., 2011; Steyvers & Tenenbaum, 2005). Researchers have been hampered in the use of this variable because of the scarcity of ratings available. This restriction has now been lifted. Having access to AoA ratings for over 30,000 content lemmas and their inflected forms means that researchers can routinely control their stimuli for this variable. Our analyses indicate that this will considerably increase the quality of stimulus matching. The AoA ratings also make it possible to include the variable in future analyses of megastudy data.

The availability of a large number of AoA ratings further makes it possible to analyze the AoA ratings themselves. For instance, a longstanding question has concerned whether AoA ratings are accurate estimates of acquisition times or rather are a reflection of the order of acquisition (see Monaghan & Ellis, 2010 and references therein). Several aspects of our data are in line with the second possibility. First, AoA estimates seem to form a normal distribution with a mode around 9 years of age and 90 % of the data points between 5 and 15 years of age (standard deviation of about 2.84 years). Importantly, this curve deviates from empirically obtained vocabulary growth curves in young responders (Biemiller & Slonim, 2001; see Fig. 5 above) and from what can be expected after the age of 15, given the massive acquisition of new words in higher education. Also, the linear relationship between AoA and lexical-decision times may point in this direction (Fig. 4). Observing a linear effect of a variable may be an indication that the variable is rank-ordered, with the order of values rather than the intervals between values driving the variable's behavioral effect; see a similar argument for ranked word frequency in Murray and Forster (2004). This topic can now be fruitfully studied using experimental and corpus-based methods against a large number of words ranging in frequency, length, and other relevant lexical properties.

Availability

Our AoA ratings are available as supplementary materials for this article. For each word, we report the number of times that it occurs in the trimmed data (OccurTotal). For most words, the count is about 19. However, for the ten calibration words and the 52 control words, this amounts to more than 1,900 presentations. Next, we provide the mean AoA ratings (in years of age) and standard deviations (Rating.Mean and Rating.SD). We also present the number of responders who gave numeric ratings to the word, rather than rated it as unknown (OccurNum). This information is useful, because it helps to avoid using unknown words in psychological experiments and indicates the degree of reliability of the mean AoA ratings. Finally, we add word frequency counts from the 50-million-word SUBTLEX-US corpus (Brysbaert & New, 2009). Words are presented in decreasing order of frequency of occurrence. The 574 words that were not present in the SUBTLEX-US frequency list were assigned the frequency of 0.5.

Author note This study was supported by an Odysseus grant awarded by the Government of Flanders (the Dutch-speaking northern half of Belgium). We thank Michael Cortese, Gregory Francis, and an anonymous reviewer for insightful comments on an earlier draft of this article, and Danielle Moed for her help with preparation of the manuscript.

References

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., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445–459. doi:10.3758/BF03193014

Barenboym, D. A., Wurm, L. H., & Cano, A. (2010). A comparison of stimulus ratings made online and in person: Gender and method effects. *Behavior Research Methods, 42,* 273–285. doi:10.3758/BRM.42.1.273

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology, 93,* 498–520.

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. doi:10.3758/BF03195349

Bonin, P., Barry, C., Méot, A., & Chalard, M. (2004). The influence of age of acquisition in word reading and other tasks: A never ending story? *Journal of Memory and Language, 50,* 456–476. doi:10.1016/j.jml.2004.02.001

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58,* 412–424.

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology, 64,* 545–559.

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition, 13,* 992–1011. doi:10.1080/13506280544000165

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990. doi:10.3758/BRM.41.4.977

Brysbaert, M., New, B., & Keuleers, E. (in press). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*. doi:10.3758/s13428-012-0190-4

Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica, 104,* 215–226. doi:10.1016/S0001-6918(00)00021-4

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers, 36,* 371–383.

Cortese, M. J., & Khanna, M. M. (2008). Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods, 40,* 791–794. doi:10.3758/BRM.40.3.791

de Boysson-Bardies, B., & Vihman, M. M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language, 67,* 297–319.

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., & Pallier, C. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: Evidence from Chronolex. *Frontiers in Psychology, 2,* 1–10.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42,* 488–496. doi:10.3758/BRM.42.2.488

Fitzsimmons, G., & Drieghe, D. (2011). The influence of number of syllables on word skipping during reading. *Psychonomic Bulletin & Review, 18,* 736–741.

Ghyselinck, M., De Moor, W., & Brysbaert, M. (2000). Age-of-acquisition ratings for 2816 Dutch four- and five-letter nouns. *Psychologica Belgica, 40,* 77–98.

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. doi:10.1016/j.actpsy.2003.11.002

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using mechanical turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass, 5,* 509–524.

Gilhooly, K. J., & Logie, R. H. (1980a). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation, 12,* 395–427. doi:10.3758/BF03201693

Gilhooly, K. J., & Logie, R. H. (1980b). Meaning-dependent ratings of imagery, age of acquisition, familiarity, and concreteness for 387 ambiguous words. *Behavior Research Methods & Instrumentation, 12,* 428–450. doi:10.3758/BF03201694

Izura, C., Pérez, M. A., Agallou, E., Wright, V. C., Marín, J., Stadthagen-González, H., & Ellis, A. W. (2011). Age/order of acquisition effects and the cumulative learning of foreign words: A word training study. *Journal of Memory and Language, 64,* 32–58.

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition, 13,* 789–845.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin, 131,* 684–712. doi:10.1037/0033-2909.131.5.684

Juhasz, B. J., Yap, M. J., Dicke, J., Taylor, S. C., & Gullick, M. M. (2011). Tangible words are recognized faster: The grounding of meaning in sensory and perceptual systems. *Quarterly Journal of Experimental Psychology, 64,* 1683–1691. doi:10.1080/17470218.2011.605150

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology, 1,* 174. doi:10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44,* 287–304. doi:10.3758/s13428-011-0118-4

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44,* 1–23.

Monaghan, P., & Ellis, A. W. (2010). Modeling reading development: Cumulative, incremental learning in a computational model of word naming. *Journal of Memory and Language, 63,* 506–525.

Munro, R., Bethard, S., Kuperman, V., Tzuyin Lai, V., Melnick, R., Potts, C., & Tily, H. (2010). Crowdsourcing and language studies: The new generation of linguistic data. In *Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk: Proceedings of the Workshop* (pp. 122–130). Stroudsburg, PA: Association for Computational Linguistics.

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

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13,* 45–52. doi:10.3758/BF03193811

Pérez, M. A. (2007). Age-of-acquisition persists as the main factor in picture naming when cumulative word-frequency and frequency trajectory are controlled. *Quarterly Journal of Experimental Psychology, 60,* 32–42.

Sailor, K. M., Zimmerman, M. E., & Sanders, A. E. (2011). Differential impacts of age of acquisition on letter and semantic fluency in Alzheimer's disease patients and healthy older adults. *Quarterly Journal of Experimental Psychology, 64,* 2383–2391. doi:10.1080/17470218.2011.596660

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija, 43,* 441–464.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In M. Lapata & H. T. Ng (Eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 254–263). Stroudsburg, PA: Association for Computational Linguistics.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods, 43,* 155–167. doi:10.3758/s13428-010-0039-7

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. doi:10.1016/j.cognition.2003.10.009

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods, 38,* 598–605. doi:10.3758/BF03193891

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

Wurm, L. H., & Cano, A. (2010). Stimulus norming: It is too soon to close down brick-and-mortar labs. *Mental Lexicon, 5,* 358–370.

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

Yap, M. J., & Brysbaert, M. (2009). *Auditory word recognition of monosyllabic words: Assessing the weights of different factors in lexical decision performance*. Unpublished manuscript. Available at http://crr.ugent.be/papers/Yap_Brysbaert_auditory_lexical_decision_regression_final.pdf

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's*N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15,* 971–979. doi:10.3758/PBR.15.5.971

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language, 47,* 1–29. doi:10.1006/jmla.2001.2834