Age-of-acquisition ratings for 30 thousand English words

Victor Kuperman [1] **Hans Stadthagen-Gonzalez** [2] **Marc Brysbaert** [3]

[1] McMaster University, Canada

[2] Bangor University, UK

[3] Ghent University, Belgium

Keywords: word recognition, age-of-acquisition, ratings, Amazon Mechanical Turk

Corresponding author: Victor Kuperman, Ph.D.

Department of Linguistics and Languages, McMaster University

Togo Salmon Hall 626

1280 Main Street West

Hamilton, Ontario, Canada L8S 4M2

phone: 905-525-9140, x. 20384

vickup@mcmaster.ca

Abstract

We present age-of-acquisition (AoA) ratings for 30,121 English content words (nouns, verbs, and adjectives). For data collection, this mega-study 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 US students reached 0.93 for a subsample of 2,500 monosyllabic words). We also show that our AoA ratings explain a substantial percentage of 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.

Age-of-acquisition ratings for 30 thousand English words

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 paper we will argue that age-of-acquisition (AoA) should be part of this list and we provide ratings for a substantial number of words to do so. First, however, we discuss the evidence in favor of the big four.

Word frequency is the most influential variable to take into account, certainly when lexical decision is the task in question (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Ferrand, Brysbaert, Keuleers, New, Bonin, Meot, Augustinova, & Pallier, 2011). If the frequency measure comes from an adequate corpus, the percentage of variance explained by this variable in lexical decision times easily exceeds 30% (Brysbaert & New, 2009; Ferrand, New, Brysbaert, Keuleers, Bonin, Méot, Augustinova, & Pallier, 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 a words 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 lexical decision times (Ferrand et al., 2010, 2011) and naming latencies (Yarkoni et al., 2008). OLD20 is a measure of orthographic similarity and calculates the minimum number of letter changes needed to transform the target word into 20 other words. For instance, the 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/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, Buchmeier, Conrad, Jacobs, Bölte, and Böhl (2011) ran a stepwise regression analysis on the lexical decision times of the English Lexicon Project (Balota, Yap, Cortese, Hutchison, Kessler, Loftis, Neely, Nelson, Simpson, & Treiman, 2007). In this project lexical decision and word naming times for over 40 thousand English words were collected. In addition, information about 20 word variables has been made available, 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 or 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
- The number of phonemes and syllables of the word
- The 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 to predict lexical decision time was word frequency, accounting for 40.5% of the variance. The second most important variable was OLD20, which accounted for additional 12.9% of variance. The unique contribution of the third variable, the number of syllables, dropped to 1.2%, and the summed contribution of the remaining variables amounted to a mere 2.0% (Brysbaert et al., 2011). Other authors also reported that the percentage of variance explained by new variables usually is less than 1% once the big four are partialled out (e.g., Baayen, Feldman, Schreuder, 2006; Juhasz, Yap, Dicke, Taylor, & Gullick, 2011).

A promising variable to add to the big four is age-of-acquisition (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 lexical decision times of English monosyllabic words in addition to

the best word frequency measure available (also see 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, Meot, & Chalard, 2004; Zevin & Seidenberg, 2002; but see Perez, 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 they have been encountered. Words learned first are easier to access than words learned later (Izura, Perez, Agallou, Wright, Marin, Stadthagen-Gonzalez, & Ellis, 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 at which age 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 mega-studies 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 (1980), 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 techniques of linguistic data collection may alleviate the situation, however. In particular, the crowdsourcing technology of Amazon Mechanical Turk as an internet market place has provided language researchers with an attractive new tool. Amazon Mechanical Turk (https://www.mturk.com/mturk/welcome) is a web-based service where a pool of anonymous web surfers can earn money by completing tasks supplied by researchers. One type of task is a questionnaire, which enables fast and cheap collection of subjective ratings, including norms of properties of words. Basic demographics, statistics and best practices of use of the Amazon Mechanical Turk have been recently reviewed in Mason and Suri (2012). Also, the last years have seen a proliferation of papers addressing the validity of the Amazon Mechanical Turk data compared to laboratory data and the procedures that need to be followed for ensuring good data quality (Gibson, Piantadosi, & Fedorenko, 2011; Mason & Suri, 2012; Munro, Bethard, Kuperman, Lai, Melnick, Potts, Schoebelen, & Tily, 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 ones and 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 one of the authors (MB) is currently compiling, we selected all 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 30,121words. 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 10 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 they were likely to encounter. 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 towards the middle of the scale.

We used the same instructions as for the collection of the Bristol norms (Stadthagen-Gonzalez & Davis, 2006). Participants were asked for each word 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 Figure 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 some 20 minutes to complete. Participants were paid half a dollar cent per rated word (i.e., $1.81 for a validly completed list).

Responders were limited to those residing in the US. No further restrictions were imposed (e.g., no requirement of English as the first language or the only language spoken by the responder). Participants were asked to also report their age, gender, their first language or

[1] Calibrator words 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; hernia 12.6.

languages, which country/state they lived in the most 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, masters degree, or doctorate.

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 words had less than 18 valid observations after this phase. They were recombined in new, comparable lists at the end of the data collection and presented to new participants until the required number of observations was reached for next to all words.

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

Results

Data trimming. About 7% of responses were empty cells, which were removed. Valid responses were defined as either a numeric AoA rating that was smaller than the responder's age, or a response "x" that signified a "Don't know" answer. AoA ratings that were equal to the responder's age were re-labeled as "Don't know" responses (less than 0.5% of all responses). About 1% of the non-empty responses were removed as they did not match our definition of a valid response or exceeded the responder's age. Participants were instructed that there was a lower boundary of a correlation with control words required to earn the 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 out of 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 we included in our study, 2,300 (7.7%) were not known to half of the respondents. For completeness, this paper and supplementary materials provide mean numeric ratings for *all* 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 only use the mean numeric ratings if they are reported to be based on at least 5 numeric responses.

All but 8 words received 18 or more valid ratings. The correlation between the mean numeric ratings for the control words and the Bristol norms was $r = 0.93$ ($N = 50$, $p < 0.0001$). The correlation between the odd-numbered and the 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, Franklin, & Howard, 2001; Stadthagen-Gonzalez & Davis, 2006) used a linear transformation procedure to homogenize the means and standard deviation of the blocks (for details see page 600 of Stadthagen-Gonzalez & Davis, 2006). We applied this procedure to a random sample of 5 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, 1136 were female and 593 male. The age ranged from 15 to 82 years, with 8% of the responders younger than 20 years; 47% between 20 and 29; 22% between 30 and 39; 12% between 40 and 49; and 11% older than 49. Twelve 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.

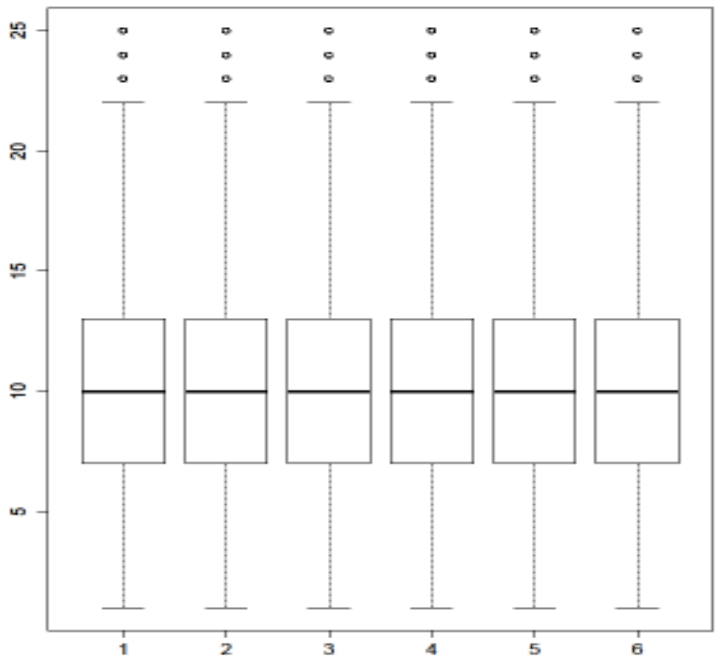
Education levels were labeled as follows: "Declined to answer" or "No high school" – 1; "High School Graduate" – 2; "Some college, no degree" – 3; "Associate degree" – 4; "Bachelors degree" – 5; "Master or higher degree" – 6. Table 1 shows the distribution of ratings and responders over the various categories.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

Table 1: Education level of the responders

Education Level	Percent 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

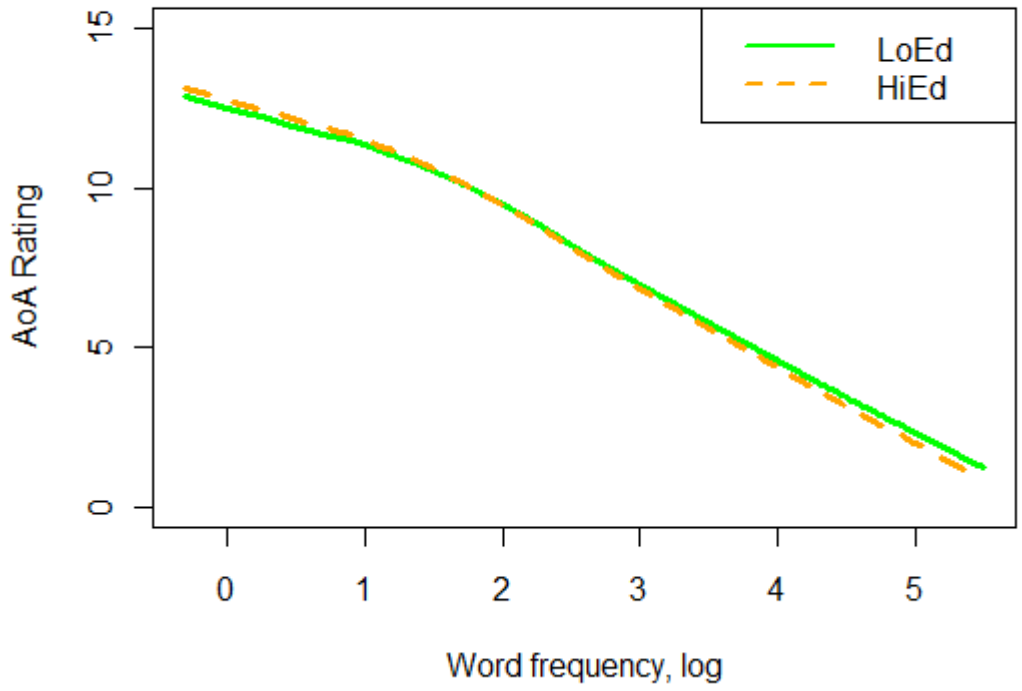
Does demography affect the numeric ratings? Women gave slightly but significantly higher AoA numeric ratings (M = 10.2, sd = 4.4) than men (M = 10.1, sd = 4.2; t = -10.27, df = 440410, p-value < 0.0001). The numeric AoA ratings did not vary by the education level of responders, as shown in the box plots of the AoA ratings in Figure 1. This null effect in subjective judgments of age-of-acquisition is surprising, given the wealth of developmental literature showing that early advantages in the vocabulary size (e.g., larger numbers of word types learned earlier) are excellent predictors of future educational achievements (e.g. Biemiller & Slonim, 2001).

Figure 1: AoA ratings as a function of education level



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 relationship between the AoA ratings and log (base 10) SUBTLEX frequency for both groups. There is a hint of an interaction (which is significant at $p < 0.05$, 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 lower-educated individuals: both differences were well within 0.2 year.

Figure 2: The association between AoA and log word frequency as a function of education level. LoEd comprises education levels 1-3 (808 responders), HiEd comprises education levels 4-6 (686 responders).



Finally, there was a weak positive correlation between AoA ratings and the age of the participants ($r = 0.07$, $t = 61.00$, $df = 615965$, $p < 0.0001$). On average, older participants gave higher AoA ratings than younger participants, presumably because they had a broader age range to choose from.

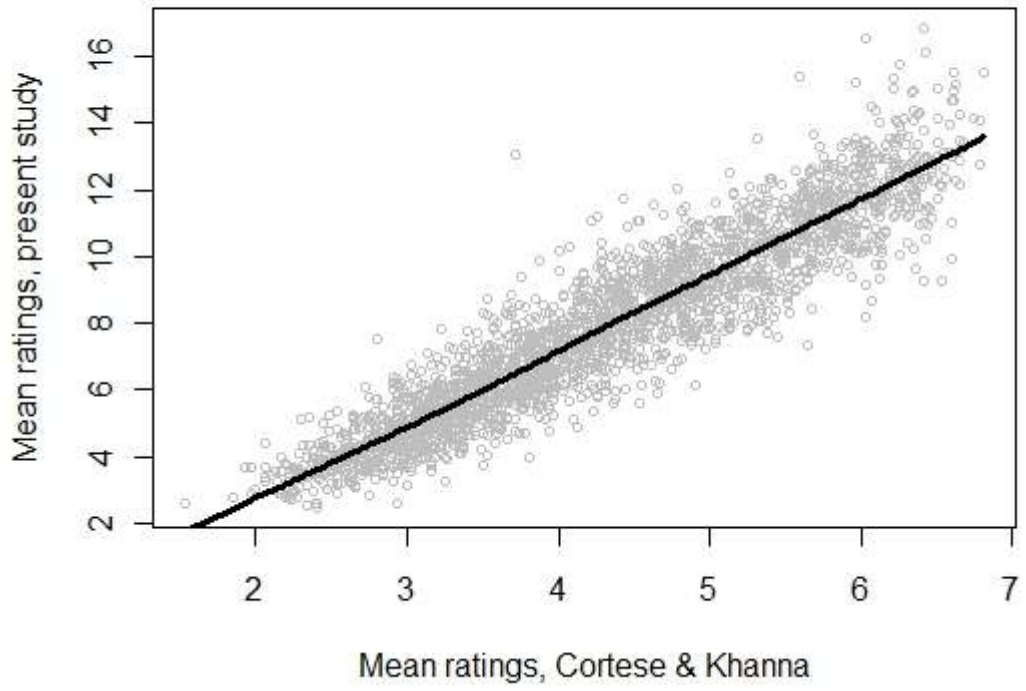
Does demography affect the number of "don't knows"? For each word, we computed the ratio of numerical responses to total responses, as an index of the responders' familiarity with this word. The ratio correlated strongly with the log frequency of the word ($r = .56$, $t = 509.9$, $df = 565587$, $p < 0.0001$) but no demographic variable was a significant predictor of the ratio. Perhaps

most surprisingly, the average percent 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 jeopardize the quality of data by using less controlled sources. There are three large-scale studies with which we can compare our mean ratings. 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 UK. Most of their participants were between 50-80 years (mean age of 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 (1980) ratings (collected in Aberdeen) for another 1,900 words.

We had 2,544 words in common with Cortese and Khanna (2008). The correlation between our ratings and theirs is $r = .93$ (Figure 3).

Figure 3: AoA ratings of Cortese and Khanna (2008; collected on the 1 to 7 Likert scale) plotted against present AoA ratings, with a solid black lowess trend line: $r = .93$, $p < 0.0001$ (based on 2,544 monosyllabic words).



There were 1,787 words in common with Bird et al. (2001), which correlated r = .83. Finally, there were 3,117 words 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. There may be some small differences in AoA-ratings between the US and the UK, 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.

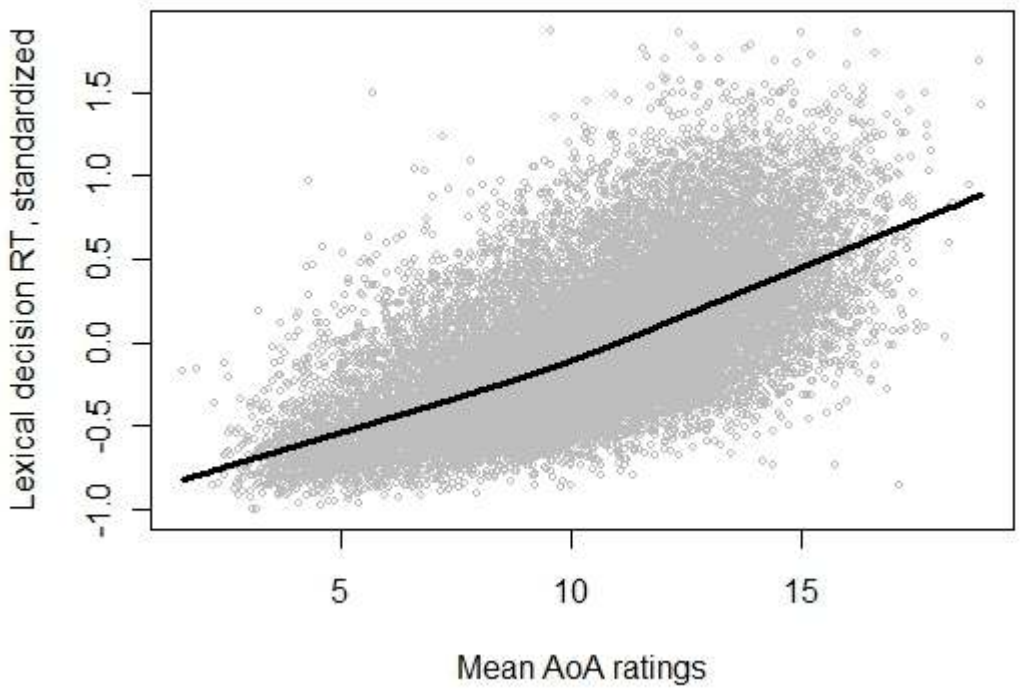
Correlation with the lexical decision data of the English Lexicon Project. Further validation of our AoA ratings is obtained by correlating them with the lexical decision data of the English Lexicon Project (ELP). There were 20,302 words in common between ELP and our list. For these words, we calculated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

Table 2: Correlations between word characteristics and the standardized reaction times and accuracy levels of the lexical decision task in the English Lexicon Project (N = 20,302 lemmas)

	zRT	Acc
AoA	.637	-.507
Log frequency (SUBTLEX)	-.685	.464
Nletters	.554	.041
Nsyllables	.537	.021
Coltheart's N	-.347	.069
OLD20	.600	-.082

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

Figure 4: Standardized ELP lexical decision response times plotted against present AoA ratings, with a solid black trend line: $r = 0.64$, $p < 0.0001$, based on 20,302 words.



The importance of the AoA variable further becomes 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 non-linear, by using restricted cubic splines for these variables. Of the many analyses we ran (and which can easily be replicated by any interested reader, as all values are freely available), we list below the ones that highlight the predictive power of AoA. For the interpretation, it is important to realize that R^2

differences of even .01 typically (and in present analyses) come with p-values below the conventional thresholds of significance (because of the large number of observations).

R^2-values for regressions on zRT:

- Freq + AoA: $R^2 = .549$
- Freq + Nlett + Nsyl + OLD20: $R^2 = .615$
- Freq + Nlett + Nsyl + OLD20 + AoA: $R^2 = .653$

R^2-values for regressions on accuracy:

- Freq + AoA: $R^2 = .318$
- Freq + Nlett + Nsyl + OLD20: $R^2 = .335$
- Freq + Nlett + Nsyl + OLD20 + AoA: $R^2 = .433$

AoA explains an extra 4% of variance in zRTs 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 reaches 10%. Compared to the influence of other variables (which usually explain less than 1% additional variance; cf. the introduction), these are substantial effects.

Are AoA ratings also predictive of inflected word forms? Having access to AoA ratings of 30 thousand 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 regular base verb has three inflected forms (3rd person, present and past participle), 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. There

were 10,011 inflected word forms in ELP 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 did 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 (cf. above). Table 3 shows the results of the inflected words.

Table 3: Correlations between word characteristics and the standardized reaction times and accuracy levels of the lexical decision task in the English Lexicon Project for inflected word forms (N = 10,011)

	zRT	Acc
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

As Table 3 suggests, there were strong correlations between lexical decision performance on inflected forms and 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 zRT:

- Freq + AoA: R^2= .488
- Freq + Nlett + Nsyl + OLD20: R^2 = .558
- Freq + Nlett + Nsyl + OLD20 + AoA: R^2 = .583
- Freq + Nlett + Nsyl + OLD20 +Freq_lemma: R^2 = .571
- Freq + Nlett + Nsyl + OLD20 + Freq_lemma + AoA: R^2 = .585

R^2-values for regressions on accuracy:

- Freq + AoA: R^2= .243
- Freq + Nlett + Nsyl + OLD20: R^2 = .271
- Freq + Nlett + Nsyl + OLD20 + AoA: R^2 = .318
- Freq + Nlett + Nsyl + OLD20 +Freq_lemma: R^2 = .297
- Freq + Nlett + Nsyl + OLD20 + Freq_lemma + AoA: R^2 = .322

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 percent 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 (Kucera & 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 the 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 was 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,
- pronunciability rating of the word,
- estimated ease of giving a definition, and estimate of whether a word has different meanings

Factor analysis suggested that the 32 variables formed 9 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 3 extra variables (log SUBTLEX frequency, our new AoA rating, and OLD20) to the list, and looked at the correlations with the standardized RT of the ELP lexical decision

task. There were values for 896 of the original 925 words. Table 4 lists the correlations in decreasing order of absolute values. This shows that the correlation with zRT was strongest for word frequency, followed by the estimated pronunceability 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. Further interesting is that our AoA ratings correlated .90 with those of Clark & Paivio (2004) and correlated slightly higher with the zRTs than the Clarke & Paivio AoA ratings.

Table 4: Correlations between word characteristics and the standardized reaction times of the lexical decision task in the English Lexicon Project for the words listed in Clark and Paivio (2004; N = 896). Ordered from high to low.

Log SUBTLEX-US frequency	-0.757 **
Estimated ease of pronunciation	-0.735 **
Familiarity rating 1	-0.727 **
Familiarity rating 2	-0.724 **
Log Thorndike-Lorge frequency	-0.714 **
Word availability (number of times word is produced as associate)	-0.711 **
Estimated ease to produce context	-0.691 **
AoA rating (current study)	0.690 **
AoA rating (Paivio)	0.657 **
Log Kucera-Francis frequency	-0.640 **
Word availability (times the word is used in dictionary definitions)	-0.625 **
Estimated ease of defining the word	-0.615 **
Log number of childhood dictionaries in which the word occurs	-0.595 **
Imageability rating 1	-0.582 **
OLD20	0.577 **
Length in letters	0.549 **
Length in syllables	0.528 **
Estimated number of similarly sounding words	-0.515 **
Estimated number of associates to the word	-0.465 **
Estimated number of similarly looking words	-0.442 **
Estimated number of rhyming words	-0.427 **
Meaningfulness (number of associates produced in 30 s)	-0.424 **

Imageability rating	-0.328 **
Estimated number of meanings of the word (ambiguity)	-0.287 **
Pleasantness rating	-0.266 **
Emotionality rating	-0.217 **
Estimated number of words that start with the same sounds	-0.201 **
Estimated goodness/badness of the word's meaning	-0.176 **
Concreteness rating	-0.166 **
Deviation emotionality rating from the mean rating	-0.122 **
Deviation goodness rating from the mean rating	-0.071 **
Estimated number of words starting with the same letters	-0.064*
Gender ladenness rating 1	-0.027
Gender ladenness rating 2	-0.017
Log number of high frequency words starting with the same two letters	0.008

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

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 less data (896 instead of 925), we failed to observe a significant contribution of the final factor (meaning ambiguity). Therefore, we worked with an 8-factor model instead of the original 9-factor model. We also included the additional variables log SUBTLEX-US frequency, OLD20, and zRT of 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 outcome of the factor analysis is 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 not significantly related to emotionality, word onset, gender ladenness, or pleasantness of the words. Interestingly, AoA loaded on exactly the same factors, just like 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, which 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 AoA measure were used (Brysbaert & Cortese, 2011; Ferrand et al., 2011).

Table 5: Factor loadings of the different variables in Clark and Paivio's (2004) study and four new variables on the words for which we had all the data (N = 896). Lexical decision times load on four factors only. 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. Variables ordered as in Table 4.

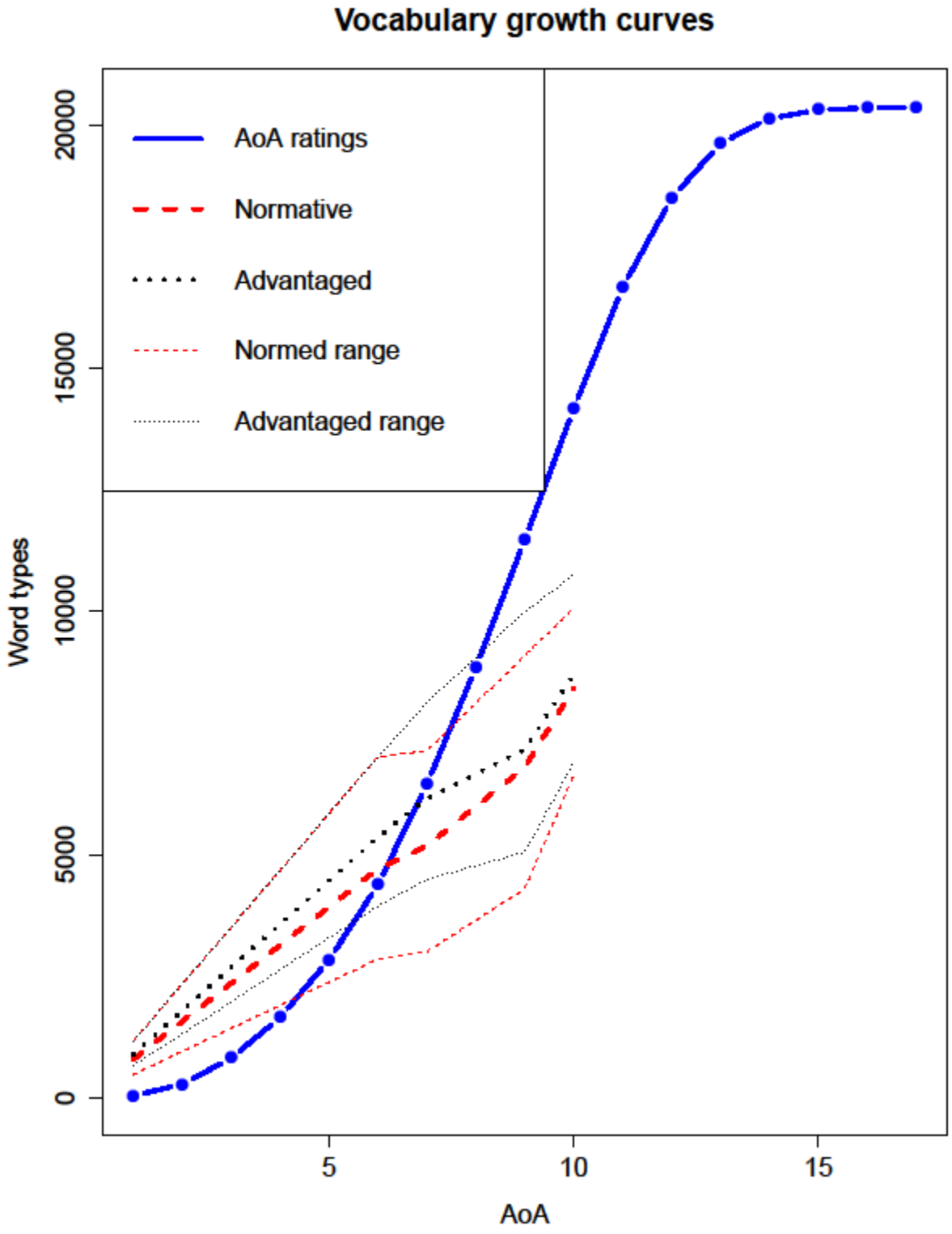
	Freq.	Len.	Fam.	Ima.	EmoDev.	Gender	Onset	Pleasant
zRT ELP Lexical Decision Task	**-0.522**	**-0.428**	**-0.526**	-0.138				
SUBTLEX-US frequency	**0.739**	0.284	**0.394**	0.127	0.178			
Estimated ease of pronunciation	**0.388**	**0.361**	**0.623**	0.138				0.107
Familiarity rating 1	**0.615**	0.131	**0.627**	0.140		0.125		0.104
Familiarity rating 2	**0.371**		**0.876**	0.112	0.111		0.117	
Thorndike-Lorge frequency	**0.795**	0.257	0.285	0.171				0.129
Word availability (produced as associate)	**0.706**	**0.381**	0.266	0.293	0.118			
Estimated ease to produce context	0.298	0.104	**0.842**	0.285	0.141			
AoA rating (current study)	**-0.432**	**-0.315**	**-0.496**	**-0.467**				
AoA rating (Paivio)	**-0.421**	**-0.326**	**-0.445**	**-0.513**		-0.108		-0.117
Kucera-Francis frequency	**0.824**	0.112	**0.305**				0.113	0.121
Word availability (used in dictionary)	**0.778**	**0.312**	0.143					
Estimated ease of defining the word	0.267		**0.729**	**0.424**				
Number of childhood dictionaries	**0.593**	0.283	0.238	**0.489**				0.106
Imageability rating 1	0.197	0.184	**0.543**	**0.715**	0.119			
OLD20	-0.259	**-0.851**		-0.104				
Length in letters	-0.256	**-0.793**		-0.186			0.273	
Length in syllables	-0.189	**-0.755**		-0.251			0.103	
Similarly sounding words (estimation)	0.185	**0.846**	0.145	0.102			0.154	
Associates to the word (estimation)	**0.419**		**0.386**		**0.381**			0.127
Similarly looking words (estimation)	0.155	**0.700**	0.199				0.251	
Rhyming words (estimation)	0.120	**0.762**	0.144				0.233	
Meaningfulness (number of associates)	0.200	0.155	0.295	**0.651**				
Imageability rating 2		0.174	0.187	**0.908**				

Meanings of the word (estimation)	0.249	0.197	0.183	**-0.306**	0.228			0.101
Pleasantness rating	0.205		0.151		0.125	0.229		**0.928**
Emotionality rating	0.143		0.204	-0.150	**0.799**			0.108
Start with the same sounds (estimate)	0.104	0.200	0.160				**0.726**	
Goodness/badness of meaning	0.174					0.240		**0.864**
Concreteness rating		0.149		**0.863**	-0.287			
Deviation emotionality from mean					**0.838**			
Deviation goodness from mean					**0.900**			
Start with same letters (estimation)							**0.785**	
Gender ladenness rating 1						**0.964**		0.184
Gender ladenness rating 2						**0.940**		0.231
High frequency words starting with same letters							**0.658**	
SS loadings	5.443	4.956	4.782	3.962	2.582	1.966	1.894	1.870
Proportion Var	0.151	0.138	0.133	0.110	0.072	0.055	0.053	0.052
Cumulative Var	0.151	0.289	0.422	0.532	0.603	0.658	0.711	0.763

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, i.e., the guesstimated vocabulary growth curve. We divided mean AoA ratings into yearly bins, from 1 to 17, and computed the cumulative sum of word types falling into each bin. This subjective estimate of vocabulary growth is compared in Figure 5 to the estimates obtained via experimental testing of children's vocabulary in Biemiller and Slonim (2001). Biemiller and Slonim presented a representative sample and a sample with an advantaged socio-economic 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 and Slonim, 2001). We relabeled grades 1-5 into ages 6 to 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 (dashed line) and the group with an advantaged socioeconomic status (dotted line), as reported by Biemiller and Slonim (2001). For each group we also include confidence intervals (based on the estimated number of lemmas known to the 0-25% and the 75-100% percentiles of the group).

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


Vocabulary growth curves
AoA ratings
Normative
Advantaged
Normed range
Advantaged range
Word types
0
5000
10000
15000
20000
5
10
15
AoA

Several aspects of the comparison between estimated and measured vocabulary growth are noteworthy. First, our responders put the main weight of word learning to the elementary school years, from 6 to 12. This underestimates the growth in the years 2-5 (the AoA estimates are lower than those in Biemiller & Slonim) and overestimates the growth after the age of 9 (AoA estimates are higher than in Biemiller & Slonim). Also, responders report that hardly any words enter their vocabulary before the age of 3 and after the age of 14. Only a small percentage (1.2%) of 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 (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 higher education years.

Discussion

In this article, we described the collection of AoA ratings for 30 thousand English content words (nouns, verbs, and adjectives) with 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 $r = .93$. In particular, the correlations with previously collected American ratings were high (Clark & Paivio, 2004; Cortese & Khanna, 2008; see Figure 3). This means that the internet

crowdsourcing technology forms a useful tool for the rapid gathering of large numbers of word characteristics (nearly 2000 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 US (or to English-speaking countries more in general) and to have some online control of the quality of the data. This is 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 correlate with the values already available. In this way, the quality of the data is 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 sharp 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 we ran, it always had a high correlation with the dependent variable (in particular 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 shown in Tables 4 and 5. The effect of AoA was not only found for the lemmas included in the rating study (Table 2), but also for the 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 the variable, because of the scarcity of ratings available. This restriction is lifted now. Having access to AoA ratings for over 30K content lemmas and their inflected forms means that researchers can routinely control their stimuli on

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, it is a long-standing question whether AoA ratings are accurate estimates of acquisition times or rather a reflection of the order of acquisition (see references above). Several aspects of our data are in line with the second possibility. First, AoA estimates seem to form a normal distribution with the mode around 9 years of age and 90% o of data points between 5 and 15 years of age (i standard deviation about 2.84 years). Importantly, this curve deviates from empirically obtained vocabulary growth curves in young age (Biemiller and Slonim, 2001; Figure 5) 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 (Figure 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 to this article. For each word, we report the number of times it occurs in the trimmed data (OccurTotal). For most words,

the count is about 19. However, for the 10 calibration words and the 52 control words, this amounts to more than 1,900 presentations. Next, we provide the mean AoA rating (in years of age) and the standard deviation (Rating.Mean and Rating.SD). We also present the number of responders that 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 SUBTLEX-US corpus (Brysbaert & New, 2009). Words are presented in the 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.

Acknowledgement

This study was supported by the 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 paper, and Danielle Moed for her help with the preparation of this manuscript.

REFERENCES

Balota, D. A., Cortese, M. J., Sergent-Marshall, S., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*(2), 283-316.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The english lexicon project. *Behavior Research Methods, 39*(3), 445-459.

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*(1), 273-285.

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*(3), 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*(1), 73-79.

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

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*(4), 456-476.

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*(5), 412-424.

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *The Quarterly Journal of Experimental Psychology, 64*(3), 545-559.

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

Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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*(4), 977-990.

Brysbaert, M., New, B., & Keuleers, E. (in press). Adding Part of Speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods.*

Brysbaert, M., Wijnendaele, I. V., & Deyne, S. D. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica, 104*(2), 215-226.

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments & Computers, 36*(3), 371-383.

Cortese, M. J., & Khanna, M. M. (2008). Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods, 40*(3), 791-794.

Ferrand L., New B., Brysbaert M., Keuleers E., Bonin P., Méot A., Augustinova M., Pallier C. (2010). The French Lexicon Project: lexical decision data for 38,840 French words and 38,840 pseudowords. *Behaviour Research Methods, 42*(2), 488.

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

Fitzsimmons, G., & Drieghe, D. (2011). The influence of number of syllables on word skipping during reading. *Psychonomic Bulletin & Review, 18*(4), 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*(2), 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*(1), 43-67.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using mechanical turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass, 5*(8), 509-524.

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*(4), 395-427.

Gilhooly, K. J., & Logie, R. H. (1980). Meaning-dependent ratings of imagery, age of acquisition, familiarity, and concreteness for 387 ambiguous words. *Behavior Research Methods & Instrumentation, 12*(4), 428-450.

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*(1), 32-58.

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition, 13*(7-8), 789-845.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin, 131*(5), 684-712.

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. *The Quarterly Journal of Experimental Psychology, 64*(9), 1683-1691.

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*.

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.

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*(4), 506-525.

Munro, R., Bethard, S., Kuperman, V., Lai, V.T., Melnick, R., Potts, C., Schnoebelen, T. and Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. *Proceedings of the NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk.* 122-130.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111*(3), 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*(1), 45-52.

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

Sailor, K. M., Zimmerman, M. E., & Sanders, A. E. (in press). 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*.

Schnoebelen, T. and Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija, 43(4),* 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 Proceedings of EMNLP 2008, 254-263.

Sprouse, J. (2011). A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods, 43*(1), 155-167.

Stadthagen-Gonzalez, H., Bowers, J. S., & Damian, M. F. (2004). Age-of-acquisition effects in visual word recognition: Evidence from expert vocabularies. *Cognition,93*(1), B11-B26.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods, 38*(4), 598-605.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science: A Multidisciplinary Journal, 29*(1), 41-78.

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

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60*(4), 502-529.

Yap, M.J. & Brysbaert, M. (2009). Auditory word recognition of monosyllabic words: Assessing the weights of different factors in lexical decision performance.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971-979.

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