Test-based Age-of-Acquisition norms for 44 thousand English word meanings

Marc Brysbaert

Ghent University, Belgium

Andrew Biemiller

University of Toronto, Canada

Keywords: age-of-acquisition, word learning, reading, psycholinguistics

Running title: Test-based AoA norms

Address:

Marc Brysbaert
Department of Experimental Psychology
Ghent University
Henri Dunantlaan 2
B-9000 Gent
Belgium
Tel. +32 9 264 94 25
Fax. +32 9 264 64 96
E-mail: marc.brysbaert@ugent.be

Abstract

Age of acquisition (AoA) is an important variable in word recognition research. Up to now, nearly all psychology researchers examining the AoA effect have used ratings obtained from adult participants. An alternative basis for determining AoA is directly testing children's knowledge of word meanings at various ages. In educational research, scholars and teachers have tried to establish the grade at which particular words should be taught by examining the ages at which children know various word meanings. Such a list is available from Dale and O'Rourke's *Living Word Vocabulary* for nearly 44 thousand meanings coming from over 31 thousand unique word forms and multiword expressions. The current paper relates these test-based AoA estimates to lexical decision times as well as AoA adult ratings and reports strong correlations between all measures. Therefore, test-based estimates of AoA can be used as an alternative measure.

Test-based Age-of-Acquisition norms

for 44 thousand English word meanings

Age-of-acquisition (AoA) is one of the most important variables in word recognition: Early-acquired words are processed more efficiently than late-acquired words even when word frequency, word length, and similarity to other words are controlled for (Brysbaert & Ellis, 2016; Brysbaert, Stevens, Mandera, & Keuleers, 2016; Johnston & Barry, 2006; Juhasz, 2005).

The existing AoA norms are based on ratings provided by adult volunteers (often students). Participants are asked to indicate the ages at which they think they have learned various words (e.g., Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). A weakness of such ratings is that they may be influenced by factors other than pure AoA. For instance, participants may be inclined to underestimate the AoA of easy words and overestimate the AoA of more difficult words. Easy words tend to be short and frequently used in the language; in contrast, difficult words tend to be long, less used words. So AoA ratings may be affected by word length and word frequency, in addition to other variables that make some words easier than others (Baayen, Milin, & Ramscar, 2016; Lété & Bonin, 2013). On the other hand, all validation studies thus far have indicated that adult AoA ratings correlate highly with test-based measures of word acquisition order (Biemiller, Rosenstein, Sparks, Landauer, & Foltz, 2014; Brysbaert, in press; Łuniewska et al., 2016; Morrison, Chappell, & Ellis, 1997).

Another limitation of AoA ratings is that they tend to be constrained in a number of ways. First, they are not available for all words. The largest collection of AoA ratings in English includes 30 thousand words (Kuperman et al., 2012), which is still short of a full vocabulary. Second, very few studies take the various meanings of ambiguous words into account (for an

exception, see Bird, Franklin, & Howard, 2001). For instance, 'wrong' can mean 'not right' but also 'to treat unfairly.' Both interpretations are unlikely to be acquired at the same age. Finally, no norms are available for familiar multiword expressions, such as phrasal verbs (give in, give over, give up, …) or compound nouns (witness stand, word of honor, …).

For the above reasons, it would be better if researchers had access to another, large-scale database of AoA estimates. Such an effort was made by Dale and O'Rourke (1981), who wanted to provide teachers with guidelines about which words to teach in which grades. Dale and O'Rourke tested nearly 44,000 meanings (31,000 different word forms) to determine at which age children were considered to "know" a meaning. This was assessed by giving the pupils three-alternative multiple-choice test items. Words were assigned to the grade level at which 67-80% of specific word meanings were passed. Adjusted for guessing on a three alternatives multiple-choice test, this amounted to an estimated 50-70% known. In other words, the assigned grade level for a word meaning was known by *half* or slightly more students. The tests were administered in grades 4, 6, 8, 10, 12, and college levels (13 and 16). The researchers estimated the grade level to test. If the result for a specific meaning fell outside of the 67-80% range, it was tested at the next higher or lower grade level.

Testing was conducted in schools throughout the U.S. Midwest. Various written tests were sent to participating classrooms. Any specific meaning was given to about 200 children across a number of schools. At the time of testing (1950-1980), most children were English-speaking. A range of socio-economic backgrounds and races were sampled.

Biemiller (2010) used the Dale and O'Rourke's (1981) list as the basis of a new list of root words worth teaching. He made an additional category of grade 2, in which he put all words known by more than 80% of the children in grade 4, based on findings in Biemiller & Slonim (2001).

The purpose of the present study is to use and slightly update the Dale and O'Rourke (1981) database and to validate its use as a source for AoA estimates based on word knowledge at different school grades. For the validation, we compare the test-based AoA data with rated AoA estimates, lexical decision times (Balota et al., 2007), and word frequency. We suggest that the available test-based data (Dale & O'Rourke, 1981) provide a useful additional estimate of word AoA in English, which has the additional advantage that multiple meanings of ambiguous words are considered.

Three other, smaller lists of test-based AoA estimates are available. First, Goodman, Dale, and Li (2008) published a list of the first 562 English word learned by toddlers, as scored by thousands of parents. Second, Morrison et al. (1997) published AoA ratings in young children on the basis of picture naming for 297 pictures. Finally, a recent website for American teachers published a list of 1,461 words to be taught in various classes from Kindergarten to grade 8 (https://www.flocabulary.com/wordlists/; retrieved on March 4, 2016).

Words present in the lists of Goodman et al. (2008) were assigned to grade 2, the lowest estimated value in Dale and O'Rourke's scale as revised by Biemiller (2010). For a few words, this meant a large change in estimated AoA. For instance, *yogurt* went from grade 10 to grade 2. Because the Morrison et al. (1997) study was run in the UK, no similar adjustment was made, although in the large majority of cases the data agreed with Dale and O'Rourke.

Validation of the Dale and O'Rourke **Test-Based AoA norms**

In the present study, we used two ways to validate the new test-based AoA norms: First by correlating them with AoA ratings; and second, by correlating them with word processing times.[1]

There were 18,139 words for which we had test-based AoA estimates, AoA ratings, and standardized lexical decision times in the English Lexicon Project (Balota et al., 2007). For words with multiple meanings, the test-based AoA measure was the youngest meaning in the database, based on the assumption that participants rated the word's first acquired meaning. Biemiller et al (2014) also found this "earliest AoA meaning" to be the one that fits with the existing AoA ratings.

We gave a rating of 14 to *Living Word Vocabulary* meanings above level 13 (Figure 1 confirms that this was the most sensible value to give).

Table 1 shows the correlations between the various variables. From this table it is clear that the test-based AoA estimates correlate highly with the ratings collected by Kuperman et al. (2012). Figure 1 shows the correlation.

- -

Insert Table 1 and Figure 1 about here

- -

[1] Some readers may wonder why we do not validate the test-based AoA norms on the basis of variables derived from word frequencies at various school ages, such as the "word frequency trajectory". More information on this can be found in Brysbaert (in press), who called word frequency trajectory one of the worst word characteristics ever introduced in psycholinguistics, because it does not correlate well with any of the validation criteria used. Although one can compare word frequencies at different grades, the differences do not correspond well to the order in which words are acquired, probably because many words are acquired after a few observations and because frequency norms at different ages come from different language registers.

There is a higher correlation between the AoA ratings and lexical decision times (r = .608) than between the test-based AoA estimates and lexical decision times (r = .525). On the other hand, the correlations between the test-based AoA estimates with word frequency and word length are smaller, indicating that the test-based estimates are less affected by these variables.

To calculate the contribution of both AoA measures to word processing times, hierarchical regression analyses were run, which additionally included word frequency and word length. They are shown in Table 2. As can be seen, the model including AoA ratings does significantly better than the model including test-based AoA estimates (z = 5.24 according to a Vuong test for non-nested models; Merkle & You, 2016), but the difference in terms of explained variance is 1.2%, instead of the 9.4% expected on the basis of the correlations listed in Table 1. This is due to the lower intercorrelations of the test-based AoA estimate with word frequency and word length.

- -

Insert Table 2 about here

- -

Discussion

In this paper, a new test-based AoA measure is introduced, largely based on the work of Dale and O'Rourke (1981), who presented words with three response alternatives to children from primary and secondary school and examined at which grade the words were known. The list was updated with the more recent CDI resource, which looked at younger ages. All in all, data are available for nearly 44 thousands meanings coming from over 31 thousand English words and multiword expressions.

Although the test-based measure has a rather crude scale (in steps of 2 grades), it does quite well to predict lexical decision times (Table 2) and, as such, takes away some of the concerns that have been raised against the use of AoA ratings to examine a genuine effect of AoA in word processing times (Baayen et al., 2016; Lété & Bonin, 2013; see also Brysbaert, in press). The Dale and O'Rourke grades are slightly inferior to the more recent and more refined Flocabulary grades, as they correlate less with the lexical decision times of the English Lexicon Project for the 1260 words with information on all variables ($r = .433$ instead of $r = .487$; Hotelling-Williams test: $t(1257) = -1.91$, $p < .06$)), but the Flocabulary grades are only available for 1,461 words.

The new measure is not perfect, but it presents an interesting alternative to the Kuperman et al. (2012) ratings. First, as indicated above, it is test-based rather than a subjective, retrospective estimate. Second, it is available for other words than the existing ratings. The regression to go from the grades to the best fitting AoA rating equals: rating = 5.72 + .554 * grade. By using this regression, both sources can be combined. Third, the measure is available for many familiar multiword expressions (in particular, phrasal verbs and compound nouns). Fourth, it is the first measure really taking into account the various meanings words may have. It is estimated that 15% of the words in English have more than one meaning (Goulden, Nation, & Read, 1990). Now, we can look at the processing of word meanings that follow earlier acquired meanings. Finally, the new measure may be particularly interesting for studies with older participants (Brysbaert & Ellis, 2016), given that the AoA values were derived at the time when they were young.

To help researchers, we have made a file with the test-based AoA measure used in the present study (available at https://osf.io/kz2px/). The file also contains the AoA ratings collected by Kuperman et al. (2012), the original LWV grades, the CDI and Morrison et al. (1997) estimates in number of months, and the Flocabulary grades (see Figure 2). The list is made

References

Baayen, R.H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology, 30*, 1174-1220.

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

Biemiller, A. (2010). *Words worth teaching: Closing the vocabulary gap.* Columbus, OH: SRA/McGraw-FINI.

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.

Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T. K., & Foltz, P. W. (2014). Models of vocabulary acquisition: Direct tests and text-derived simulations of vocabulary growth. *Scientific Studies of Reading, 18(2),* 130-154.

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.

Brysbaert, M. (in press). Age of acquisition ratings score better on criterion validity than frequency trajectory or ratings 'corrected' for frequency. *Quarterly Journal of Experimental Psychology.*

Brysbaert, M. & Ellis, A.W. (2016). Aphasia and age-of-acquisition: Are early-learned words more resilient? *Aphasiology, 30*, 1240-1263.

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(4),* 977-990.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance ;42(3),*441-458.

Dale, E., & O'Rourke, J. (1981). *The Living Word Vocabulary, the Words We Know: A National Vocabulary Inventory*. Chicago, IL: World book.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language, 35(3),* 515-531.

Goulden, R., Nation, I.S.P. and Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics 11(4),* 341-363.

Johnston, R .A. & Barry, C. (2006) Age of Acquisition and lexical processing: A review. *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.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods, 44*, 978-990.

Lété, B., & Bonin, P. (2013). Does frequency trajectory influence word identification? A cross-task comparison. *The Quarterly Journal of Experimental Psychology, 66(5),* 973-1000.

Łuniewska, M., Haman, E., Armon-Lotem, E., Etenkowski, B., Southwood, F., ... & Unal-Logacev, O.. (in press). Ratings of age of acquisition of 299 words across 25 languages: Is there a crosslinguistic order of words? *Behavior Research Methods, 48*, 1154-1177.

Merkle, E., &, D. (2016). *Package 'nonnest2'*. (Retrieved from https://cran.r-project.org/web/packages/nonnest2/nonnest2.pdf, March 4, 2016)

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology: Section A, 50(3),* 528-559.

	AoA_{rating}	Frequency	Length	LDT
$AoA_{test\text{-}based}$	0.757	-0.587	0.262	0.525
AoA_{rating}		-0.675	0.395	0.608
Frequency			-0.401	-0.640
Length				0.510

Table 1: Pearson correlations between the variables (N = 18,139). $AoA_{test\text{-}based}$ is the present scale based on actual performance. AoA_{rating} are the ratings collected by Kuperman et al. (2012). Frequency is the log SUBTLEX-US frequency (Brysbaert & New, 2009). Length is the number of letters in the word. LDT is the standardized lexical decision time from the English Lexicon Project (Balota et al., 2007). Very similar values are obtained when Spearman correlations are calculated.

	R^2	ΔR^2
Test-based AoA estimates		
LDT = frequency + length	51.9%	51.9%
LDT = frequency + length + AoA	54.4%	2.5%
AoA ratings		
LDT = frequency + length	51.9%	51.9%
LDT = frequency + length + AoA	55.2%	3.3%

Table 2: Percentage of variance accounted for by the various variables. Non-linear estimates for word frequency and word length (restricted cubic splines). A linear estimate for AoA, as more was not required.

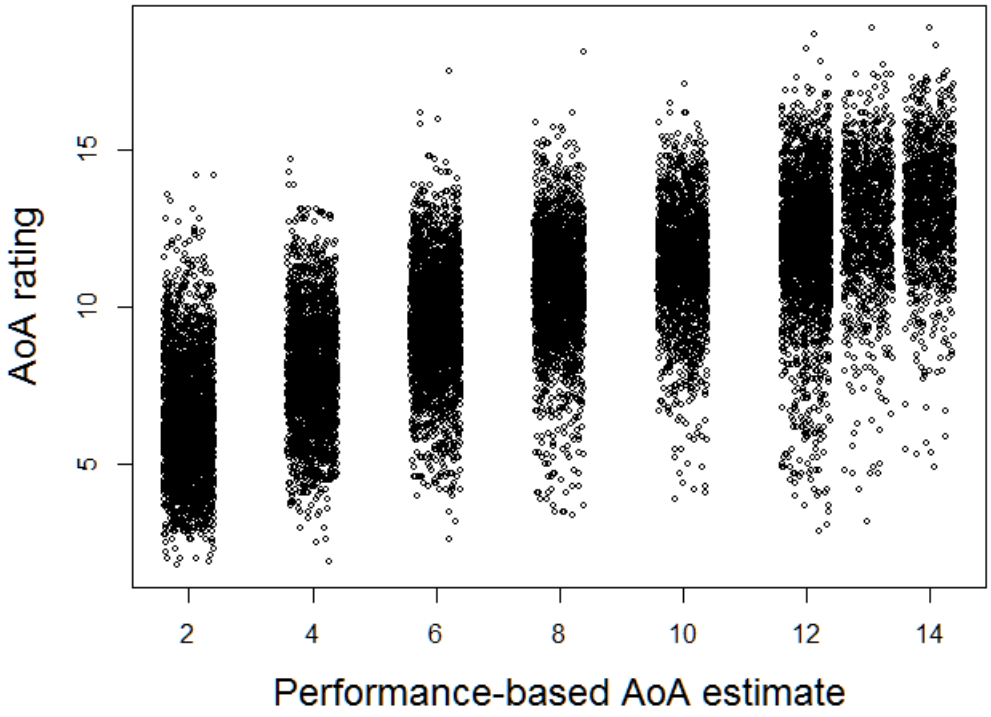

Figure 1: Correlation between the test-based AoA estimates and the AoA ratings. Jitter added to the test-based AoA estimates to diminish the overlap of the observations.

	A	B	C	D	E	F	G	H
1	WORD	MEANING	AoAtestl	AoArating	LWV	CDI	Morr	Floc
3080	bastille	a prison	12		12	#N/A	#N/A	#N/A
3081	bastings	long stitches	8		8	#N/A	#N/A	#N/A
3082	bastion	fortification	14	15.4	16	#N/A	#N/A	#N/A
3083	bat	hard blow	2	4.7	4	25	56.5	#N/A
3084	bat	ball-player's stick	2	4.7	4	25	56.5	#N/A
3085	bat	small flying animal	4	4.7	4	25	56.5	#N/A
3086	bat	wild good time	14	4.7	16	25	56.5	#N/A
3087	bat boy	takes care of team's bats	4		4	#N/A	#N/A	#N/A
3088	bat your eyes	to close and open eyelids quickly	4		4	#N/A	#N/A	#N/A
3089	batch	lot of	4	7.2	4	#N/A	#N/A	1
3090	batch	unmarried man	12	7.2	12	#N/A	#N/A	1
3091	batch	male housekeeping	13	7.2	13	#N/A	#N/A	1
3092	bate	to reduce in intensity	12		12	#N/A	#N/A	#N/A
3093	bateau	flat-bottomed boat	12		12	#N/A	#N/A	#N/A
3094	bath	wash	2	3.5	4	17	23.4	#N/A

Figure 2: Screenshot of the file containing the test-based AoA measures used in the present paper. It shows the words with their different meanings, as tested by Dale and O'Rourke (1981). The figure also shows how the LWV grades were adapted (4 became 2 when more than 80% of the children in grade 4 knew the meaning of the word or when the word was part of the 562 first learned words according to the CDI database; 16 became 14). The file further contains the AoA ratings collected by Kuperman et al. (2012), the age at which children can name pictures according to Morrison et al. (1997), and the suggested grade to teach the word according to the Flocabulary website. Because only Dale and O'Rourke provide different values for the various meanings of homographs, the other measures always have the same value for all meanings of a word. The Flocabulary grades in general are lower than the Dale and O'Rourke grades, suggesting that children now learn words at an earlier age than 40-50 years ago.