Behav Res (2017) 49:1520–1523
DOI 10.3758/s13428-016-0811-4

Test-based age-of-acquisition norms for 44 thousand English word meanings

Marc Brysbaert[1] · **Andrew Biemiller**[2]

Published online: 22 September 2016

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 (1981) *Living Word Vocabulary* for nearly 44 thousand meanings coming from over 31 thousand unique word forms and multiword expressions. The present article relates these test-based AoA estimates to lexical decision times as well as to AoA adult ratings, and reports strong correlations between all of the measures. Therefore, test-based estimates of AoA can be used as an alternative measure.

Keywords Age of acquisition · Word learning · Reading · Psycholinguistics

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).

Existing AoA norms are based on ratings provided by adult volunteers (often students), in which participants are asked to indicate the ages at which they think they 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. Thus, AoA ratings may be affected by word length and word frequency, as well as by 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, 2016; 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 these reasons, it would be better if researchers had access to another, large-scale database of AoA estimates.

✉ Marc Brysbaert
marc.brysbaert@ugent.be

[1] Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Gent, Belgium

[2] University of Toronto, Toronto, Ontario, Canada

Such an effort was made by Dale and O'Rourke (1981), who wanted to provide teachers with guidelines regarding which words to teach in which grades. Dale and O'Rourke tested nearly 44,000 meanings (31,000 different word forms) to determine at what age children were considered to "know" a meaning. This was assessed by giving pupils three-alternative multiple-choice test items. Words were assigned to the grade level at which 67 %–80 % of the 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, and 12, and at two college levels (13 and 16). The researchers estimated the grade levels 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, and 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. Also, 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 the words known by more than 80 % of the children in grade 4, on the basis of the findings in Biemiller and Slonim (2001).

The purpose of the present study was 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 compared the test-based AoA data with the 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.

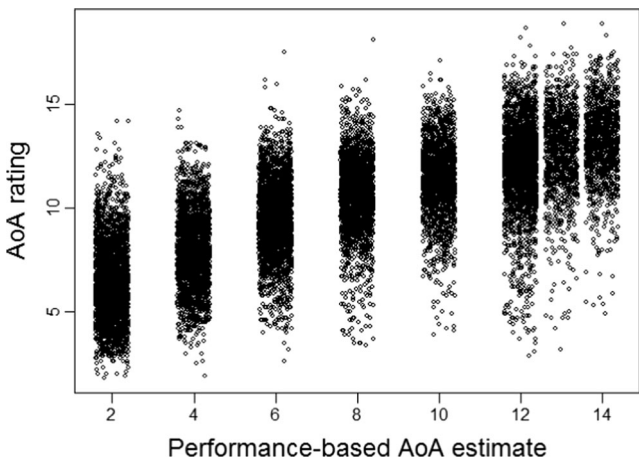

Fig. 1 Correlations between the test-based AoA estimates and the AoA ratings. Jitter is added to the test-based AoA estimates to diminish the overlap of the observations

Table 1 Pearson correlations between the variables (N = 18,139)

	AoA_{rating}	Frequency	Length	LDT
$AoA_{test\text{-}based}$	.757	–.587	.262	.525
AoA_{rating}		–.675	.395	.608
Frequency			–.401	–.640
Length				.510

$AoA_{test\text{-}based}$ 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 were obtained when Spearman correlations were calculated.

Three other, smaller lists of test-based AoA estimates are also available. First, Goodman, Dale, and Li (2008) published a list of the first 562 English words learned by toddlers, as scored by thousands of parents. Second, Morrison et al. (1997) published AoA ratings in young children based on 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 March 4, 2016).

The words present in the lists by 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 the 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 for that study, although in the large majority of cases the data agreed with those of 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]

[1] Some readers may wonder why we did 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 (2016); he 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 the frequency norms at different ages come from different language registers.

Table 2 Percentages of variance accounted for by the various variables

	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 %

Nonlinear estimates were used for word frequency and word length (restricted cubic splines). A linear estimate is given for AoA, because more was not required.

We had test-based AoA estimates, AoA ratings, and standardized lexical decision times in the English Lexicon Project for a total of 18,139 words (Balota et al., 2007). For words with multiple meanings, the test-based AoA measure was taken as the youngest meaning in the database, on the basis of the assumption that participants were rating 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 (Fig. 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 individual correlations.

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 also smaller, indicating that the test-based estimates are less affected by these variables.

To calculate the contributions of both AoA measures to word-processing times, hierarchical regression analyses were run, which additionally included word frequency and word length. The results 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 nonnested models; Merkle & You, 2016), but the difference in terms of explained variance is 0.8 %, instead of the 9.4 % that would be 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.

Discussion

In this article, a new test-based AoA measure was introduced, based largely on the work of Dale and O'Rourke (1981), who presented words with three response alternatives to children from primary and secondary schools and examined at which grades the words were known. The list was updated with the more recent communicative development inventories (CDI)

	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

Fig. 2 Screenshot of the file containing the test-based AoA measures used in the present article. It shows the words with their different meanings, as tested by Dale and O'Rourke (1981). The figure also shows how the *Living Word Vocabulary* grades were adapted (grade 4 became grade 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; in addition, a rating of 16 became 14). The file further contains the AoA ratings collected by Kuperman et al. (2012), the age at which children could 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 they did 40–50 years ago

resource, which looked at younger ages. All in all, data are available for nearly 44,000 meanings coming from over 31,000 English words and multiword expressions.

Although the test-based measure has a rather crude scale (in steps of two grades), it does quite well predicting lexical decision times (Table 2), and as such, this takes away some of the concerns that have been raised regarding 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, 2016). The Dale and O'Rourke grades are slightly inferior to the more recent and more refined Flocabulary grades, since they correlate less with lexical decision times from the English Lexicon Project for the 1,260 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 we indicated above, it is test-based rather than a subjective, retrospective estimate. Second, it is available for words not included in the existing AoA ratings. The regression to go from the grades to the best-fitting AoA rating is: 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, for phrasal verbs and compound nouns). Fourth, it is the first measure to really take into account the various meanings that 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 the 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 measures 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 *Living Word Vocabulary* grades, the CDI and Morrison et al. (1997) estimates in numbers of months, and the Flocabulary grades (see Fig. 2). The list is made available for research purposes under the Creative Commons Non-Commercial License (https://creativecommons.org/); it must not be used for commercial purposes.

References

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

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

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

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,* 130–154.

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

Brysbaert, M. (2016). Age of acquisition ratings score better on criterion validity than frequency trajectory or ratings "corrected" for frequency. *Quarterly Journal of Experimental Psychology*. Electronic preprint. doi:10.1080/17470218.2016.1172097

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,* 977–990. doi:10.3758/BRM.41.4.977

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,* 441–458. doi:10.1037/xhp0000159

Dale, E., & O'Rourke, J. (1981). *The living word vocabulary, the words we know: A national vocabulary inventory*. Chicago: 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,* 515–531.

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

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition, 13,* 789–845. doi:10.1080/13506280544000066

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

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44,* 978–990. doi:10.3758/s13428-012-0210-4

Lété, B., & Bonin, P. (2013). Does frequency trajectory influence word identification? A cross-task comparison. *Quarterly Journal of Experimental Psychology, 66,* 973–1000. doi:10.1080/17470218.2012.723725

Łuniewska, M., Haman, E., Armon-Lotem, E., Etenkowski, B., Southwood, F., Anđelković, D., … Ünal-Logacev, Ö. (2016). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*, *48*, 1154–1177. doi:10.3758/s13428-015-0636-6

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

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal of Experimental Psychology, 50A,* 528–559. doi:10.1080/027249897392017