Journal of Child Language (2019), **46**, 199–213
doi:10.1017/S0305000918000363

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Subjective ratings of age-of-acquisition: exploring issues of validity and rater reliability

Carla WIKSE BARROW[1], Kristina NILSSON BJÖRKENSTAM[2],
and Sofia STRÖMBERGSSON[1]*

[1]Division of Speech and Language Pathology, Karolinska Institutet, Stockholm, Sweden and [2]Department of Linguistics, Stockholm Universitet, Stockholm, Sweden
*Corresponding author: Department of Clinical Science, Intervention and Technology, Division of Speech-Language Pathology, Karolinska Institutet, Huddinge F67, Stockholm 14186, Sweden.
E-mail: sofia.strombergsson@ki.se

(Received 28 November 2017; revised 12 June 2018; accepted 11 August 2018;
first published online 23 October 2018)

Abstract

This study aimed to investigate concerns of validity and reliability in subjective ratings of age-of-acquisition (AoA), through exploring characteristics of the individual rater. An additional aim was to validate the obtained AoA ratings against two corpora – one of child speech and one of adult speech – specifically exploring whether words over-represented in the child-speech corpus are rated with lower AoA than words characteristic of the adult-speech corpus. The results show that less than one-third of participating informants' ratings are valid and reliable. However, individuals with high familiarity with preschool-aged children provide more valid and reliable ratings, compared to individuals who do not work with or have children of their own. The results further show a significant, age-adjacent difference in rated AoA for words from the two different corpora, thus strengthening their validity. The study provides AoA data, of high specificity, for 100 child-specific and 100 adult-specific Swedish words.

Keywords: age-of-acquisition; lexical development; corpus linguistics

Introduction

In spite of child language development being characterized by tremendous individual variation, there are developmental milestones that appear to be universal, for instance canonical babble at 6–10 months, first words at 12 months, telegraphic speech at 24–30 months, etc. (Stoel-Gammon, 2011; Toppelberg & Shapiro, 2000; Vihman, 2014). Moreover, there is an emerging body of evidence showing several striking, cross-linguistic similarities in early vocabulary development with regard to pace of vocabulary growth (Bleses *et al.*, 2008) and distribution of word types in the early vocabulary (Caselli *et al.*, 1995).

Age-of-acquisition (henceforth AoA) is a psycholinguistic construct that refers to the age at which the average child learns a given word (Carroll & White, 1973). The AoA effect, referring to the processing advantage that words acquired early in life hold over those learnt later, is well established in the word recognition literature (Brysbaert, 2017), affecting performance on tasks such as picture naming (Juhasz, 2005), word naming (Cortese & Khanna, 2007), semantic classification, lexical decision, and more (Łuniewska et al., 2016). AoA has furthermore been suggested to affect words' resilience in Alzheimer's disease (Cuetos, Herrera, & Ellis, 2010) and aphasia (Brysbaert & Ellis, 2015).

The most desirable estimate of AoA is arguably objective data, based on the analysis of children's recorded speech production (Morrison, Chappell, & Ellis, 1997). Objective AoA can be determined as the age at which a word appears in a given percentage of the children sampled, or when it reaches a predetermined cumulative frequency criterion (Łuniewska *et al.*, 2016). Test-based AoA, for example picture-elicited production, has also been proposed as an alternative method, the advantage of which being a more direct measure of children's knowledge of word meanings (Morrison *et al.*, 1997). However, such objective AoA estimates are restricted by the context and time during which the speech recordings occur (Łuniewska *et al.*, 2016), and test-based scores may be affected by the task performed (Brysbaert, 2017). Furthermore, many young children refuse to cooperate with strangers, and therefore, eliciting responses in picture-naming tasks as well as spontaneous speech may be both difficult and time-consuming (Law & Roy, 2008).

Many other estimates of AoA have been proposed, of which one of the most widely distributed is the MacArthur-Bates Communicative Development Inventory (CDI) (Fenson, Marchman, Thal, Dale, Reznick, & Bates, 2007). CDI is a parental report instrument, designed to capture information about children's developing vocabulary, and other linguistic abilities (Frank, Braginsky, Yurovsky, & Marchman, 2017). These parental questionnaires generate data from which AoA can be calculated, that is, the age in months where a percentage of children reportedly produce or comprehend a given word (Fenson, Dale, Reznick, Bates, Thal, & Pethick, 1994), as applied in for instance Goodman, Dale, and Li (2008) and Hansen (2017). Although CDI-questionnaires are widely used as instruments in descriptions of child language acquisition, questions regarding validity, sensitivity, and parental bias have been raised; for example the educational level of the parents has been found to influence the accuracy of their ratings (Law & Roy, 2008).

Most previous AoA studies have enlisted adult raters, predominantly undergraduate students, to report on their own age of acquisition for a given sample of words (Alario & Ferrand, 1999; Bakhtiar, Nilipour, & Weekes, 2013; Cortese & Khanna, 2007; Ferrand *et al.*, 2008; Stadthagen-Gonzalez & Davis, 2006), allowing collection of AoA estimates for a larger number of words (Birchenough, Davies, & Connelly, 2017). However, the use of adult raters is one of the main reasons why the AoA construct is contested (Brysbaert, 2017). The target question in such studies is habitually posed with reference to the subjects themselves, and the age at which they acquired a given word. The issue of self-reported, subjectively rated AoA has been described as problematic (Brysbaert, 2017), as these ratings may be influenced by other information, as AoA ratings often correlate with other lexical variable such as frequency, imageability, familiarity, and concreteness (Bird, Franklin, & Howard, 2001; Ferrand *et al.*, 2008; Schröder, Gemballa, Ruppin, & Wartenburger, 2012; Stadthagen-Gonzalez & Davis, 2006).

It has, moreover, been found that raters in AoA studies tend to underestimate the number of words learned before the age of four, and after the age of fifteen (Brysbaert, 2017). This observation corresponds with infantile amnesia, a

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

well-documented psychological phenomenon, manifested in adults as rarely recalling events from their childhood prior to the age of three, and furthermore having scarce memories of events occurring between the ages of three and seven, in which time a tremendous number of words are acquired (Madsen & Kim, 2016). In fact, Lind, Simonsen, Hansen, Holm, and Mevik (2015) found that adult raters underestimate their own acquisition of words, when compared to parental reports of AoA through the Norwegian adaptation of CDI, such that adults rated that they, at age three, had only acquired 122 of the 442 words that were reported as present in children at that age (i.e., in 50% of children). While this does not necessarily pose a problem in studies of word processing in adult subjects, if the object is to assess the effects of order of acquisition, it poses a threat to the validity of AoA ratings, if the aim is to study early vocabulary development. A child's lexical development undergoes considerable acceleration during the first few years of life (Stoel-Gammon, 2011), and data purporting to reflect this development is naturally constrained if language acquisition prior to four years of age is not captured.

In their large cross-linguistic study, Łuniewska and colleagues (2016) proposed that the target question be changed to concern children in general, as opposed to the subjects themselves, that is, "When do children learn the word … ?" rather than "When did you learn the word… ?" In fact, they found that individuals who rate when children acquire words report significantly lower AoA than those reporting their own experience (Łuniewska *et al.*, 2016). While changing the target question might circumvent the issue of infantile amnesia, there remain other concerns regarding the validity of utilising undergraduate students as raters of AoA. Regardless of whether the target question is directed at their own language development or at that of children in general, the average undergraduate student is not to be expected to have extensive familiarity with children in the midst of early language development. In the United States of America, a clear majority (77%) of undergraduate students, were, in 2010, younger than 30 years (Chronicler of Higher Education, 2010), while the mean age for first births is rising and was 26.4 years in 2015 in the USA (Martin, Hamilton, Osterman, Driscoll, & Mathews, 2017). Similar demographics are visible in other areas of the world. For example, in Sweden, the mean age for first births was 29.3 years for mothers and 31.6 years for fathers in 2017 (Statistiska Centralbyrån [SCB], 2018b), while 70.9% of higher education students were under 30 years in 2008 (SCB, 2018a). Hence, it can be questioned whether the average undergraduate student will have sufficient experience with children to be considered a valid rater of AoA. The authors of the present study recognize the benefits of rating of AoA, in terms of practicality, objectivity, and cost-efficiency, and therefore take a different approach to the problems facing this method: perhaps it is not the process of rating that is questionable, but rather the raters themselves.

One group of interest for subjective ratings of AoA are individuals with familiarity with children. In the present study, we have chosen to focus on parents and preschool teachers. Łuniewska and colleagues (2016) hypothesised that parenting young children might influence an individual's ability to assess their own AoA. Indeed, they found that parents of children under 10 years of age rated earlier AoA than the control group for 99% of the words presented, while the order of acquisition was very similar. Although parents arguably have valuable insight into their children's development, another group of individuals that can be expected to have high familiarity with children amidst early language development is preschool teachers. In Sweden, the average preschool teacher

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

encounters 16 children between the ages one and six, on a daily basis (Skolverket, 2016). While a parent may lack a precise frame of reference or be prone to bias (Law & Roy, 2008), the average preschool teacher will, through years of contact with children in the midst of early language development, have accumulated both knowledge and experience of the variability and trends in children's developing language. This unique experience makes preschool teachers particularly interesting to involve in studies obtaining AoA ratings.

A limited number of studies have explored rater characteristics and potential effects of demographic variables on AoA ratings. A recent addition to the research area is reported by Birchenough and colleagues (2017) who, in their large-scale web-based collection of AoA ratings from adult German individuals, found that neither multilingualism nor educational level influenced the ratings of AoA, whilst age and gender did have a weak effect, such that men and older participants were found to provide slightly higher subjective AoA ratings. However, Luniewska and colleagues (2016) found no significant differences in AoA ratings based on the raters age, education, or gender.

When assessing the validity of a construct, values generated from one instrument are often correlated with previously obtained values, generated from an instrument designed to measure the same, or a similar, construct – through so-called 'convergent validity'. Previous validations of subjective AoA have correlated obtained ratings with more objective measures of AoA (Carroll & White, 1973; Gilhooly & Gilhooly, 1980; Morrison *et al.*, 1997), with CDI parental reports (Łuniewska *et al.*, 2016), and with existing subjective ratings of AoA in the same or other languages (Brysbaert, Stevens, De Deyne, Voorspoels, & Storms, 2014; Ferrand *et al.*, 2008; Raman, Raman, & Mertan, 2014), often producing high correlations. However, as much of the criticism of AoA estimates from adult ratings are based on the subjectivity of the data (Brysbaert, 2017; Hansen, 2017), employing more objective estimates of AoA as an anchor may be a more appropriate option. We argue that the most valid, as well as the most practicable, reference material is existing recordings of multiple individuals at different ages. This type of data allows for analyses of what words (and other linguistic features) are characteristic at different ages (see e.g. Daland, 2013; Geirut & Dale, 2007; Zevin & Seidenberg, 2004). The most ecologically valid kind of such data is, arguably, recordings based on spontaneous conversation. Although many researchers have relied on linguistic resources representing language DIRECTED TO children and/or adults at different ages (see e.g. Brysbaert, 2017; Zevin & Seidenberg, 2004), a more direct validation of AoA ratings of PRODUCTION of words is through correlation with linguistic resources representing language PRODUCED BY children and adults, respectively. The present study aims to explore the relation between production frequency and rated AoA, by investigating whether the most overused words in a child-speech corpus were rated as being acquired earlier than the most overused words in an adult-speech corpus, and vice versa.

While validity and reliability are occasionally presented within and across groups (Łuniewska *et al.*, 2016; Moors *et al.*, 2013; Schröder *et al.*, 2012), intra-rater reliability, that is, whether the rater is consistent in his or her ratings, and the validity of each individual's ratings is seldom reported in studies obtaining subjective ratings of AoA (Alario & Ferrand, 1999; Bird *et al.*, 2001; Bonin, Peereman, Malardier, Méot, & Chalard, 2003; Cortese & Khanna, 2007; Ferrand *et al.*, 2008; Łuniewska *et al.*, 2016; Moors *et al.*, 2013; Raman *et al.*, 2014; Stadthagen-Gonzalez & Davis, 2006; Zevin & Seidenberg, 2004). Exploring the validity of the instruments

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

(i.e., the individual raters) should be a high priority, in particular when the number of raters is limited, often around 18–35 individuals (Alario & Ferrand, 1999; Brysbaert *et al.*, 2014; Cortese & Khanna, 2007; Della Rosa, Catricalà, Vigliocco, & Cappa, 2010; Ferrand *et al.*, 2008; Łuniewska *et al.*, 2016; Moors *et al.*, 2013; Stadthagen-Gonzalez & Davis, 2006). In their large web-based study, Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) did explore the validity of each individual rater, by correlating their subjective AoA ratings with previously obtained AoA norms (Stadthagen-Gonzalez & Davis, 2006). This process led to the exclusion of 15% of the obtained ratings, indicating that many individuals are not suitable for ratings of AoA. Motivated by a concern that this potential problem may undermine the validity of earlier AoA research, we further aim to investigate the role of the individual rater, treating each subject as an instrument of measurement, and also exploring the potential influence of familiarity with children on the accuracy of said instrument.

Hypothesis 1: Individuals with a higher familiarity with children in the midst of early language development (defined here as between one and six years) are more appropriate as raters of AoA, and will score better on validity and reliability criteria, than individuals with low familiarity with children of that age.

Hypothesis 2: Words overused in the child-speech corpus (child-specific words) will be rated as being acquired earlier than words overused in the adult-speech corpus (adult-specific words), thus validating the obtained ratings against authentic child-speech.

Method

Participants

Adult speakers of Swedish (n = 145) were recruited for anonymous participation via email or social media. Child-care professionals were enlisted by email through their preschool managers or supervising teachers, whose email addresses were obtained via the online search engine Google. Around 331 preschools, surrounding the 13 largest cities in Sweden, were approached with a letter of interest. Forty-five preschools responded, of which 30 were positive to participation. An invitation to participate was sent to all preschool managers who did not respond negatively (n = 321). All other participants were recruited through the social media platform Facebook. Participants who reported they did not speak Swedish with the children they encountered were excluded from analyses (n = 5).

A total of 140 individuals were included in the study, of which 79 reportedly worked with one- to six-year-old children every day, 16 did so occasionally, and 45 individuals did not work with children within that age range at all.

Materials

The linguistic data were obtained by comparing a frequency list extracted from a child-speech corpus to a corresponding list extracted from an adult-speech corpus. The child-speech frequency list was extracted from the Strömqvist–Richthoff corpus (Strömqvist, Richthoff, & Andersson, 1993), which is available through CHILDES (MacWhinney, 2000) as the Lund corpus. This longitudinal corpus (5 children; age 1;0 to 6;0; approximately 125,000 tokens) consists of transcripts of spontaneous interaction with family members and/or a researcher in a home environment. The orthographic transcripts also include vocalizations, largely following the CHILDES conventions. The adult-speech frequency lists were extracted from a set of orthographically transcribed corpora (approximately 2.3 million tokens combined, hereinafter referred to as THE ADULT-SPEECH CORPUS): (1) the Gothenburg dialogue corpus, available through Språkbanken at Gothenburg University (Allwood, Grönqvist, Björkberg, Ahlsen, & Ottesjö, 2000); (2) Spontal: 120 dyads of spontaneous interaction in a lab environment (Edlund, Beskow, Elenius, Hellmer, Strömbergsson, & House, 2010); and (3) Swedia 2000: interview transcripts of speakers of Swedish dialects (Eriksson, 2004).

A spreadsheet was generated containing information on the raw frequency of each word in the child-speech corpus and the adult-speech corpus, respectively. Furthermore, the spreadsheet included information on the relative frequency of each word in child speech, given the total number of tokens in the child-speech corpus, and the corresponding value for that word in adult speech, given the size of the adult-speech corpus. Following Rayson and Garside (2000), the log-likelihood ratio was calculated to capture differences in relative word frequency between the child- and adult-speech corpora, where a log-likelihood ratio of > 3.84 signified a significant difference. Over- or under-use of a word in the child-speech corpus was determined by comparing the relative frequency of the word in the child-speech corpus with the relative frequency of that word in the adult-speech corpus. By sorting the over-used words in the frequency lists according to the log-likelihood ratio in decreasing order, two lists of the 100 most significantly over-used words from each corpus were compiled, for inclusion in a web-based survey (see below). Nonsense words and onomatopoetic sounds were excluded. Table 1 shows the 10 most frequent, of the 100 most over-used words from the child-speech corpus and adult-speech corpus, respectively.

Procedure

An internet-based survey was created using a standard format in Google forms, accepting responses during four weeks in the summer of 2017. The initial two sections of the survey contained demographic questions concerning the social and linguistic environment of each participant, in addition to questions pertaining to their familiarity with one- to six-year-old children. The final section of the survey comprised a list of 200 words, for which the participants were asked to estimate children's age of acquisition. Half of the above-mentioned words (n = 100) were retrieved from the list of over-used words in the child-speech corpus and half (n = 100) from the corresponding list based on the adult-speech corpus. The stimuli will hereinafter be referred to as 'child-specific words' and 'adult-specific words'. Six of the words were duplicated, to allow assessment of intra-rater reliability (#25, #50, and #75 from each corpus). The order in which the words were presented was randomized using a Perl-script. The duplicated words were presented with 26–150 words between them (M = 89 words).

A 12-point scale was utilized for ratings of AoA, ranging from one to six years of age, in equidistant 6-month intervals. The last point of the scale was termed 'later' (i.e., after six years of age). The rating scale was restricted to six years or 'later' as the group in focus were preschool teachers, who encounter children of one to six years of age. Participants were instructed to choose the age-alternative that best corresponded to their estimation of the age at which children learn to say each word. Participants were further instructed to include simplifications of pronunciation, such as /des/

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

Table 1. The 10 most frequent, over-used words from the child-speech corpus and the adult-speech corpus, presented with their respective log-likelihood ratio values, representing the degree of over-use

Child-specific words	Log-likelihood ratio	Adult-specific words	Log-likelihood ratio
mamma 'mum'	4407.65	*det* 'it/that'	11413.59
den 'it/that one'	2993.51	*jag* 'me'	3931.60
pappa 'dad'	1142.08	*ju* 'as you know'*	3053.52
titta 'look'	1112.15	*att* 'to/that'	2868.58
denna 'this one'	1089.65	*är* 'is/are'	1777.39
sitta 'sit'	883.77	*för* 'for'	1654.55
min 'my/mine'	808.95	*och* 'and'	1621.58
docka 'doll'	781.04	*med* 'with'	1466.25
hon 'her'	726.93	*var* 'where'	1247.44
bada 'take a bath'	464.40	*som* 'as/like'	1192.92

Note. * = discourse marker.

instead of /dres/ for 'dress', in their ratings. In case of uncertainty, participants were encouraged to guess. The target question read "När uppskattar du att barn lär sig att säga … ?" 'When do you estimate that children learn to say …? ', followed by the list of words.

Statistical analyses were performed in IBM SPSS Statistics for Macintosh, Version 24.0. Microsoft Excel version 15.32 was utilised for data-formatting throughout.

Data analyses

All AoA ratings were recoded into a corresponding 12-point ordinal scale, for analysis of validity and reliability. One hundred and twenty of the words included in the AoA survey were also present in the Swedish adaptation of the MacArthur-Bates Communicative Development Inventory: Words and Sentences (SECDI w&s) (Berglund & Eriksson, 2000). SECDI w&s (henceforth, SECDI) is based on 900 parental reports of the linguistic development of 336 children aged 16–28 months. SECDI Words and Sentences includes 710 words, as well as questions regarding the child's grammatical and pragmatic development (Berglund & Eriksson, 2000). SECDI allows derivation of age-based norms of productive skills as well as pragmatic and grammar skills. Data from SECDI was downloaded from Wordbank (Frank *et al.*, 2017) on 29 June 2017.

The overlapping set of words enabled analysis of convergent validity, by comparing the obtained AoA ratings to SECDI norms. AoA norms derived from SECDI were determined as the age at which each given word was reported as present in $\geqslant$ 75% of children (i.e., at 16, 19, 22, 25, or 28 months of age) (similar to the procedure used in Goodman and colleagues, 2008). All words that were not present in $\geqslant$ 75% of children at 28 months ($n = 71$) were excluded from analysis as AoA could not be determined. Spearman's rank correlation coefficient was calculated between each individual participant's ratings and the SECDI-norms, where the rater validity criterion was set at moderate agreement $\rho \geqslant 0.5$ (Mukaka, 2012).

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

For assessment of intra-rater reliability, Cohen's weighted kappa (Cohen, 1968) was calculated to determine agreement between each individual participant's ratings of the six duplicated words. Reliability criterion was set at $\kappa \geqslant 0.61$, substantial agreement (Landis & Koch, 1977). Participants were naive to the words' duplication.

The ratings of all participants meeting the validity and reliability criteria were tallied, generating a mean AoA rating for each word. A two-tailed independent samples *t*-test was conducted on the mean AoA ratings of all 200 words (including the mean rating of the six duplicated words), with corpus type (child-speech vs. adult-speech) as a grouping variable. Alpha level was set at 0.05 for all analyses.

Ethics

As the survey was anonymous, and no personal data were processed, ethical vetting was not necessary. All present research was, nonetheless, conducted in accordance with the WMA Declaration of Helsinki (World Medical Association [WMA], 2013). All participants were informed that involvement was voluntary and anonymous.

Results

Table 2 shows the proportional distribution of participating raters to meet both the validity and reliability criteria, across groups separated by vocational experience and familiarity with one- to six-year-old children.

As shown in Table 2, 55 of 140 individual raters met the validity criterion of $\rho \geqslant 0.5$ (min $\rho = 0.5$, max $\rho = 0.73$) and 77 individuals met the reliability criterion of moderate agreement $\kappa \geqslant 0.61$ (min $\kappa = 0.61$, max $\kappa = 0.96$). Only 37 individuals met both the validity and the reliability criteria. All agreement was found significant at the 0.01 level.

As seen in Table 2, vocational experience, and to a lesser degree frequency and duration of said experience, was found to affect the proportion of participants to meet the validity and reliability criteria, such that the highest proportion of individuals deemed valid and reliable work with children every day, closely followed by individuals who occasionally work with children. However, Table 2 also shows that the group of subjects who have children under the age of three years reaches the highest proportion of individuals to meet rater criteria for validity and reliability. The group of subjects who do not work with children and have no children of their own shows the lowest percentage of individuals to meet the validity and reliability criteria. Only one individual, of 18, from this group meets the criteria for validity and reliability.

AoA ratings versus corpus distributions

An independent samples *t*-test was conducted to compare the mean ratings of the child- and adult-specific words, obtained from the 37 individuals who were both valid and reliable. A statistically significant, age-adjacent difference in subjectively rated AoA was found [$t(195) = -12.48$ $p < .001$], such that child-specific words were rated with lower AoA ($M = 29.4$ months, $SD = 7.7$) and adult-specific words with higher AoA ($M = 43.9$ months, $SD = 8.7$). The distribution of ratings, for 100 words from the child-speech corpus and 100 words from the adult-speech corpus, can be seen in Figure 1.

Table 3 presents a sample of the AoA data collected, for illustration; the 10 words with the lowest mean rated AoA, and the 10 words with the highest mean rated

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

Table 2. Number (percentage) of raters meeting validity and reliability criteria, across groups separated by vocational experience and familiarity with preschool-aged children

Group	Reliable	Valid	Reliable and valid
Work with children every day (n = 79)	42 (53%)	38 (48%)	25 (32%)
Work with children occasionally (n = 16)	11 (69%)	7 (44%)	5 (31%)
Do not work with children (n = 45)	24 (53%)	19 (42%)	7 (16%)
Have children < 3 years (n = 11)	6 (55%)	7 (64%)	4 (36%)
Have children < 10 years (n = 6)	5 (83%)	1 (17%)	1 (17%)
Have children > 10 years (n = 10)	6 (60%)	1 (10%)	1 (10%)
No children (n = 18)	8 (44%)	1 (6%)	1 (6%)
TOTAL	77 (55%)	55 (39.3%)	37 (26.4%)

AoA, respectively. It can be noted that all 10 words with lowest-rated AoA were from the child-speech corpus, and all the 10 highest-rated words were from the adult-speech corpus. Further, it can be observed that the variability in ratings is considerably larger for the words rated with the highest AoA compared to the words rated with the lowest AoA, in line with previous findings (Moors *et al.*, 2013).

Discussion

The debate regarding the validity of AoA as a variable in psycholinguistic research is largely based on questions regarding the derivation of AoA data, that is, adult self-reported ratings (Brysbaert, 2017). The present study has presented an investigation of subjective ratings of AoA, exploring the suspicion that both validity and reliability vary between raters. The reliability of individual raters was assessed by comparing their ratings on duplicated word items. The validity of each participant's ratings was determined through comparison with existing AoA-norms, derived from the Swedish McArthur-Bates CDI. The obtained AoA group ratings were moreover compared to frequency distributions in two corpora – one representing child- speech and the other representing adult speech – to test their validity.

The finding that a clear majority of the participants did not meet the predetermined validity and reliability criteria indicates that theoretical issues of certain individuals' suitability as raters of AoA (Łuniewska *et al.*, 2016) are in fact problematic also in practice. Particularly noteworthy, however, was the finding that only one individual (of 18) who did not work with or have children (hence matching the profile of a standard participant in many studies) met the criteria for validity and reliability.

More individuals were found to be reliable than valid across groups, suggesting that many raters were consistent in their ratings, despite not providing valid estimates of AoA. Only participants who met the predetermined criterion for both validity and reliability were included in the study's analyses, as consistency in over- or underestimation of AoA should not be considered a criterion for inclusion. As validity and reliability scores varied across all groups, the authors conclude that both validity and reliability analyses are necessary to determine an individual's appropriateness as a rater of subjective AoA.

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

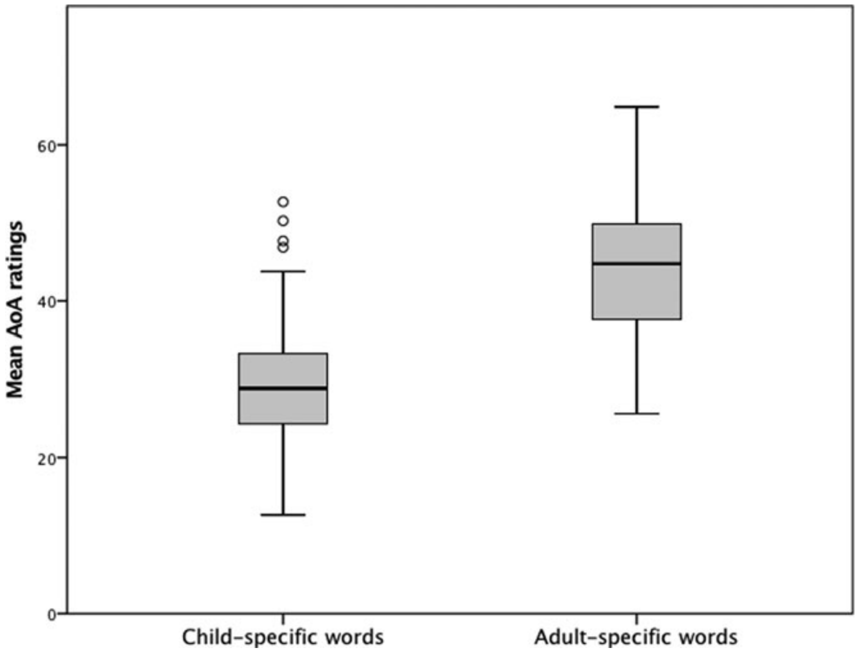

Figure 1. Boxplot showing the median and inter-quartile range of the mean AoA ratings of 100 child-specific and 100 adult-specific words, as rated in months of age, by the 37 individuals who were valid and reliable.

The discovery that familiarity with one- to six-year-old children influenced the participants' validity and reliability in ratings of AoA confirms the hypothesis that individuals with high familiarity with children in the midst of early language development are more appropriate raters of AoA than individuals with low familiarity with children in that age. These results may serve, for example, in recruitment for future studies obtaining subjective ratings of AoA.

The comparison between AoA ratings of words from the child-speech corpus and the adult-speech corpus confirmed the hypothesis that child-specific words would elicit lower AoA ratings than the adult-specific words. Not only does this finding confirm the validity of the obtained AoA ratings, it also confirms the validity of these corpora as sources of valuable linguistic information about child speech and adult speech. Considering the relatively small size of the child-speech corpus – 125,000 tokens in the present study compared to, for instance, 17 million tokens in Zevin and Seidenberg (2004) and Brysbaert (2017) – this was perhaps not self-evident. This result also aligns with suggestions of deriving AoA information from frequency measures at different ages (Zevin & Seidenberg, 2004), but differs from earlier work in that the analysis is based on language PRODUCTION at different ages. Considering that AoA in production is the focus of the present investigation, production data is arguably the most direct and ecologically valid data source.

In addition to the above investigations, the study contributes AoA data of high specificity (ratings in 6-month intervals, as compared to 1- to 2-year intervals in many previous studies (Bonin *et al.*, 2003; Ferrand *et al.*, 2008; Łuniewska *et al.*,

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

Table 3. An illustration of the extreme ends of the AoA mean ratings, showing the 10 words with the lowest mean AoA ratings, and the 10 words with the highest mean AoA ratings, as rated in months, by the 37 individuals who were valid and reliable

Rank order (from lowest to highest AoA)	Word (*Swedish* ‘English’)	Mean rated AoA (in months)*
1	*mamma* ‘mum’	12.65 (SD: 1.89)
2	*pappa* ‘dad’	12.97 (SD: 2.24)
3	*titta* ‘look’	16.22 (SD: 5.27)
4	*hej* ‘hello/hi’	16.38 (SD: 4.83)
5	*boll* ‘ball’	17.51 (SD: 5.36)
6	*lampa* ‘lamp’	17.51 (SD: 5.72)
7	*bil* ‘car’	18.00 (SD: 5.10)
8	*katt* ‘cat’	18.65 (SD: 3.95)
9	*apa* ‘monkey’	19.78 (SD: 5.63)
10	*mat* ‘food’	20.43 (SD: 6.55)
…		
191	*ändå* ‘anyway’	55.14 (SD: 13.85)
192	*väl* ‘probably/well’	57.08 (SD: 14.00)
193	*liksom*** ‘like/as well	58.05 (SD: 11.41)
194	*egentligen* ‘actually/really’	58.22 (SD: 12.48)
195	*just* ‘exactly’	58.70 (SD: 13.35)
196	*alltså* ‘so/then as’	59.35 (SD: 15.15)
197	*män* ‘men’	61.49 (SD: 13.37)
198	*exempel* ‘example’	62.27 (SD: 11.52)
199	*rå* ‘raw’	62.43 (SD: 14.73)
200	*vidare*** ‘further/furthermore’	65.84 (SD: 12.84)

Notes. * Note that the last increment of the rating scale, termed ‘later’ was recoded to 78 months; ** = duplicated words, the first presentation shown here.

2016; Moreno-Martínez, Montoro, & Rodríguez-Rojo, 2014; Stadthagen-Gonzalez & Davis, 2006)) for 200 Swedish words, up to the age of six years. For words acquired later, the ratings are – as previously discussed – both less specific and less reliable, as the boundaries of the last increment was unspecified in terms of age. These AoA data are publically available at <www.ling.su.se/english/nlp/corpora-and-resources/>.

AoA norms that reflect the age at which children IN GENERAL learn a given set of words are a potentially valuable resource for the research of disordered language development, and could provide guidance to speech-language pathologists and other professionals who encounter children at risk for language disorders. On the other hand, if AoA data are to be used as a proxy for lexical processing in adults, it can certainly be argued that the precise age at which a given word is learnt is not crucial.

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

Limitations and future research

The recruitment procedure did not allow controlling the participating raters' characteristics, and consequently, the raters were not balanced with regards to their familiarity with children. However, the number of participants is considerably larger than in most previous studies obtaining subjectively rated AoA (Alario & Ferrand, 1999; Bird *et al.*, 2001; Bonin *et al.*, 2003; Ferrand *et al.*, 2008; Łuniewska *et al.*, 2016; Raman *et al.*, 2014; Stadthagen-Gonzalez & Davis, 2006). However, the AoA norms presented here are based on the ratings of 37 valid and reliable individuals, which, in terms of numbers, is comparable to many previous studies (Cortese & Khanna, 2007; Della Rosa *et al.*, 2010; Łuniewska *et al.*, 2016; Moors *et al.*, 2013; Moreno-Martínez *et al.*, 2014). Additional research should further examine rater validity and reliability, in balanced categories of raters, to make inter-categorical comparisons possible.

The present study utilised a highly specific scale (6-month intervals, as opposed to the 1- to 3-year intervals frequently used for ratings of AoA (Bonin *et al.*, 2003; Cortese & Khanna, 2007; Łuniewska *et al.*, 2016; Morrison *et al.*, 1997), presumably making the task more difficult for the participating raters, and consequently may have affected their validity and reliability scores. Furthermore, as the words included in the survey were generated from both a child-speech corpus and an adult-speech corpus, the nature of the stimuli may have enhanced the complexity of the task at hand. Indeed, considering the low imageability of function words (Bird *et al.*, 2001), one may speculate that rating AoA of function words is a more difficult task than that of rating concrete nouns and verbs. It is, moreover, conceivable that the threshold for CDI-based AoA applied here – present in 75% of children, as opposed to 50% in previous studies (Goodman *et al.*, 2008; Hansen, 2017) – could have affected validity scores. In the light of these complicating factors, the finding that some participants nonetheless meet the posited validity and reliability criteria is perhaps surprising.

Questions concerning validity, sensitivity, and reporter bias have been raised (Law & Roy, 2008) regarding the parental report instrument CDI (Fenson *et al.*, 2007). As the present study used SECDI-norms for convergent validity analyses, the validation process may have been affected, if similar reporter bias was present in the parents who participated in this study. It is also conceivable that individuals with children under three may have been favoured in the validity analyses, as all AoA norms obtained from SECDI were acquired at 28 months, or prior. However, due to the lack of available objective measures of AoA in Swedish, SECDI was deemed the best available option as a reference. Also, this potential concern is remediated through the additional validation against the two corpora. The observation that validity was variable across all categories of raters, however, suggests that other variables, unaccounted for in this study, may affect participant suitability in the subjective rating of AoA. Future research should further explore what demographic features make an individual appropriate for subjective ratings of AoA.

Conclusions

The current data provide insight into flaws of the current practice of obtaining subjective ratings of AoA, nonetheless confirming that some individuals are both valid and reliable in their ratings, even with a more fine-grained rating scale than is most often used. The results from this study suggest that familiarity with children is an important factor in determining an individual's appropriateness as a rater of AoA.

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

As such, the results of this study strengthen the hypothesis that familiarity with children in the midst of language development enhances validity and reliability in raters of AoA. Based on the current results, existing subjectively rated AoA norms may benefit from supplementing their data with information regarding the raters' familiarity with young children, if such information is available.

Acknowledgements. Our sincere thanks to all informants for participating in this survey. Thanks also to Eva Berglund and Mårten Eriksson for sharing their advice regarding the Swedish CDI data. This work was funded by the Swedish Research Council (VR 2015-01525).

References

Alario, F. X., & Ferrand, L. (1999). A set of 400 pictures standardized for French: norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, *31*(3), 531–52.

Allwood, J., Grönqvist, L., Björkberg, M., Ahlsen, E., & Ottesjö, C. (2000). The Spoken Language Corpus at the Linguistics Department, Gothenburg University. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, *1*(3).

Bakhtiar, M., Nilipour, R., & Weekes, B. (2013). Predictors of timed picture naming in Persian. *Behavior Research Methods*, *45*(3), 834–41.

Berglund, E., & Eriksson, M. (2000). Communicative development in Swedish children 16–28 months old: the Swedish Early Communicative Development Inventory–Words and Sentences. *Scandinavian Journal of Psychology*, *41*(2), 133–44.

Birchenough, J. M. H., Davies, R., & Connelly, V. (2017). Rated age-of-acquisition norms for over 3,200 German words. *Behavior Research Methods*, *49*(2), 484–501.

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), 73–9.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P. I. A., Madsen, T. O., & Basboll, H. (2008). Early vocabulary development in Danish and other languages: a CDI-based comparison. *Journal of Child Language*, *35*(3), 619–50.

Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 158–67.

Brysbaert, M. (2017). Age of acquisition ratings score better on criterion validity than frequency trajectory or ratings 'corrected' for frequency. *Quarterly Journal of Experimental Psychology*, *70*(7), 1129–39.

Brysbaert, M., & Ellis, A. W. (2015). Aphasia and age of acquisition: Are early-learned words more resilient? *Aphasiology*, *30*(11), 1–24.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*(C), 80–4.

Carroll, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture naming latency. *Quarterly Journal of Experimental Psychology*, *25*(1), 85–95.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Wier, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development*, *10*, 159–99.

Chronicler of Higher Education, The (2010). Who are the undergraduates? *The Chronicler of Higher Education*. Retrieved 22 November 2017 from <http://www.chronicle.com/article/Who-Are-the-Undergraduates-/123916/>.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–20.

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical- decision performance above and beyond 22 other predictor variables: an analysis of 2,342 words. *Quarterly Journal of Experimental Psychology*, *60*(8), 1072–82.

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

Cuetos, F., Herrera, E., & Ellis, A. W. (2010). Impaired word recognition in Alzheimer's disease: the role of age of acquisition. *Neuropsychologia*, *48*(11), 3329–34.

Daland, R. (2013). Variation in the input: a case study of manner class frequencies. *Journal of Child Language*, *40*(5), 1091–122.

Della Rosa, P., Catricalà, E., Vigliocco, G., & Cappa, S. (2010). Beyond the abstract–concrete dichotomy: mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods*, *42*(4), 1042–8.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (pp. 2992–5). Valetta, Malta. Online <http://www.speech.kth.se/prod/publications/files/3399.pdf>.

Eriksson, A. (2004). SweDia 2000: a Swedish dialect database. In P. J. Henrichsen (Ed.), *Babylonian confusion resolved: proceedings of the Nordic Symposium on the Comparison of Spoken Languages* (Copenhagen Working Papers in LSP 1) (pp. 33–48).

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Reserach in Child Development*, *59*(5), 1–185.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: user's guide and technical manual*, 2nd ed. Baltimore, MD: Brookes.

Ferrand, L., Bonin, P., Méot, A., Augustinova, M., New, B., Pallier, C., & Brysbaert, M. (2008). Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic French words and their relation with other psycholinguistic variables. *Behavior Research Methods*, *40*(4), 1049–54.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: an open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–94.

Gierut, J. A., & Dale, R. A. (2007). Comparability of lexical corpora: word frequency in phonological generalization, *Clinical Linguistics & Phonetics*, *21*(6), 423–33.

Gilhooly, K. J., & Gilhooly, M. L. M. (1980). The validity of age-of-acquisition ratings. *British Journal of Psychology*, *71*(1), 105–10.

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), 515–31.

Hansen, P. (2017). What makes a word easy to acquire? The effects of word class, frequency, imageability and phonological neighbourhood density on lexical development. *First Language*, *37*(2), 205–25.

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*, 978–90.

Landis, R. J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–74.

Law, J., & Roy, P. (2008). Parental report of infant language skills: a review of the development and application of the Communicative Development Inventories. *Child and Adolescent Mental Health*, *13*(4), 198–206.

Lind, M., Simonsen, H. G., Hansen, P., Holm, E., & Mevik, B.-H. (2015). Norwegian Words: a lexical database for clinicians and researchers. *Clinical Linguistics & Phonetics*, *29*(4), 276–90.

Łuniewska, M., Haman, E., Armon-Lotem, S., 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*(3), 1154–77.

MacWhinney, B. (2000). *The CHILDES Project: tools for analyzing talk*, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Madsen, H. B., & Kim, J. H. (2016). Ontogeny of memory: an update on 40 years of work on infantile amnesia. *Behavioural Brain Research*, *298*(A), 4–14.

Martin, J. A., Hamilton, B. E., Osterman, M. J. K., Driscoll, A. K., & Mathews, T. J. (2017). *Births: final data for 2015*. (National Vital Statistics Reports, 66(1)). Hyattsville, MD: National Center for Health Statistics.

https://doi.org/10.1017/S0305000918000363 Published online by Cambridge University Press

Moors, A., Houwer, J., Hermans, D., Wanmaker, S., Schie, K., Harmelen, A.-L. … Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, *45*(1), 169–77.

Moreno-Martínez, F., Montoro, P., & Rodríguez-Rojo, I. (2014). Spanish norms for age of acquisition, concept familiarity, lexical frequency, manipulability, typicality, and other variables for 820 words from 14 living/nonliving concepts. *Behavior Research Methods*, *46*(4), 1088–97.

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*, 528–59.

Mukaka, M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*(3), 69–71.

Raman, I., Raman, E., & Mertan, B. (2014). A standardized set of 260 pictures for Turkish: norms of name and image agreement, age of acquisition, visual complexity, and conceptual familiarity. *Behavior Research Methods*, *46*(2), 588–95.

Rayson, P., & Garside, R. (2000). *Comparing corpora using frequency profiling*. Paper presented at the Proceedings of the Workshop on Comparing Corpora, Hong Kong. Online <http://ucrel.lancs.ac.uk/people/paul/publications/rg_acl2000.pdf>.

Schröder, A., Gemballa, T., Ruppin, S., & Wartenburger, I. (2012). German norms for semantic typicality, age of acquisition, and concept familiarity. *Behavior Research Methods*, *44*(2), 380–94.

Skolverket (2016). Statistik om förskolan [Preschool statistics]. Retrieved 22 November 2017 from <https://www.skolverket.se/statistik-och-utvardering/statistik-i-tabeller/forskola>.

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

Statistiska Centralbyrån (SCB) (2018a). Antal studenter i högskoleutbildning på grundnivå och avancerad nivå efter universitet/högskola, område/ämne, kön, ålder och läsår [Number of students in higher education at undergraduate and graduate level, by university, study topis, gender, age and academic year]. Retrieved 4 June 2018 from <http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__UF__UF0205/RegLasarN/table/tableViewLayout1/?rxid=f45f90b6-7345-4877-ba25-9b43e6c6e299>.

Statistiska Centralbyrån (SCB) (2018b). Medelåldern vid barnets födelse efter ordningsnummer, region och kön. År 2000–2017 [Mean age at child's birth, by order number, region and gender. Year 2000–2017]. Retrieved 4 June 2018 from <http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__BE__BE0101__BE0101H/MedelAlder1/?rxid=f45f90b6-7345-4877-ba25-9b43e6c6e299>.

Stoel-Gammon, C. (2011). Relationships between lexical and phonological development in young children. *Journal of Child Language*, *38*(1), 1–34.

Strömqvist, S., Richthoff, U., & Andersson, A.-B. (1993). Strömqvist's and Richthoff's corpora: a guide to longitudinal data from four Swedish children. *Gothenburg Papers in Theoretical Linguistics*, *66*.

Toppelberg, C. O., & Shapiro, T. (2000). Language disorders: a 10-year research update review. *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(2), 143–52.

Vihman, M. (2014). *Phonological development: the first two years*. Malden, MA: Wiley-Blackwell.

World Medical Association (WMA) (2013). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*, *310*(20): 2191–4.

Zevin, J., & Seidenberg, M. (2004). Age-of-acquisition effects in reading aloud: tests of cumulative frequency and frequency trajectory. *Memory & Cognition*, *32*(1), 31–8.

Cite this article: Wikse Barrow C, Nilsson Björkenstam K, Strömbergsson S (2019). Subjective ratings of age-of-acquisition: exploring issues of validity and rater reliability. *Journal of Child Language* **46**, 199–213. https://doi.org/10.1017/S0305000918000363