AGE-OF-ACQUISITION RATINGS FOR 2816 DUTCH
FOUR- AND FIVE-LETTER NOUNS

Mandy GYSELINCK, Wendy DE MOOR, and Marc BRYSBAERT
Ghent University

Studies on object and word naming have shown that the age at which words are acquired is an important factor in processing times. Research on the issue in Dutch has been hampered by the fact that only teacher ratings were available about which words should be known by 4-year-olds. As a supplement to these teacher ratings, we conducted a large-scale study in which 550 students rated the age-of-acquisition of 2816 four- and five-letter nouns. Reliability of the ratings is high, and correlations with word frequency and word inselagility are in the same order as those reported for English.

Virtually all psycholinguistic researchers agree that word frequency (i.e., the number of times a person is likely to come across a particular word) is an essential variable in word processing. High frequency words are easier to process than low frequency words. This is true for all sorts of word processing tasks (e.g., word naming, lexical decision, perceptual identification) and no model of word recognition has a chance of being accepted in the literature if it does not account for the frequency effect (for a review, see Mousmou, 1991). Also, experiments on the effects of other variables in word processing are unlikely to be published if word frequency has not been taken into account. As a result, researchers have invested major efforts to collect frequency norms for their language. In Dutch, for a long time the frequency norms of Uit den Bogaart (1975) were used. These were based on a corpus of 720,000 written words. Nowadays, the Uit den Bogaart corpus has been replaced by the electronic Cellex Database (Bazeyre, Piepenbrok, & Van Rijn, 1993), which is based on a corpus of 42,360,000 written words. Similarly, in English the old Kucera and Francis (1967) measures are currently being replaced by the Cobuild frequencies from the Cellex Database. These frequency measures have their limitations (e.g., they are nearly all based on written corpora; and there are always choices to be made about which texts to include, which types of word derivatives to combine in the frequency measures, etc.), but in general it is thought that the

This research was made possible by a BOF-project granted to the authors by the Research Council of Ghent University. We thank the anonymous reviewers for their helpful suggestions on an earlier version of the manuscript.

Correspondence concerning this article should be addressed to Mandy Gyselinck, Department of Experimental Psychology, Ghent University, E. Deurne 2, 9000 Ghent. Electronic mail may be sent to mandy.gyselinck@rug.ac.be
existing measures are reliable enough for practical purposes (at least in Dutch and English) and that further gains would not outweigh the efforts needed to collect additional data.

During the last five years or so, evidence is rapidly growing that the robust frequency effects in word processing tasks are actually a compound of two variables: the frequency with which words are encountered in texts (i.e., the usual integration of the word frequency effect), and the age at which the words have been acquired by individuals. Although it is true that most high frequency words have been acquired early in life and most low frequency words have been acquired late in life, giving rise to a strong positive correlation between word frequency and word age-of-acquisition (see below), there are exceptions (e.g., in English puppy is a low frequency word that is known by infants, whereas income is a high frequency word that is unlikely to be known by children). Using such words, Morrison and Ellis (1995) were able to disintegrate the effects of word frequency and word age-of-acquisition (AoA), and they presented evidence that at least part of the usual frequency effect in word naming and lexical decision is due to AoA. These results have been replicated by Gerhard and Barry (1998, 1999a) and Turner, Valentine and Ellis (1998) for English by Yamazaki, Ellis, Morrison, and Lamber-Ralph (1997) for Japanese, and by Brysbaert (1996) and Brysbaert, Lange, and Van Wijndendeale (2000) for Dutch.

Inspired by these findings, authors have started to examine AoA effects in other tasks. The importance of word AoA in addition to word frequency had already been suggested a long time ago for picture naming (Carroll & White, 1973), and new experiments with better controlled materials have indeed established that a large proportion of the variability in picture naming latencies is due to the age at which the object names have been acquired. This is not only true for English (Barry, Morrison, & Ellis, 1997; Ellis & Morrison, 1998), but also for Spanish (Cuetos, Ellis, & Alvarez, 1999) and French (Alario & Ferrand, 1999). Using a speeded naming task in which participants were instructed to name visual stimuli as fast as they could, Gerhard and Barry (1999b) showed that the effect of AoA is stronger than in a normal naming task. Investigating the importance of AoA in the semantic system, van Loon-Nervoorn (1998) and Brysbaert, van Wijndendeale, and De Deyne (in press) reported significant AoA effects in a word association task in which participants were asked to produce the first associate that came to their mind when they saw a stimulus word. Brysbaert et al. additionally reported an effect in a task in which participants had to decide whether a word belonged to the category of nouns with a definable meaning or to the category of first names. Finally, Lewis (1999) reported independent effects of frequency and AoA in a face categorisation task. In this task, participants had to indicate to which TV-soap pictures of characters belonged. The variables that were manipulated were the time since the first appearance in the soap (AoA) and the average exposure time during an episode (frequency). Both variables had an effect.

Other research has indicated that AoA is unlikely to be a confound of a third variable. For instance, both Brysbaert, Lange, and Van Wijndendeale (2000) and Coltheart, Laxon, and Kening (1988) showed that the AoA effect in visual word processing is not due to the imagability of the words. Although AoA and imagability are intercorrelated (see below), the effect of AoA on word naming and lexical decision remained significant when stimulus lists were matched on both word frequency and word imagability.

Ghyzelink and Brysbaert (submitted) examined the correlations between word familiarity, word frequency, and word AoA, in an attempt to solve a long-lasting debate in the literature on word recognition. In 1984, Gernsbacher showed that the subjective measure of word familiarity (based on the question: "how often have you come across this word?") explained significantly more variance in word processing times than word frequency. Ever since, there has been a debate about what exactly word familiarity measures in addition to word frequency (see e.g., Balota, 1994). Ghyzelink and Brysbaert's series of experiments strongly suggest that the missing variable is AoA, as the variance in word processing times explained by word familiarity equals the variance explained by word frequency plus the variance explained by word AoA.

Finally, there are theoretical reasons to expect AoA as a crucial variable in word processing in addition to word frequency. Lewis (1999), for instance, pointed out that AoA and frequency are two different measures of how often a person has encountered a particular word. The amount of experience with a word that has been acquired early in life will on the average be greater than the amount of experience with a word that has been acquired only recently. According to Lewis, it may very well be that not only the recent experiences with a word counts, but the cumulative frequency of all encounters with a particular stimulus (which would be evidence for an instance-based organisation of the mental lexicon). This line of reasoning agrees with Ghyzelink and Brysbaert's finding of the relationship between word familiarity, word frequency, and word AoA. On the basis of simulations with connectionist networks, studies by Ellis and Lamber-Ralph (in press) reported that the effect of AoA is likely to go beyond mere cumulative frequency. They argued that words learned first by a network have a privileged status in the network because they can be implemented in all possible units and connections between units, and because early training makes for larger weight changes than later learning. The activation function in a typical connectionist network follows a sigmoid curve with small changes towards the extremes and large changes in the middle of the curve.
Due to the characteristics of the words that are learned, and the mechanism by which they are stored in memory, it is likely that differences in frequency, word length, and word type are reflected in the performance of children learning to read. The frequency of words is an important factor in the acquisition of reading skills, and as a result, the frequency effect has been studied extensively. This bias of frequency of occurrence in the acquisition of new words is known as the frequency effect. This effect suggests that the frequency effect will be present in all processing stages, which are described in the next section.

Different AQA Measures

A first problem with AQA measures is that words are unlikely to be defined in the same way by different researchers. For example, a word like "cat" might be defined as a "noun" in some cases, but as an "animal" in others. This ambiguity can lead to inconsistencies in the results, as different researchers may interpret the same word differently.

The second problem with AQA measures is that the measure is not standardized. This means that different researchers are likely to use different definitions of AQA, which can lead to inconsistencies in the results. For example, some researchers might use the frequency of a word in a particular language, while others might use the frequency of a word in a particular domain (e.g., science, literature, etc.).

Finally, a language typically consists of a few tens or thousands of words. This means that the frequency of a word in a language is likely to be influenced by a variety of factors, including the age of the child, the language they are learning, and the frequency of the word in their environment. As a result, the frequency effect is likely to be complex and difficult to measure accurately.

Different AQA Measures

A second problem with AQA measures is how to define the population. For example, it is unclear whether AQA measures are applicable to all children, or only to those who are learning to read. This is important because the frequency effect is likely to vary across different populations, and different measures may be needed to accurately assess the frequency effect in different groups.
of 13 or more (see Gilhooly & Logie, 1980). This resulted, for example, in a score of 1.80 for apple. The correlation between both AoA measures was 0.75. In addition, further research showed that both measures had virtually the same correlation with object naming times (Ellis & Morrison, 1998). On the basis of these and other studies (e.g., Gilhooly & Gilhooly, 1980; Lyon, Tier, & Rubenstein, 1978), Morrison et al. (1997) concluded that student ratings are a valid measure of AoA (see also De Moor et al., 2000). Therefore, most work on AoA effects in English has relied on the student ratings published by Gilhooly and Logie (1980). These authors asked 36 students to rate the AoA of different words on the 7-point scale described above. To assess the reliability of their ratings, Gilhooly and Logie divided the participants randomly in two groups, balanced for sex, and found an intergroup correlation of .98.

The studies done in Dutch so far all made use of a different AoA measure. In 1981, Kohnstamm, Schaerlaeckens, de Vries, Akkerhuis, and Froonincx published a book that indicated for 6,785 words to what extent each word should be known by 6-year-olds (see also Schaerlaeckens, Kohnstamm, & Leijenburger, 1999, for a very recent update of the list). The estimates were obtained by asking a representative sample of teachers from the Netherlands and the Dutch speaking part of Belgium to mark for each word whether it should be known by a 6-year-old or not. For the Belgian data, 40 teachers of kindergarten and 41 teachers of the first year of primary school from all over the Dutch speaking region took part in the assessment. The advantages of the Kohnstamm et al. measure, defined as the percentage of teachers indicating that a word should be known by a pupil who starts primary school, are that the measure is based on persons who have daily experience with the children they are judging, and that the measure has been obtained some 20 years ago (when the undergraduates of current studies were born). In addition, van Loon-Vervoorn (1989) obtained a correlation of .92 between the Kohnstamm et al. measure and student ratings on an 8-point scale, based on 44 nouns. The Kohnstamm et al. ratings have proven very useful in several studies, but they are limited because they only provide information about one moment in time: the transition from kindergarten to primary school. There is no information about whether a word that is known by a 6-year-old, has been acquired at the age of 2 or at the age of 5; similarly, a word that is not known at the age of 6 according to the teacher ratings, can be acquired at the age of 7 or at the age of 15 and later. Also, the different measures used in English and in Dutch studies make it difficult to interpret deviating findings between both languages. Therefore, we decided to collect student ratings of AoA for the Dutch language as well.

Method

Stimulus Materials and Rating Procedure

We selected 2816 four- and five-letter nouns from the CELEX Database (Baayen et al., 1993). The selection was limited to four- and five-letter nouns for practical reasons and because we typically use words of these lengths in our studies of visual word processing. Nouns were chosen partly because they can be used in picture naming experiments. Words with a frequent non-

noun interpretation and nouns with multiple frequent meanings or with a frequency less than 1 per 42,380,000 were excluded.

Of the stimulus set, we created 10 lists of 281 or 282 words that were matched on frequency and word length. Because our previous experience had shown that rating scales like Gilhooly and Logie’s (1980) were sometimes confusing for the participants, we simply asked them to indicate for each word from which age they estimated they knew the word. If they did not know the word, they could write an “N”. Of each list, three different permutations were made to minimise sequence effects. The lists were handed out at the beginning of a course, and completion of a list took about half an hour. Each participant completed but one list.

Participants

Participants were 558 undergraduates (310 females and 248 males). They were students from the faculties Political and Social Sciences, Criminal Sciences, Philosophy or Moral Sciences at Ghent University. All were native Dutch speakers. Average age was 19 (range 16-42).

Results and Discussion

The AoA data are shown in the Appendix. The full matrix of AoA, frequency, Log(frequency) and % of answers is available on the internet: http://cellerm.ugent.be/a-skusen/savkgوء/ (Research, available documents, data, etc.).

For each list we computed the correlation between the individual ratings and the mean AoA measures. Sixteen participants who correlated less than .68 with the means were excluded from further analyses. All in all, the minimal number of raters per list was 50. To assess the reliability of our ratings, we calculated the intraclass correlation of Shout and Fleiss (1979). For this analysis, words that were not known by at least 80% of the
participants were excluded and the missing AOA values for the remaining words were estimated on the basis of the means of the rows and the columns. The reliability of the individual rating scale varied from .85 to .98, and the reliability of the overall rating scale for the materials varied from .85 to .98.

To further check the reliability of our AOA measure, we attempted to replicate the results of University of London tests reported by Coltheart et al. (1992) with a larger database of word meanings. The results of the replication were consistent with the previous findings.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.

The reliabilities of the rating scales were estimated using the split-half method. The split-half reliabilities ranged from .85 to .98, which is nearly the maximum possible reliability. This analysis shows that the reliabilities of the rating scales were high and consistent.
DUTCH AGE-OF-ACQUISITION NORMS


