# Supplementary materials

## RPVT Word Selection

The word prompts for the RPVT were gathered from three sources. The first was Parent's (2012) list of the 75 most common English homonyms. The second was Wikipedia's list of 248 true homonyms. The third was screening for word meanings of polysemous words with a grade of acquisition of grade 2 to 6 (Brysbaert & Biemiller, 2017). Words selected conformed to six criteria. Words had to be either true homonyms (i.e. both spelt and pronounced the same) or polysemous words with at least two clearly distinguishable meanings (as identified from Brysbaert and Biemiller, 2017). The estimated grade of acquisition (from Brysbaert and Biemiller, 2017) of the earliest acquired meaning of the word had to be 8th grade or lower. Both meanings had to be unlikely to have different interpretations of meaning across dialects of English. For each word, both meanings were appropriate for children (i.e. no references to drugs or violence), and neither of the meanings of the word were judged to be very infrequent (e.g. ear was excluded, because ear as in a unit of corn is a highly infrequent meaning, Parent, 2012). Both meanings of the word also needed to be easily visualised in images. Thus, for example, the word step was not included, because step as in ‘step-mother’ is difficult to visualise. Items were created for the pool of selected prompt words and these were piloted to identify the most suitable for the final test.

Limited data were available on the age at which children acquire the meanings of polysemous words, consequently, two pilot studies were conducted with children from the target population to select the most effective word stimuli from a larger pool. The first pilot was conducted with 37 children (18 female) in year groups 2 and 5 (aged 6;2-10;1, *M* = 7.85, *SD* = 1.51), 11% of whom had EAL, and an initial set of 50 items. Due to time constraints, each child completed one half of the items. Following this, some ineffective items were removed, either due to being too easy (*n*=5) or correlating poorly with the overall score (*n* =2). Some new items were added (*n* =9), which were anticipated to be more difficult (due to a higher estimated grade of acquisition according to Brysbaert et al, 2017). Thus, a test with 52 items was created. A second pilot was then conducted with 26 children (10 female) in year groups 1 and 4 (aged 5;3- 9;0, *M* = 7.00, *SD* = 1.46), 39% of whom had EAL. Items which showed poor correlations with total scores and/or low variability based on all child pilot data were excluded (*n* =8).

To ensure that final test items were comprehensible to adults, a further pilot was conducted with adult participants. An opportunity sample of 34 EL1 adults (Age: Range= 20-69, *M* = 31.35, *SD* = 10.9; 24 female) completed the test. Participants were sent a link to the test and completed it online via Qualtrics (Qualtrics Labs Inc., Provo, UT). Participants received 54 items, which consisted of the 52 from the second pilot, plus two easier items previously removed. Items on which adults showed any difficulty (scoring lower than 95% accuracy, *n* = 8) were excluded, to ensure that items with ambiguous images were removed and that children were being judged against a fair standard, that is performance of EL1 adults. Remaining items were ranked in order of difficulty based on the child data and a subset selected to represent a range of difficulty levels, to create a test of 30 items.

## RPVT Stimuli Psycholinguistic Characteristics

Imageability ratings were obtained for each of the two meanings of the word. Adults with EL1 (*N* = 40) aged 19 to 63 years completed imageability ratings for each of the 60 word meanings on a scale from 1 (low imageability) to 7 (high imageability). The average imageability rating across the word meanings was 5.21 out of 7 (*SD* = 1.07, range: 2.22 to 6.72). Thus, the word meanings were generally of moderate to high imageability. Imageability was not significantly different between primary (*M* = 5.30, *SD* = 1.09) and secondary (*M* = 5.11, *SD* = 1.06) meanings (*t* (29) = 0.67, *p* = .51).

Adult ratings of meaning relatedness were also obtained following methods from previous research (Rodd, Gaskell & Marslen-Wilson, 2002; Azuma & Van Orden, 1997). The same adults (*N* = 40) who completed imageability ratings subsequently rated each of the 30 words from the polysemy test on the similarity of its two meanings, on a scale from 1 (not at all related) to 7 (highly related). The average relatedness rating across the words was 2.81 out of 7 (*SD* = 1.10, range: 1.43 to 4.59). Thus, the word meanings were generally perceived to be only slightly related.

Dominance, or the psychological salience of one meaning of a word over another, can also be estimated for the stimuli. One way to estimate this is to calculate the difference in children’s score between the primary and secondary meanings. Dominance can theoretically range from 0 to 1. The average of the dominance scores was 0.30 (*SD*= 0.23, range: 0.00 to 0.79). Thus, the words varied in the dominance of their first meaning and showed low to moderate dominance overall.