SUPPLEMENTARY MATERIALS: METHODS AND RESULTS

This document provides supplementary methods and results material for “Validation of two measures for assessing English vocabulary knowledge on web-based testing platforms: Long-form assessments.”

1. Sample size determination and participant exclusions

As described in the main text, Experiment 1 (Vocabulary Size Test, VST) and Experiment 2 (Word Familiarity Test, WordFAM) each tested a different sample of 100 participants. The sample size (n = 100 in each experiment) was determined by convention; specifically, this sample size reflects approximately 2–3 times our evaluation of the common sample sizes in between-subjects conditions in the psycholinguistic research domain. This sample size exceeds 80% power to detect an effect size of the magnitude \( r = 0.30 \).

One additional participant was tested in Experiment 1 but excluded from analysis due to failure to perform the task as directed. This participant showed a total completion time of less than one minute, with mean accuracy (0.19 proportion correct) near chance (0.25, given the four-alternative, forced-choice design of the assessment). Two additional participants were tested in Experiment 2 and subsequently excluded from analysis due to failure to perform the task as directed. One of these participants showed reaction times of less than 50 ms for most of the trials; the other participant showed a flat response function with many consecutive strings of repeated ratings (e.g., ratings of 3, 4, 5, 3, 4, 5, 3, 4, 5 over a series of trials).

2. Modifications to the VST

As described in the main text, stimuli for the VST consisted of the 100 items on Form A of the monolingual (20,000) version of the VST (available at https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/VST-version-A.pdf). The spelling of one item (yoghurt) was changed to reflect American English spelling conventions (i.e., yogurt), as was one of the response options (group of players gathered round the ball in some ball games was changed to group of players gathered around the ball in some ball games; underline added here for clarity). Aside from these two cases, the stimuli in the present study were identical to those of the original VST.

3. Analysis to examine accuracy as a function of frequency bin for the VST (Experiment 1)

To examine accuracy as a function of frequency bin, trial-level responses (0 = incorrect, 1 = correct) were submitted to a generalized linear mixed effects model with the binomial response family as implemented with the glmer() function of the lme4 package in R. Frequency bin was entered into the model as a fixed effect, coded to reflect sliding contrast comparisons (i.e., low vs. mid-low, mid-low vs. mid-high, mid-high vs. high). The random effects structure consisted of random intercepts by subject, random slopes for frequency bin by subject, and random intercepts by item. The results of the model showed no significant change in accuracy between the low and mid-low bins (\( \hat{\beta} = 0.859, SE = 0.537, z = 1.600, p = 0.110 \)), and monotonic improvement in accuracy from the mid-low bin to the mid-high bin (\( \hat{\beta} = 1.882, SE = 0.558, z = 3.372, p < 0.001 \)) and from the mid-high bin to the high frequency bin (\( \hat{\beta} = 1.327, SE = 0.592, z = 2.240, p = 0.025 \)).

4. Analysis to examine ratings as a function of frequency bin for the WordFAM (Experiment 2)

To examine familiarity ratings as a function of frequency bin, trial-level ratings were submitted to a linear mixed effects model as implemented with the lmer() function of the lme4 package in R. Frequency bin was entered into the model as a fixed effect, coded to reflect sliding contrast comparisons (i.e., low vs. middle, middle vs. high). The random effects structure consisted of random intercepts by subject, random slopes for frequency bin by subject, and random intercepts by item. The results of the model showed a monotonic increase
in ratings from the low to middle frequency bin ($\hat{\beta} = 1.204, SE = 0.221, t = 5.455, p < 0.001$) and from the middle to high frequency bin ($\hat{\beta} = 2.568, SE = 0.227, t = 11.314, p < 0.001$).

5. Analysis to compare WordFAM performance between the Prolific and Hoosier samples (Experiment 2)

Mean item ratings were submitted to ANOVA with the between-subjects factor of frequency bin and the within-subject factor of sample. Because this is a by-items analysis, item is treated as subject. Accordingly, frequency bin is a between-subjects factor (a given item can only be in one frequency bin) and sample is a within-subject factor (because each item is present in both samples). The ANOVA revealed the expected main effect of frequency bin [$F(2, 147) = 476.46, p < 0.001$]. There was no main effect of sample [$F(1, 147) = 0.02, p = 0.888$]. However, there was a significant interaction between frequency bin and sample [$F(2, 147) = 4.71, p = 0.010$]. To explain the interaction, paired $t$-tests examined the difference between the two samples for each frequency bin. There was no significant difference between the two samples for the low frequency bin [$t(49) = -1.87, p = 0.068$] or the middle frequency bin [$t(49) = 1.25, p = 0.219$]. For the high frequency bin, the Hoosier sample showed a slightly higher mean rating [mean difference = 0.21, $t(49) = 2.923, p = 0.005$].