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: Brief assessments.”

1. Sample size determination and participant exclusions

As described in the main text, the sample size was determined by the sample size used in Drown et al. (Under review). As described in the main text, the results reflect performance of 85 participants who completed both test sessions. Five additional participants completed both sessions but were excluded from analysis due to failure to perform the task as directed, as evident by exhibiting most reaction times less than 50 ms and/or pressing only one or two buttons for an entire task.

2. Creation of the brief versions of the VST and WordFAM assessments

For the VST, each brief version contained 42 of the original VST items, representing 10-11 items in each of the four lexical frequency bins (low, mid-low, mid-high, high). First, items from the full assessment were assigned to either the A or B version as described for the split-half reliability analysis reported in Drown et al. (Under review; i.e., odd-numbered items were assigned to the A version, even-numbered items were assigned to the B version). Second, two items were removed from each frequency bin for each version based on the by-item distribution of accuracy means reported in Drown et al (Under review); specifically, we removed items that deviated most substantially from the median item accuracy for a given bin. The Supplementary Material (https://osf.io/pcsu6/) provides item lists for each brief version. This procedure yielded Brief-A and Brief-B versions of the VST that contained an equal number of items, equivalent sampling of the original items across frequency bins, and mutually exclusive items between the two brief assessments.

Creating the brief versions of the WordFAM assessment followed a similar protocol. Each brief version contained 72 of the original WordFAM items, including 22 items in the low frequency bin and 25 items in each of the middle and high frequency bins. First, items from the full assessment were assigned to either the A or B version as described in Drown et al. (Under review; i.e., alternate assignment of items to either the A or B version based on ordered familiarity rating of the original assessment). Second, three items were removed from the low frequency bin for each version based on the by-item distribution of rating means from Drown et al. (Under review). Specifically, we removed items that deviated most substantially from the median item rating for the low frequency bin. The Supplementary Material (https://osf.io/pcsu6/) provides item lists for each brief version. As for the VST, this procedure yielded Brief-A and Brief-B versions of the WordFAM that contained an equal number of items, equivalent sampling of the original items across frequency bins, and mutually exclusive items between the two brief assessments.

3. Analysis to examine accuracy as a function of frequency bin for the VST

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. The model included fixed effects of frequency bin and test version. 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). Test version was entered into the model as a mean-centered contrast (Brief-A = -0.5, Brief-B = 0.5). The random effects structure consisted of random intercepts by subject, random slopes for frequency bin and version by subject, and random intercepts by item. The results of the model showed no significant difference in accuracy between the low and mid-low frequency bins ($\hat{\beta} = 0.631$, $SE = 0.362$, $z = 1.745$, $p = 0.081$), and a monotonic increase in accuracy between the mid-low and mid-high frequency bins ($\hat{\beta} = 2.224$, $SE = 0.388$, $z = 5.735$, $p < 0.001$) and the mid-high and high...
4. Analysis to examine ratings as a function of frequency bin for the WordFAM

To examine 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. The model included fixed effects of frequency bin and version. 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). Version was entered into the model as a sum-coded contrasts (Brief-A = -0.5, Brief-B = 0.5). The random effects structure consisted of random intercepts by subject, random slopes for frequency bin and version by subject, and random intercepts by item. The results of the model showed a monotonic increase in familiarity ratings from the low to middle frequency bin ($\hat{\beta} = 1.580, SE = 0.182, t = 8.665, p < 0.001$) and from the middle to high frequency bin ($\hat{\beta} = 2.515, SE = 0.184, t = 13.673, p < 0.001$). The main effect of version was not significant ($\hat{\beta} = 0.131, SE = 0.148, z = 0.887, p = 0.377$), nor did version interact with frequency bin ($p \geq 0.067$ for all three interaction contrasts).

5. Analysis to compare WordFAM performance between the Prolific and Hoosier samples

Mean item ratings were submitted to an ANOVA with the between-subjects factor of frequency bin and version and the within-subject factor of sample. Because this is a by-items analysis, item is treated as subject. Accordingly, frequency bin and version are between-subjects factors (i.e., a given item can only be in one frequency bin or one test version) and sample is a within-subject factor (i.e., each item is present in both samples). The ANOVA revealed the expected main effect of frequency bin [$F(2, 138) = 571.52, p < 0.001$]. There was no main effect of sample [$F(1, 138) = 0.88, p = 0.349$], no main effect of version [$F(1, 138) = 0.54, p = 0.463$], nor were any of the interactions statistically reliable ($p \geq 0.199$ in all cases).