Validation of two measures for assessing English vocabulary knowledge on web-based testing platforms

Lee Drown,[1,2] Nikole Giovannone,[1,2] David B. Pisoni[3] and Rachel M. Theodore[1,2]

[1]Department of Speech, Language, and Hearing Sciences

University of Connecticut

[2]Connecticut Institute for the Brain and Cognitive Sciences

University of Connecticut

[3]Psychological and Brain Sciences

Indiana University

Author to whom correspondence should be addressed:

Rachel M. Theodore

Department of Speech, Language, and Hearing Sciences; University of Connecticut

2 Alethia Drive, Unit 1085

Storrs, CT 06269-10825

rachel.theodore@uconn.edu

(860) 486-3477

Abstract

The goal of the current work was to develop and validate web-based measures for assessing English vocabulary knowledge. Two existing paper-and-pencil assessments, the Vocabulary Size Test (VST) and the Word Familiarity Test (WordFAM), were modified for web-based administration. In Experiment 1, participants (n = 100) completed the web-based VST. In Experiment 2, participants (n = 100) completed the web-based WordFAM. Results from these experiments confirmed that both tasks (1) could be completed online, (2) showed expected sensitivity to English frequency patterns, and (3) revealed high split-half reliability, suggesting that stable vocabulary assessment could be achieved with fewer test items. Based on the results of Experiments 1 and 2, two "brief" versions of the VST and WordFAM were developed. Each version consisted of approximately half of the items from the full assessment, with novel items presented across the two brief versions of each assessment. In Experiment 3, participants (n = 85) completed one brief version of both the VST and WordFAM at session one, followed by the other brief version of each task at session two. The results showed high test-retest reliability for both the VST (r = 0.68) and WordFAM (r = 0.82). The two brief assessments also showed moderate convergent validity (ranging from r = 0.38 to r = 0.59) indicative of construct validity for each assessment. This work provides open-source vocabulary knowledge assessments with normative data that researchers and clinicians can use to foster high quality data collection in web-based environments.

Introduction

Reliable, valid measures of language proficiency are useful for both the clinical and research domains. In clinical practice, such measures serve as a benchmark for initial diagnosis and for tracking rehabilitation outcomes. In research, measures of language proficiency can serve to describe the research sample, group participants based on a given proficiency, or serve as predictors for examining individual differences in performance. Vocabulary is a core aspect of linguistic knowledge (e.g., Bloom, 2002). Deficits in vocabulary expression and comprehension can contribute to adverse social and academic outcomes (e.g., Bleses et al., 2016; Irwin et al., 2002; Landi, 2010). For example, low vocabulary knowledge is linked to lower academic achievement in several domains, including written language comprehension and second language learning (e.g., Bleses et al., 2016; Landi, 2010; Mancilla-Martinez et al., 2014; Snow & Kim, 2007). Lower vocabulary production early in life is also related to higher rates of depression and social withdrawal, and weaker abilities to engage in social reciprocity and pretend play (Irwin, Carter & Briggs-Gowan, 2002). Vocabulary knowledge also has cascading effects on reading ability, which is a critical skill for children and adults alike (e.g., Wasik et al., 2016), and is related to other cognitive skills, including phonological working memory, word recognition, lexical access, and language comprehension (Gathercole & Baddeley, 1993; Lewellen et al., 1993; Tamati & Pisoni, 2014). More recently, vocabulary knowledge has been identified as a source of individual differences in other aspects of language processing, including perceptual learning for speech (Colby et al., 2018; Giovannone & Theodore, 2021; Rotman et al., 2020; Theodore et al., 2019).

Standardized assessments exist to measure English vocabulary proficiency, including the Clinical Evaluation of Language Fundamentals, Fifth Edition (CELF-5; Wiig et al., 2013), the

Expressive Vocabulary Test, Second Edition (EVT-2; Williams, 1997), and the Peabody Picture Vocabulary Test, Third Edition (PPVT-3; Dunn & Dunn, 1997). These assessments provide normative data that clinicians and researchers can use to quantify an individual's vocabulary knowledge relative to their peers. These assessments provide critical tools for clinicians and researchers; however, they are not without limitations. First, standardized assessments often require substantial training for administration, and some standardized assessments require administrators to possess an advanced and/or specialized degree (Wiig et al., 2013). Second, standardized vocabulary assessments can be long in duration, often exceeding 30 minutes for administration (Dunn & Dunn, 1997; Wiig et al., 2013; Williams, 1997). Third, many standardized vocabulary assessments – including the CELF-5, EVT-2, and PPVT-3 – only provide normative data for a limited age range (6 – 22 years of age). Though some aspects of language proficiency may meet a maturational ceiling by age 22 (e.g., speech sound proficiency), vocabulary knowledge in principle has no ceiling across the healthy lifespan (e.g., Kavé et al., 2010). Fourth, many standardized assessments are licensed by for-profit companies, which introduces a financial barrier to their access. Finally, most standardized assessments are designed to be administered in-person, with the administrator and participant in a shared physical space. Recently, in-person administration has been identified as a limitation of existing standardized assessments due to safety concerns resulting from the continuing COVID-19 pandemic and geographic constraints that may limit access to clinical services and research participation for individuals who reside in rural and other underserved areas.

Web-based technologies have the potential to address some of these limitations. For example, telepractice, or the remote administration of clinical services, has been approved by the American Speech-Language-Hearing Association, which allows audiologists and speech-

language pathologists to use virtual platforms for service delivery (American Speech-Language-Hearing-Association, n.d.). In addition, the rapid rise of web-based research technologies has provided researchers with tools that remove physical and geographical barriers in the research process (e.g., Anwyl-Irvine et al., 2020; Palan & Schitter, 2018). For example, when participation in a research study is not limited to a physical lab, participant samples can reflect diversity beyond a particular university community. However, remote administration of existing standardized vocabulary assessments is not possible for all those who may wish to use them due to the identified training and financial barriers. Moreover, not all existing vocabulary assessments transfer well to a web-based format, particularly for researchers who may wish to use these assessments for non-clinical purposes. For example, the CELF-5 requires multi-modal responses, including a combination of expressive (e.g., requests for repetition) and receptive (e.g., pointing, gestural imitation) responses. In addition, the normative data for the CELF-5 were established based on in-person administration and thus may not be applicable to virtual performance. Further, administration of the CELF-5 can take approximately 30 – 45 minutes. In the research domain, tests of this length may not be ideal given that researchers who use vocabulary assessments are likely to use them in conjunction with other experimental tests of interest.

As described above, web-based research has proliferated in recent years, reflecting the emergence of new tools for remote data collection. Though these tools show strong promise for the scientific process, some challenges remain, particularly for research that draws on anonymous participant pools (Godinho et al., 2020; Griffin et al., 2021; Palan & Schitter, 2018; Storozuk et al., 2020). Specifically, web-based research methods afford the possibility of automated enrollment in online studies by software applications, known as “bots.” Bots pose

threats to data integrity and thus the validity of conclusions drawn from web-based studies. Researchers need to be aware of the presence of bots on online data collection platforms and employ strategies during both the data collection and data analysis stages to prevent bots from diminishing the integrity of their studies (Godinho et al., 2020; Griffin et al., 2021; Storozuk et al., 2020). Even when an actual human may be completing a web-based study, concerns remain regarding whether self-reported demographic information is accurate. That is, researchers may have concerns that participants misrepresent demographic variables (e.g., native language) that may be critical to the integrity of a given study. In the psycholinguistic domain, language experience and proficiency are often foundational characteristics of the participant sample that are needed to accurately interpret research findings. In principle, standardized assessments of vocabulary knowledge could provide researchers with a means to verify self-reported language proficiency. For example, one would expect most individuals who report English as their native language and no history of speech or language disorders to exhibit a standardized vocabulary score within one standard deviation of the mean standard score of a given assessment. For the reasons described above, however, existing standardized assessments are not ideal for this purpose.

In this context, the goal of the current work was to develop and validate two web-based measures that assess English vocabulary knowledge. We aimed to meet four criteria for each measure. First, the assessment should be openly available for free and public re-use in the research domain. Second, the assessment should be fast and easy to complete without requiring real-time interaction between a participant and a researcher. Third, the assessment should yield acceptable psychometric properties including split-half reliability and test-retest reliability. Fourth, the two assessments should show strong convergent validity as an indicant of construct

validity for each assessment. Meeting these criteria would yield vocabulary assessments that could be used to verify self-reported language proficiency and/or examine vocabulary in web-based research. To meet this goal, we developed web-based versions of two existing paper-and-pencil assessments, the Vocabulary Size Test (Beglar & Nation, 2007) and the Word Familiarity Test (Lewellen et al., 1993; Pisoni, 2007), and then submitted the web-based versions to validation testing. Below we describe each of the existing assessments in turn, and then introduce the validation testing executed in the current work.

The Vocabulary Size Test (VST; Beglar & Nation, 2007) is a multiple-choice test designed to estimate an individual's English vocabulary size (Beglar, 2010; Beglar & Nation, 2007; Coxhead, 2016; Coxhead et al., 2015). The VST has numerous forms, including versions of various lengths for use with monolingual and bilingual individuals (Beglar & Nation, 2007; Coxhead et al., 2014, 2015). The current work adapted Form A of the 20,000 word families VST, available at https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/VST-version-A.pdf (Nation, 2012). This VST consists of 100 multiple-choice items that assess vocabulary from the twenty most frequent word families as they occur in the British National Corpus (Bauer & Nation, 1993). Each family consists of 1000 words, and five words from each of 20 families are presented. The first family represents the most frequent 1000 words in the corpus, the second family represents the second most frequent 1000 words in the corpus, and so on to the twentieth family. The assessment is designed to sample vocabulary knowledge across word families to infer an individual's vocabulary size. For example, if a person shows ceiling performance on the VST, then vocabulary size is estimated to be 20,000 words. Or, for example, if a person shows ceiling accuracy through the first 10 families and floor performance for the second 10 families, then the vocabulary size is estimated to be 10,000

words. In the paper-and-pencil version of the assessment, each word is presented in a neutral sentential context (e.g., cabaret: We saw the cabaret) followed by four response options (e.g., painting covering a whole wall; song and dance performance; small crawling creature; person who is half fish, half woman). Participants are directed to circle the option that best defines the target word.

The Word Familiarity Test (WordFAM; Lewellen et al., 1993; Pisoni, 2007) is a subjective word familiarity rating questionnaire. The WordFAM was developed based on normative data in the Hoosier mental lexicon corpus (Nusbaum et al., 1984). This corpus consists of word familiarity ratings for 19,750 English words that reflect a wide range of lexical frequencies. A total of 600 participants provided ratings for this corpus, with each subject providing ratings for 395 words. The rating scale ranged between 1 and 7, with 1 corresponding to "You have never seen or heard this word before" and 7 corresponding to "You recognize the word and are confident that you know the meaning of the word." The normative data in this corpus consist of the mean familiarity rating for each word as derived by collapsing across the 12 unique participants who rated each word.

Using the Hoosier mental lexicon corpus, the WordFAM was developed to sample 150 words from the corpus that span a wide range of normative familiarity ratings. Specifically, 50 items were selected to represent low, middle, and high familiarity words. The paper-and-pencil version of the WordFAM lists 150 words (in a single randomized order) next to the digits one through seven. Participants are asked to rate their familiarity with each word by circling the appropriate digit corresponding to the rating scale that is provided on the first page of the assessment. For a given participant, ratings can be averaged over the 150 words to calculate an overall word familiarity score; in addition, ratings can be averaged separately for the three

frequency bins to devise word familiarity scores for low, middle, and high frequency lexical items. Previous research has demonstrated that subjective word familiarity ratings predict confrontation word naming, lexical decision, and semantic categorization performance, suggesting that lexical familiarity is associated with language processing efficiency (Lewellen et al., 1993).

In some ways, the VST and WordFAM assessments are particularly well-suited for the current goal. Specifically, both tests are currently open access, do not require advanced training to administer or interpret, and lend themselves well to self-guided completion. They both make use of a lexical frequency manipulation to assess breadth of vocabulary knowledge, and past research provides some evidence to suggest that these assessments are valid measures of vocabulary knowledge (Beglar, 2010; Coxhead et al., 2014; Lewellen et al., 1993; Nusbaum et al., 1984; Tamati & Pisoni, 2014). Their differences, too, make these assessments ripe for joint consideration. That is, though both measures assess vocabulary knowledge, they do so in different ways. The VST is a closed-choice test with objectively correct answers, whereas the WordFAM elicits a subjective measure of perceived word familiarity. Together, these two vocabulary measures can provide a global picture of an individual's vocabulary knowledge, through both an objective and subjective lens.

However, the utility of the VST and the WordFAM could be enhanced through a better understanding of the psychometric characteristics of each assessment. Construct validity, or the degree to which a test measures what it is intended to measure, is necessary to ensure valid interpretation of performance (Anastasi & Urbina, 1997). Construct validity can be measured, at least in part, by assessing convergence between individuals' performance on two tasks intended to measure the same construct. Reliability, or the extent to which a task is a stable measure of a

given construct, can likewise be assessed in multiple ways (e.g., split-half reliability, test-retest reliability; Anastasi & Urbina, 1997). Split-half reliability offers insight into the internal consistency of a measure; test-retest reliability assesses the degree to which consistent results can be obtained each time the task is administered, thus promoting a better understanding of the measurement error intrinsic to the task.

Standardized tests used for clinical purposes are subjected to rigorous testing to ensure validity and reliability of construct measurement; however, the same is not true for many of the commonly used tasks in psycholinguistics and cognitive sciences research. Unknown task validity and reliability pose a formidable threat to the integrity of research in this domain; without an understanding of these properties, it is difficult to know how much of a participant's performance on a given task is related to characteristics of that participant versus characteristics of the task itself. A lack of convergent validity across tasks assumed to measure the same construct is a particular issue for research in cognitive sciences. Recent studies have demonstrated that many common tasks used in the domains of perceptual adaptation (Heffner et al., 2022), audiovisual integration (Wilbiks et al., 2022), and listening effort (Strand et al., 2018) are only weakly associated with each other despite being purported to measure the same underlying constructs. Thus, it is difficult for researchers to ascertain whether results found with one specific task are generalizable to other tasks or even the broader construct itself, severely limiting the scope of interpretation. The test-retest reliability of tasks used in the domain of cognitive sciences also a concerning issue. For example, a recent study of commonly used infant speech perception tasks assessed their test-retest reliability across 13 separate samples and found that across these samples, only three showed significant, positive associations across test sessions (Cristia et al., 2016). Without adequate test-retest reliability, researchers cannot be confident that

their task is measuring a stable trait of the test subject. Taken together, a firm understanding of the validity and reliability of a given measure gives clinicians and researchers alike the power to make the strongest claims possible given constraints introduced by the psychometric properties of a given test or task.

The goal of the current work was to develop and validate two web-based measures that assess English vocabulary knowledge. To this end, three experiments were conducted. Experiment 1 tested participants (n = 100) on a web-based administration of the VST and Experiment 2 tested a different group of participants (n = 100) on a web-based administration of WordFAM. In both experiments, analyses were conducted to determine the suitability of each measure for web-based platforms, including an analysis of split-half reliability as a potential means to reduce administration time while maintaining high validity. In Experiment 3, participants (n = 85) completed brief versions of both the VST and WordFAM at two points in time. Analyses were conducted to examine the stability of performance at the individual subject level over time and the degree to which performance on the two measures were associated.

Experiment 1

Methods

Participants. Participants (n = 100) were recruited from the Prolific participant pool (https://www.prolific.co; Palan & Schitter, 2018). The inclusion criteria (set using standard Prolific filters) were: monolingual English speaker, born in the United States, currently residing in the United States, between 18 and 35 years of age, and no history of language-related disorders. The resulting sample consisted of 47 men and 53 women who had a mean age of 25 years (SD = 6 years). Comprehensive demographic characteristics of the sample including race, ethnicity, and self-reported dialect are available in the Supplementary Material. One additional

participant was tested but was excluded from final analyses 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).

Stimuli. Stimuli 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/vocabulary-tests/the-vocabulary-size-test/VST-version-A.pdf). Each item consisted of a semantically-neutral prompt (e.g., *veer: The car veered*) and four response options (e.g., *moved shakily*, *changed course*, *made a very loud noise*, *slid without the wheels turning*). The 100 items were designed to sample five English words from each of 20 frequency categories. The categories ranged from extremely high frequency items (e.g., *see*, *time*) that would be expected to be known if a person has a small vocabulary (e.g., 1000 words) to extremely low frequency items (e.g., *sagacious*, *casuist*) that would be expected to be known if a person has a large vocabulary (e.g., 20,000 words). The 20 frequency categories of the VST are coded as groups that range from 1,000 (lowest frequency items) to 20,000 (highest frequency items) in 1,000 unit bins. 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 three cases, the stimuli in the present study were identical to those of the original VST.

Procedure. The experiment was programmed using Gorilla Experiment Builder (https://gorilla.sc), which was also used to control online data collection. The task is available to preview and clone for re-use in Gorilla Open Materials (https://app.gorilla.sc/openmaterials/245615). Each trial consisted of the presentation of a visual

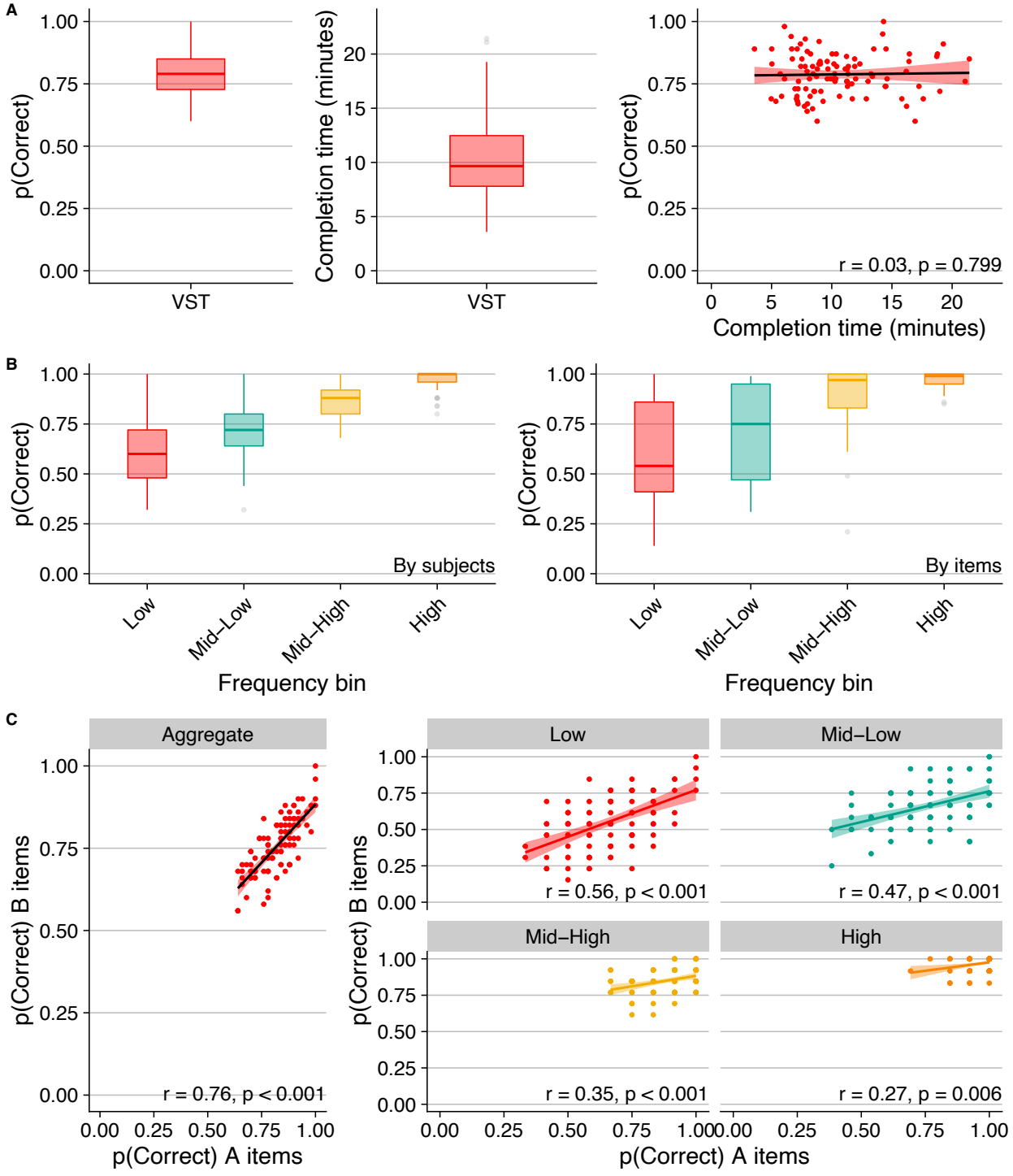
display. The item prompt appeared at the top of the display, the four response options appeared in the middle of the display as clickable buttons beneath the prompt, and a progress bar appeared at the bottom of the display. On each trial, participants were directed to choose which of the four response options best defined the word shown in the item prompt. A response was required on every trial and participants were encouraged to guess if they were unsure. Participants completed 100 trials in total, reflecting one unique randomization of the 100 test items. The ISI was 500 ms, timed from the participant's response. Participants were compensated with $1.67 following the completion of the experiment, an amount that reflects pro-rated compensation equivalent to $10/hour based on an estimated completion time of 10 minutes.

Results

Three sets of analyses were conducted. The first analysis examined mean accuracy and completion time, the second analysis examined the relationship between accuracy and word frequency, and the third analysis examined split-half reliability. Each is addressed in turn, below. We note that the Supplementary Material provides figures illustrating performance for each individual participant in addition to providing normative data (i.e., *mean* and *SD*) for each item on the VST.

Mean accuracy and completion time. In the first analysis, accuracy (mean proportion correct across the 100 test trials) was calculated for each participant, in addition to total completion time. Figure 1, panel A shows the boxplot distribution of accuracy scores across the 100 participants. There is considerable variability in accuracy across participants (*range* = 0.60 – 1.00). However, mean accuracy across participants was relatively high (0.79, *SD* = 0.08), and even the lowest scoring participant showed accuracy above chance level (defined as 0.25, given the four-alternative, forced-choice task). Figure 1, panel A also shows the boxplot distribution of

Figure 1. Results of the VST examined in Experiment 1. Panel A shows the boxplot distribution of accuracy (proportion correct) and completion time across participants, and their relationship. Panel B shows the accuracy boxplot distributions for each frequency bin by subjects (left) and by items (right). Panel C shows split-half reliability for accuracy in the aggregate (left) and by frequency bin (right). Individual points show by-subject means; functions indicate the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit.



completion times across participants. Mean completion time was 11 minutes (*SD* = 4 minutes). The plot at right in Figure 1, panel A shows the relationship between accuracy and completion time across participants; there was no evidence to suggest a speed-accuracy tradeoff ($r = 0.03$, $p = 0.799$).

Accuracy by frequency bin. The second analysis considered performance with respect to the frequency groups of the VST. Recall that the VST was designed to present five items from each of 20 frequency groups, which could be used to estimate a person's vocabulary size. For example, the 1,000 group contains high frequency words that are expected to be in the mental lexicon of a person with a very small vocabulary. In contrast, the 20,000 group contains words with very low frequencies, and thus are expected to be in the mental lexicon of a person who has a very large vocabulary. If the frequency norms used to develop the VST reflect current word usage, then we should observe a relationship between accuracy and frequency group.

To promote more direct comparison to the WordFAM assessment, which is arranged into low, middle, and high frequency bins, the 20 frequency groups of the VST were each assigned to one of four frequency bins that consisted of successive groupings of five consecutive frequency groups. The low frequency bin contained the 16,000 – 20,000 frequency groups, the mid-low frequency bin contained the 11,000 – 15,000 frequency groups, the mid-high frequency bin contained the 6,000 – 10,000 frequency groups, and the high frequency bin contained the 1,000 – 5,000 frequency groups. Thus, each frequency bin contained 25 items (5 frequency groups * 5 items). The boxplot distribution of accuracy scores for each frequency bin is shown in Figure 1, panel B by subjects (at left) and by items (at right). Visual inspection suggests a monotonic increase in accuracy across the four frequency bins for both the by-subject and by-item distributions. As viewed in Figure 1, panel B, the two lower frequency bins show greater

variability in participants' mean accuracy scores compared to the two higher frequency bins. Some of this variability likely reflects individual differences in vocabulary knowledge. However, variability within each frequency bin could also reflect the degree to which individual items are representative of each frequency bin. That is, English word usage changes over time and, consequently, items that were low frequency when the VST was developed may not be low frequency in present times. Indeed, visual inspection of the by-items distributions reveals substantial by-item variability for the low and mid-low frequency bins relative to the by-item variability observed for the mid-high and high frequency bins.

To examine accuracy as a function of frequency bin, trial-level responses (0 = incorrect, 1 = correct) were submitted to a generalized linear mixed effect model (GLMM) 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 ($\hat{\beta} = 1.327$, $SE = 0.592$, $z = 2.240$, $p = 0.025$).

Split-half reliability. The results presented thus far suggest that the web-based version of the VST developed here has promise for use in online research studies. Specifically, mean accuracy across participants is well above chance and accuracy exhibits the expected frequency effect. Mean completion time across participants was 11 minutes (SD = 4 minutes), which clearly

constitutes a "quick" task. However, given that researchers may choose to use the VST as a means to screen and/or verify English word knowledge in participants who also complete an additional experimental task, the web-based VST may be of greater utility to the field if it were shorter in duration. Indeed, current best practice suggests that keeping total testing time minimal supports higher quality data in web-based studies (e.g., Rodd, 2019). To examine whether reliable VST scores could be obtained with fewer trials, we analyzed the split-half reliability of VST scores across participants. The item split was achieved by calculating mean proportion correct for each participant separately for odd- and even-numbered items, which we refer to the A and B items, respectively. In the paper version of the VST, items are numbered consecutively (i.e., 1 – 100) across ascending frequency groups. As such, making a split based on odd vs. even item numbers yields equal frequency representation between the two halves.

Figure 1, panel C shows the association between accuracy on the A and B items in the aggregate (at left) and by frequency bin (at right). In the aggregate, the VST yielded high split-half reliability ($r = 0.76$, $p < 0.001$). Split-half reliability was variable across the frequency bins, with numerically higher split-half reliability for the low ($r = 0.56$, $p < 0.001$) and mid-low ($r = 0.47$, $p < 0.001$) bins compared to the mid-high ($r = 0.35$, $p < 0.001$) and high ($r = 0.27$, $p = 0.006$) frequency bins.

Experiment 2

Methods

Participants. A different sample of participants ($n = 100$) was recruited from the Prolific participant pool following the same inclusion criteria as described for Experiment 1; no participant completed more than one experiment across the three experiments presented in this manuscript. The resulting sample consisted of 48 men, 51 women, and one participant who

declined to report gender. The mean age of participants was 27 years (*SD* = 5 years). Comprehensive demographic characteristics of the sample including race, ethnicity, and dialect are available in the Supplementary Material. Two additional participants were excluded from the final analyses 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).

Stimuli. Stimuli consisted of the 150 items on the Word Familiarity Test (WordFAM). Each item is a single word. The 150 items were designed to sample the English lexicon across a wide range of lexical frequency, with 50 items in each of three frequency categories. The categories included low frequency items (e.g., *inrush*, *shibboleth*), mid frequency items (e.g., *radioisotope*, *undulant*), and high frequency items (e.g., *mother*, *educate*). The stimuli used in the present study were identical to those on the paper-and-pencil version of the WordFAM (Lewellen et al., 1993; Pisoni, 2007).

Procedure. The task was programmed using Gorilla Experiment Builder (https://gorilla.sc), which was also used to control online data collection. The task is available to preview and clone for re-use in Gorilla Open Materials (https://app.gorilla.sc/openmaterials/245615). Each trial consisted of a visual array. The word appeared in the middle of the display, the Likert scale response options appeared as clickable buttons beneath the word, and a progress bar appeared at the bottom of the array. The Likert scale remained at the top of the display across trials. On each trial, participants were directed to rate their familiarity with the word according to the provided scale, which is shown in Table 1. A response was required on every trial and participants were encouraged to guess if they were

unsure. This procedure yielded a web-based version of the WordFAM that was identical to the paper WordFAM version except that (1) the 150 items were randomized separately for each participant, instead of being presented in a single fixed randomization, (2) participants saw one word at a time, instead of lists of words, and (3) participants clicked a button from an array of digits instead of circling a digit on a paper form. Participants completed 150 trials in total, reflecting one unique randomization of the 150 test items. The ISI was 500 ms, timed from the participant's response. Participants were compensated with $2.50 following the completion of the experiment, an amount that reflects pro-rated compensation equivalent to $10/hour based on an estimated completion time of 15 minutes.

Table 1. Likert scale used to elicit familiarity ratings for the WordFAM assessment. As described in the main text, this scale was displayed on the screen during administration of the WordFAM assessments in Experiments 2 and 3.

Rating	Reference
1	You have never seen or heard the word before.
2	You think that you might have seen or heard the word before.
3	You are pretty sure that you have seen or heard the word but you are not positive.
4	You recognize the word as one you have seen or heard before, but you don't know the meaning of the word.
5	You are certain that you have seen the word but you only have a vague idea of its meaning.
6	You think you know the meaning of the word but are not certain that the meaning you know is correct.
7	You recognize the word and are confident that you know the meaning of the word.

Results

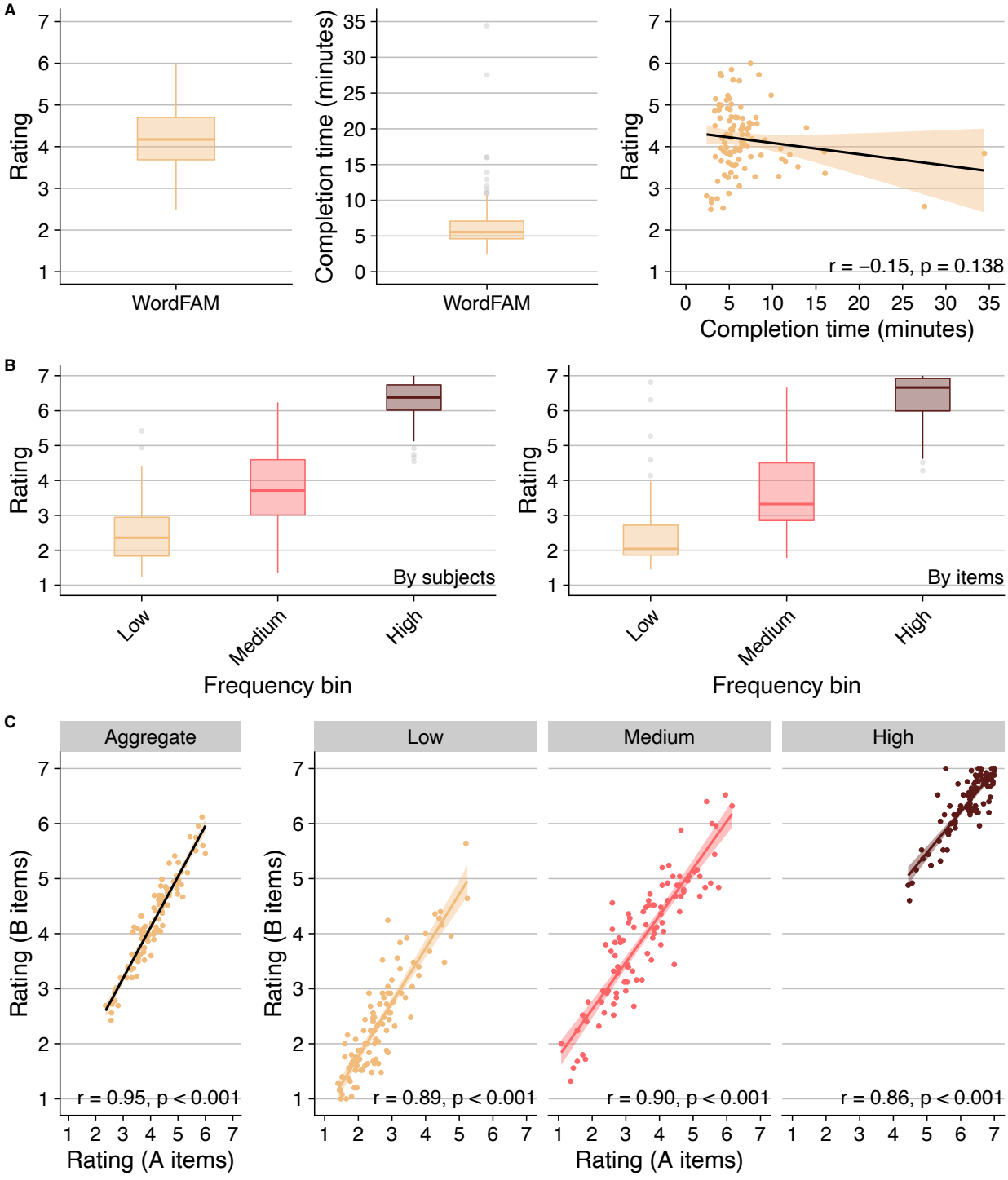
Three sets of analyses were conducted that were parallel to the analyses outlined in the Experiment 1. The first analysis examined mean familiarity ratings and completion times, the second analysis examined the relationship between familiarity ratings and word frequency, and

the third analysis examined split-half reliability. A final analysis was conducted to directly compare performance in the current sample to the existing WordFAM norms. Each is addressed in turn, below. We note that the Supplementary Material provides figures illustrating performance for each individual participant in addition to providing normative data for each item on the WordFAM for the current sample.

Mean rating and completion time. In the first analysis, mean familiarity rating was calculated for each participant, in addition to total completion time. Figure 2, panel A (left) shows the boxplot distribution of mean rating scores across the 100 participants. There is considerable variability in mean rating across participants (*range* = 2.5 – 6.0). The mean rating across participants was at the center of the Likert scale (4.2, *SD* = 0.8). Figure 2, panel A (middle) also shows the boxplot distribution of completion times across participants. Mean completion time was 7 minutes (*SD* = 4 minutes). The plot at right in Figure 2, panel A shows the relationship between mean rating and completion time across participants; there was no evidence to suggest an association between participants' mean ratings and completion times (r = -0.15, p = 0.138). We note that this relationship is further attenuated when the two participants exceeding completion times of 20 minutes are excluded (r = -0.04, p = 0.714).

Ratings by frequency bin. The second analysis considered performance with respect to the frequency bins of the WordFAM. Recall that the WordFAM contains 50 items in each of three frequency bins. If the frequency norms used to develop the WordFAM reflect current word usage, then we should observe a relationship between familiarity ratings and frequency bin. The boxplot distribution of mean ratings for each frequency bin are shown in Figure 2, panel B by subjects (at left) and by items (at right). Visual inspection suggests a monotonic increase across the three frequency bins for both the by-subject and by-item rating distributions.

Figure 2. Results of the WordFAM test examined in Experiment 2. Panel A shows the boxplot distribution of mean ratings and completion time across participants, and their relationship. Panel B shows the rating boxplot distributions for each frequency bin by subjects (left) and by items (right). Panel C shows split-half reliability for mean ratings in the aggregate (left) and by frequency bin (right). Individual points show by-subject means; functions indicate the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit.



To examine familiarity ratings as a function of frequency bin, trial-level ratings were submitted to a linear mixed effect model (LMM) 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, $p < 0.001$).

Split-half reliability. The results presented thus far suggest that the web-based version of the WordFAM developed here has promise for use in online research studies. Specifically, familiarity ratings exhibit the expected frequency effect and mean completion time was relatively quick. However, as in Experiment 1, we examined split-half reliability of the WordFAM assessment in order to assess whether reliable WordFAM scores may be obtained with fewer trials. The item split was achieved as follows. First, the 150 words were sorted by their original normed score. Second, items were alternately assigned to A and B versions moving from the lowest to highest normed scores. In this way, 25 items in each frequency bin were assigned to the A version, 25 items in each frequency bin were assigned to the B version, and normed scores were equivalent between the A and B versions within each frequency category. Figure 2, panel C shows the association between familiarity ratings on the A and B items in the aggregate (at left) and separately by the three frequency bins (at right). Split-half reliability for the WordFAM was extremely high in the aggregate ($r = 0.95$, $p < 0.001$) and for each of the low ($r = 0.89$, $p < 0.001$), middle ($r = 0.90$, $p < 0.001$), and high ($r = 0.86$, $p < 0.001$) frequency bins.

Comparison between Prolific sample and existing norms. Recall that normative data exist

for the WordFAM. These data were collected in the late 1990s in a laboratory setting and consist of a mean familiarity rating for each of the 150 items as derived across participants from the Indiana University community (the Hoosier sample). Because normative data exist for the WordFAM assessment, we compared performance between the current sample and the existing norms. Figure 3, panel A, shows a strong association between the Hoosier and Prolific samples in terms of the mean familiarity rating for each item ($r = 0.82$, $p < 0.001$).

Figure 3. Comparison between results of the WordFAM test in Experiment 2 and existing norms from the Hoosier mental lexicon corpus. Panel A shows the association between mean by-item ratings in the Hoosier sample and the current Prolific sample. Individual points show mean by-item ratings; the black function indicates the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit. Panel B shows mean by-item ratings for each frequency bin in both samples; error bars indicate standard error of the mean.

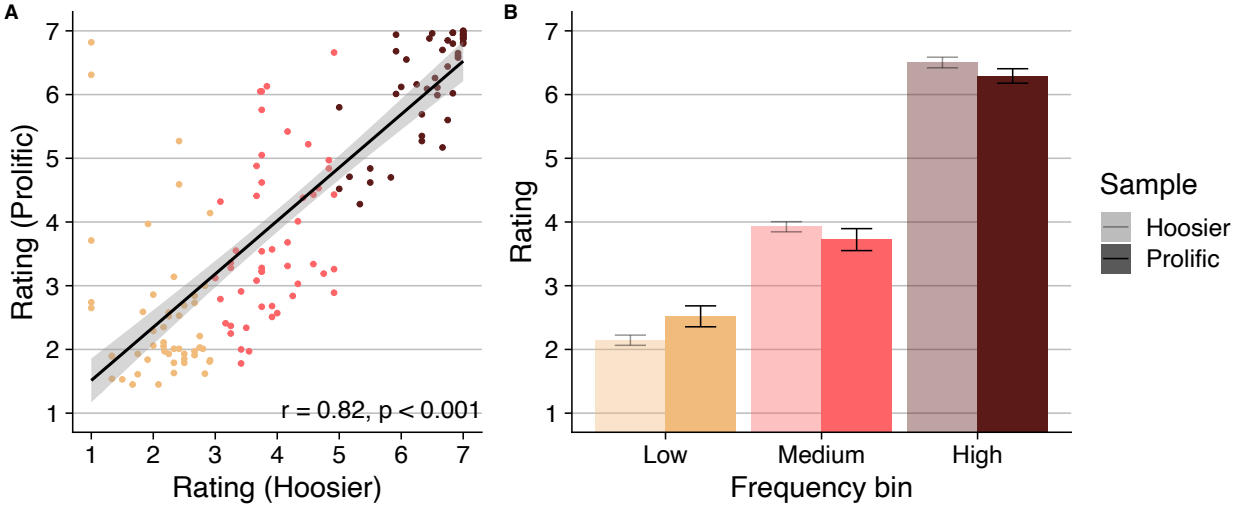


Figure 3, panel B shows the mean familiarity rating for each frequency bin for each sample, which reveals similar ratings between the two samples. Mean item ratings were submitted to ANOVA with the between-subjects factor of frequency bin and the within-subject factor of sample. Because this is a by-items 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, $t(49) = 2.923$, $p = 0.005$].

Experiment 3

The results of Experiments 1 and 2 confirm that the VST and WordFAM assessments are well-suited for web-based administration and provide normative data for their administration with a large, diverse sample. Attrition rate due to failure to comply with task instructions was incredibly low. The results from both tasks exhibited the expected influence of word frequency on performance in both the aggregate and at the level of individual subjects and items; moreover, both tasks exhibited robust individual variation demonstrating that the assessments did not yield only floor or ceiling performance. The results of the WordFAM in Experiment 2 closely matched existing norms for this assessment. Though the results of Experiments 1 and 2 indicate that administration of the full assessments is relatively brief, high split-half reliability for each assessment suggests that stable assessment may be achieved in each task with half the number of trials. However, the normative data gathered in Experiments 1 and 2 did reveal some outlier items for a given lexical frequency bin, consistent with word usage changing over time. Moreover, Experiments 1 and 2 do not allow for assessment of test-retest reliability, nor do they afford assessment of convergent validity across tasks.

In this context, Experiment 3 was conducted to validate two brief versions of the VST

and WordFAM assessments. Each brief version consists of approximately half of the items from the respective full assessment, with a different item subset presented across the brief versions of each assessment. Participants completed a brief version of each assessment at two time points. In addition to providing normative data for the brief versions of each assessment, the goal of Experiment 3 was to measure test-retest reliability across the brief versions of each assessment and to measure convergent validity between the two assessments at each time point.

Methods

Participants. A different sample of participants (n = 85) was recruited from the Prolific participant pool following the same inclusion criteria as described for Experiment 1; no participant completed more than one experiment across the three experiments presented in this manuscript. The resulting sample consisted of 46 men and 39 women. The mean age of participants was 26 years (SD = 5 years). Comprehensive demographic characteristics of the sample including race, ethnicity, and dialect are available in the Supplementary Material. Fifteen additional participants completed session one but declined the invitation to also participate in session two; these participants were excluded from all analyses. Five additional participants were excluded from analyses 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.

Stimuli. Two versions of each assessment, which we refer to as the Brief-A and Brief-B versions, respectively, were created as follows. 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 earlier for Experiment 1 (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 shown in Figure 1, panel B (at right); specifically, we removed items that deviated most substantially from the median item accuracy for a given bin. The Supplementary Material indicates which specific items from the full assessment were removed for the brief versions, in addition to providing 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 for Experiment 2 (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 shown in Figure 2, panel B (at right); specifically, we removed items that deviated most substantially from the median item rating for the low frequency bin. The Supplementary Material indicates which specific items from the full assessment were removed for the brief versions, in addition to providing 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.

Procedure. The VST and WordFAM tasks were programmed using Gorilla Experiment Builder (https://gorilla.sc), which was also used to control online data collection. The tasks are

available to preview and clone for reuse in Gorilla Open Materials (https://app.gorilla.sc/openmaterials/245615). Procedural details of each task were identical to those described in Experiment 1 (VST) and Experiment 2 (WordFAM) with the exception that fewer trials were presented for each task, as described above.

Participants completed two experimental sessions. In session one, participants completed the Brief-A version of each task, with task order counterbalanced across participants (n = 42 for VST followed by WordFAM order, n = 43 for WordFAM followed by VST order). In session two, participants completed the Brief-B version of each task; task order was again counterbalanced across participants (n = 40 for VST followed by WordFAM order, n = 45 for WordFAM followed by VST order). To promote a higher rate of return for session two, participants were given a 60-day window in which to complete session two following session one. The mean time between session was 16 days (SD = 10 days, *range* = 1 – 53 days). As described in the results section, the time between sessions did not predict the difference in performance across sessions. Participants were compensated with $1.67 following the completion of each session, an amount that reflects pro-rated compensation equivalent to $10/hour based on an estimated completion time of 10 minutes.

Results

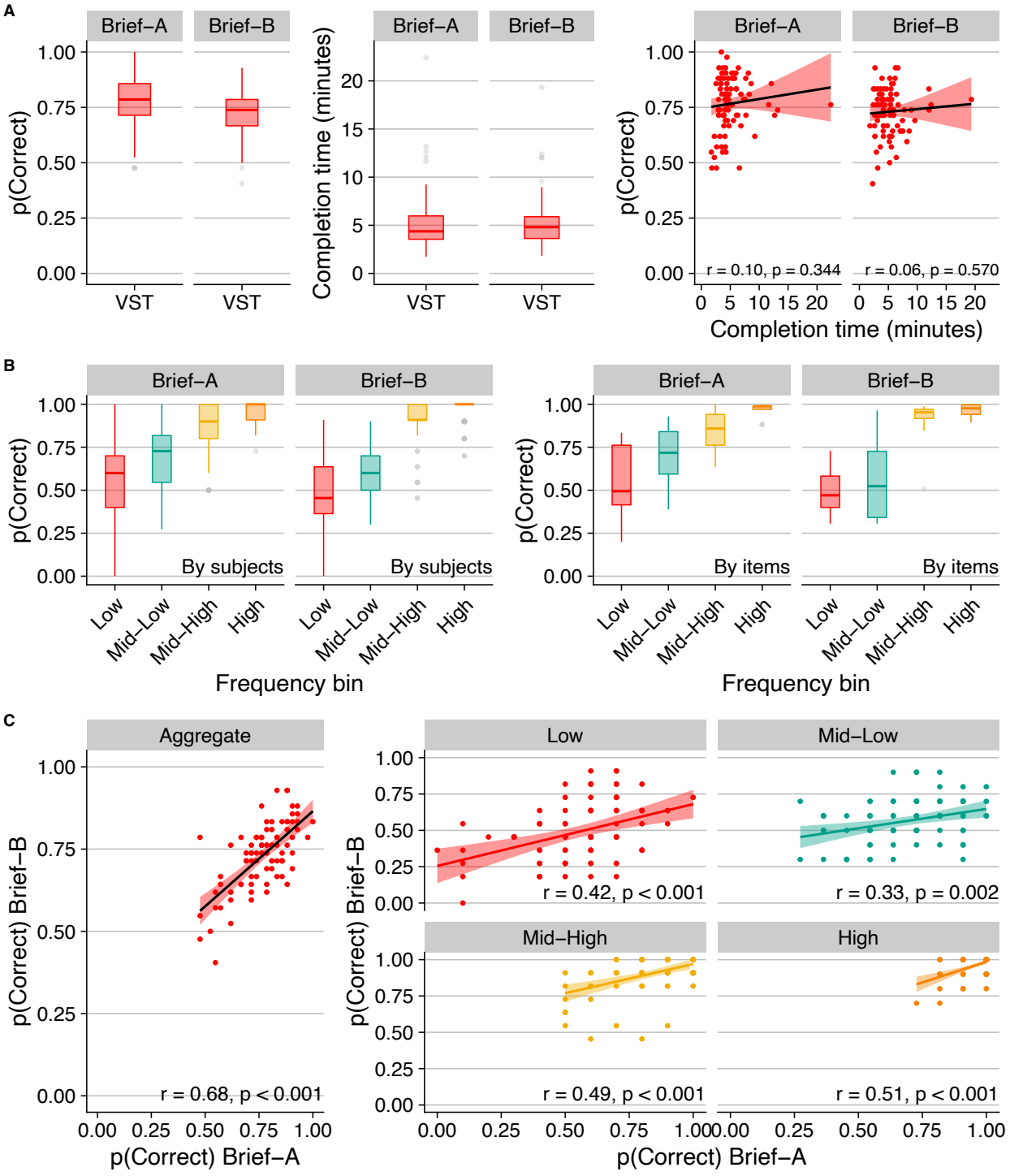
Three sets of analyses were conducted. The first two analyses parallel those presented in Experiment 1 and Experiment 2 for the VST and WordFAM assessments, respectively, to provide complementary analyses for the brief versions of each assessment that were tested in Experiment 3. The third analysis examined performance between the VST and WordFAM assessments. Each is addressed in turn, below. We note that the Supplementary Material presents figures illustrating performance for each participant in all assessments in addition to providing

normative data (i.e., *mean* and *SD* for the sample) for each item on the Brief-A and Brief-B versions of the VST.

Vocabulary Size Test (VST). Completion time and mean proportion correct were calculated for each participant separately for the Brief-A and Brief-B versions of the VST. As shown in Figure 4, panel A, mean proportion correct across participants was high for both the Brief-A (0.77, $SD = 0.12$) and Brief-B (0.73, $SD = 0.10$) versions; likewise, mean completion time was fast (*mean* = 5 minutes, $SD = 3$ minutes for both versions). There was no evidence of a speed-accuracy trade-off for either the Brief-A ($r = 0.10$, $p = 0.344$) or Brief-B ($r = 0.06$, $p = 0.570$) version.

For each VST version, the boxplot distribution of accuracy scores for each frequency bin are shown in Figure 4, panel B by subjects (at left) and by items (at right). Visual inspection suggests a monotonic increase in accuracy across the four frequency bins in each version for both the by-subject and by-item distributions. To examine this pattern statistically, trial-level responses (0 = incorrect, 1 = correct) were submitted to a generalized linear mixed effect 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

Figure 4. Results of the Brief-A and Brief-B VST versions examined in Experiment 3. Panel A shows the boxplot distribution of accuracy (proportion correct) and completion time across participants, and their relationship. Panel B shows the accuracy boxplot distributions for each frequency bin by subjects (left) and by items (right). Panel C shows test-retest reliability for accuracy in the aggregate (left) and by frequency bin (right). Individual points show by-subject means; functions indicate the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit.

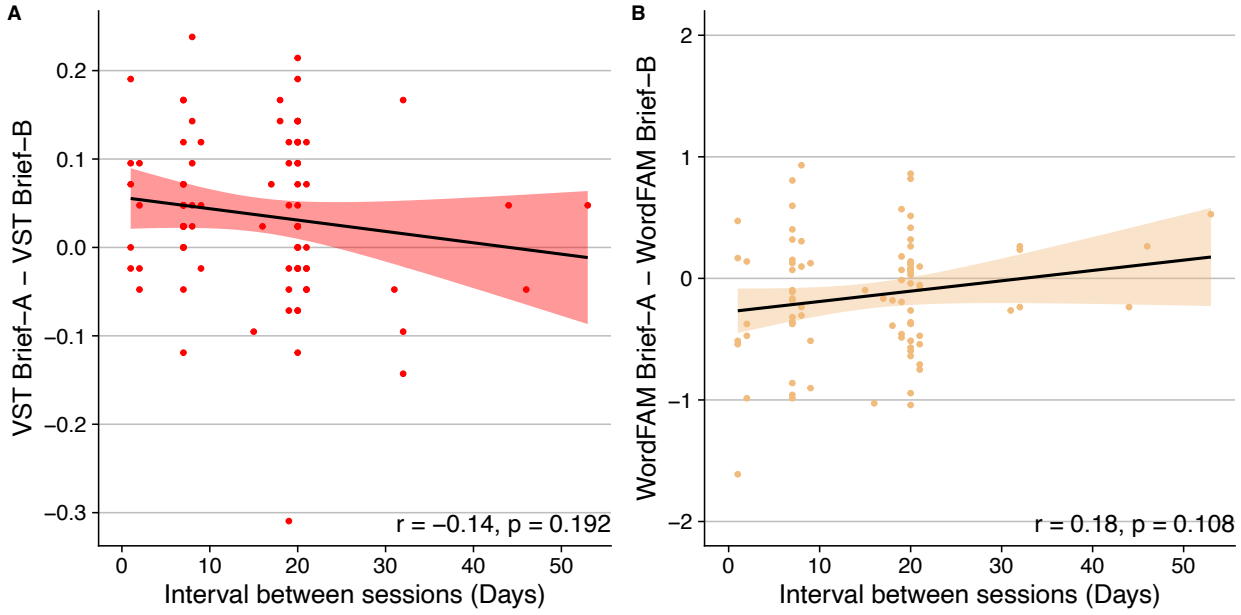


frequency bins ($\hat{\beta}$ = 2.224, *SE* = 0.388, *z* = 5.735, *p* < 0.001) and the mid-high and high frequency bins ($\hat{\beta}$ = 1.752, *SE* = 0.449, *z* = 3.898, *p* < 0.001). The main effect of version was not significant ($\hat{\beta}$ = -0.114, *SE* = 0.277, *z* = -0.412, *p* = 0.680), nor did version interact with frequency bin (p ≥ 0.067 for all three interaction contrasts).

To examine test-retest reliability of the brief VST assessments, we examined the association between individuals' performance on each of the brief versions of the VST. Figure 4, panel C shows the association between accuracy on the brief assessments in the aggregate (at left) and separately by the four frequency bins (at right). In the aggregate, the brief assessments yielded high test-retest reliability (*r* = 0.68, *p* < 0.001). Test-retest reliability was comparable across the frequency bins, all of which showed numerically lower associations compared to the aggregate association (low: *r* = 0.42, *p* < 0.001; mid-low: *r* = 0.33, *p* = 0.002; mid-high: *r* = 0.49, *p* < 0.001; high: *r* = 0.51, *p* < 0.001). Recall that the interval between completion of the two brief versions was highly variable across participants (*mean* = 16 days, *SD* = 10 days, *range* = 1 – 53 days). To examine whether the time between sessions influenced participants' performance, we examined the association between interval and the difference in mean accuracy between the two versions. As shown in Figure 5 (panel A), there was no significant association between the time between sessions and the difference in mean accuracy of the two test versions of the VST (*r* = -0.14, *p* = 0.192).

Word Familiarity Test (WordFAM). Mean familiarity rating was calculated for each participant, in addition to total completion time, for each test version. As shown in Figure 6, panel A, the mean rating across participants was at the center of the Likert scale for both the Brief-A (4.2, *SD* = 0.8) and Brief-B (4.3, *SD* = 0.8) versions of the assessment. In addition, mean completion time was very fast (for both versions: *mean* = 3 minutes, *SD* = 1 minute). The plot at

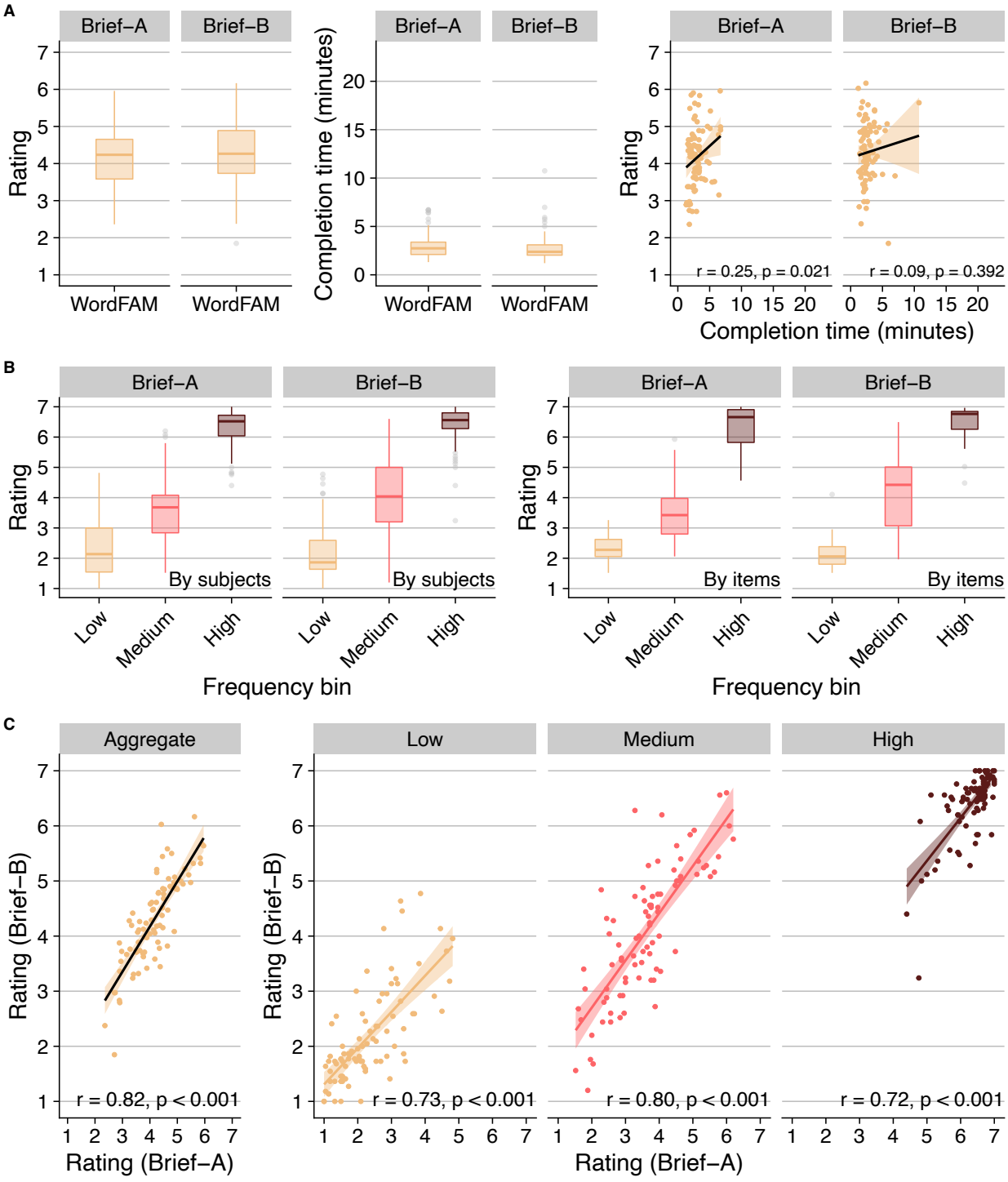
Figure 5. Relationship between time between sessions and difference in performance between sessions for the two brief versions of the VST (panel A) and the two brief versions of the WordFAM (panel B). Individual points show mean by-subject familiarity ratings; the black function indicates the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit.



the right in Figure 6, panel A shows the relationship between mean familiarity rating and completion time across participants. There was a small but statistically reliable positive association between mean rating and completion time for the Brief-A version ($r = 0.25$, $p = 0.021$); no reliable association was observed for the Brief-B version ($r = 0.09$, $p = 0.392$).

The boxplot distribution of mean ratings across subjects for each frequency bin are shown in Figure 6, panel B by subjects (at left) and by items (at right). In both versions, visual inspection suggests a monotonic increase across the three frequency bins for both the by-subject and by-item rating distributions. To test this observation statistically, 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

Figure 6. Results of the Brief-A and Brief-B versions of the WordFAM test examined in Experiment 3. Panel A shows the boxplot distribution of mean ratings and completion time across participants, and their relationship. Panel B shows the rating boxplot distributions for each frequency bin by subjects (left) and by items (right). Panel C shows split-half reliability for mean ratings in the aggregate (left) and by frequency bin (right). Individual points show by-subject means; functions indicate the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit.



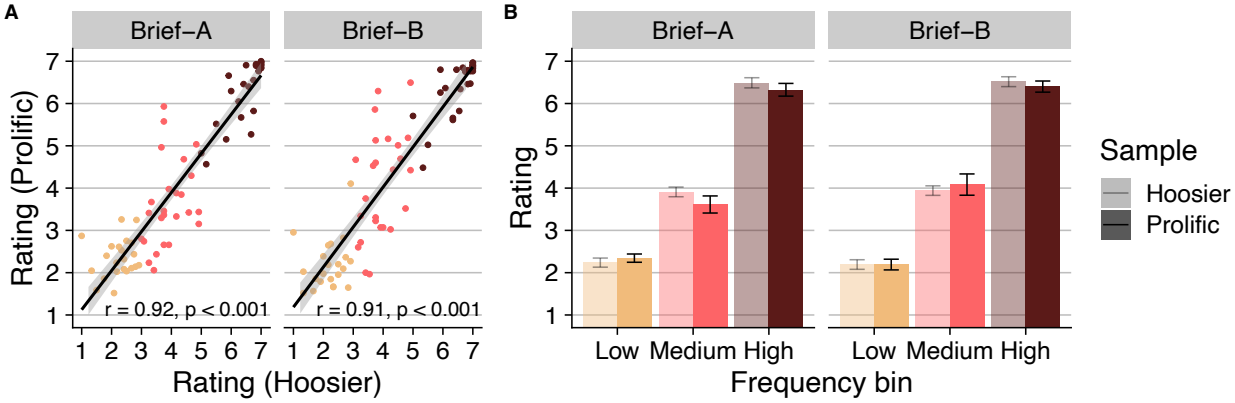
(Brief-A = -0.5, Brief-B = 0.5). The random effects structure consisted of random intercepts by subject, random slopes for frequency bin and version by subject, and random intercepts by item. The results of the model showed a monotonic increase in familiarity ratings from the low to middle frequency 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 ($p \geq 0.077$ for both interaction contrasts).

To examine test-retest reliability of the brief WordFAM assessments, we examined the association between individuals' performance on each of the brief versions of the WordFAM. Figure 6, panel C shows the association between familiarity ratings on the brief assessments in the aggregate (at left) and by each of the three frequency bins (at right). Test-retest reliability for the brief assessments was extremely high in the aggregate ($r = 0.82$, $p < 0.001$) and for each of the low ($r = 0.73$, $p < 0.001$), middle ($r = 0.80$, $p < 0.001$), and high ($r = 0.72$, $p < 0.001$) frequency bins. Because the interval between completion of the two brief versions was highly variable across participants (*range* = 1 – 53 days), we also examined whether the time between sessions influenced participants' performance. As shown in Figure 5 (panel B), there was no significant association between the time between sessions and the difference in mean familiarity ratings of the two test versions ($r = 0.18$, $p = 0.108$).

Finally, we compared performance between the current sample and the existing norms from the Hoosier sample, as described for Experiment 2. Figure 7, panel A (left), shows a strong association between the Hoosier and Prolific samples in terms of the mean item rating for both the Brief-A ($r = 0.92$, $p < 0.001$) and Brief-B ($r = 0.91$, $p < 0.001$) versions of the WordFAM. Figure 7, panel B shows the mean item rating for each frequency bin for each sample, which

reveals similar ratings between the two samples for each test version. 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 (a given item can only be in one frequency bin or one test version) 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 [$F(2, 138) = 571.52$, $p < 0.001$]. 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).

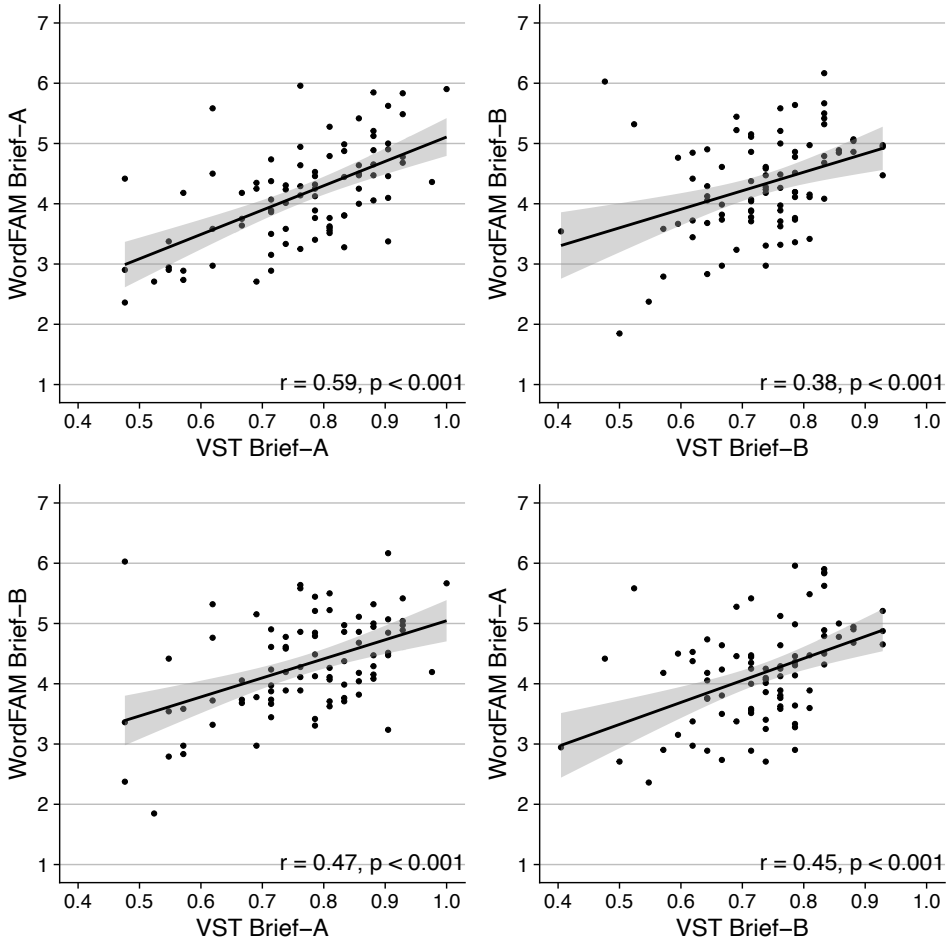
Figure 7. Comparison between results of the Brief-A and Brief-B versions of the WordFAM test in Experiment 3 and existing norms from the Hoosier mental lexicon corpus. Panel A shows the association between mean by-item ratings in the Hoosier sample and the current Prolific sample. Individual points show mean by-item ratings; the black function indicates the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit. Panel B shows mean by-item ratings for each frequency bin in both samples; error bars indicate standard error of the mean.



Convergent validity of the VST and WordFAM measures. To assess convergent validity across the VST and WordFAM brief assessments, four correlations were calculated. Values for each correlation consisted of by-subjects mean accuracy on the respective VST assessment and by-subjects mean rating on the respective WordFAM assessment. There was a significant association between the Brief-A versions of the VST and WordFAM ($r = 0.59$, $p < 0.001$) and

the Brief-B versions of the VST and WordFAM ($r = 0.38$, $p < 0.001$); the association was numerically weaker in the latter compared to the former. Recall that the A versions of each assessment were completed at session one and the B versions of each assessment were completed at session two. Moderate associations were also observed between the two assessments across sessions. Specifically, there was a moderate association between the VST Brief-A (completed at session one) and the WordFAM Brief-B (completed at session two, approximately two weeks later); $r = 0.47$, $p < 0.001$) and between the VST Brief-B (completed at session two) and the WordFAM Brief-A (completed at session 1; $r = 0.45$, $p < 0.001$).

Figure 8. Relationship between performance on the VST and WordFAM assessments. Individual points show mean accuracy (VST) and mean rating (WordFAM) for individual subjects. In all plots, the black function indicates the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit.



Discussion

The purpose of the current study was to develop and validate two web-based measures for assessing English vocabulary knowledge. In the introduction, we outlined four criteria for each measure. First, the assessment should be openly available for free and public re-use in the research domain. Second, the assessment should be fast and easy to complete without requiring real-time interaction between a participant and a researcher. Third, the assessment should yield acceptable psychometric properties including split-half reliability and test-retest reliability. Fourth, the two assessments should show strong convergent validity as an indicant of construct validity for each assessment. In this discussion, we consider the results against the four criteria, and then conclude with broader implications of the current work.

In Experiment 1, participants completed a web-based version of the existing long-form VST. We observed moderate variability in completion time and accuracy across participants, with relative fast mean completion times and relatively high accuracy. Accuracy was linked to lexical frequency, as expected. Split-half reliability of this task was strong ($r = 0.78$). In Experiment 2, participants completed a web-based version of the existing long-form WordFAM assessment. We observed minimal variability in completion time, which was relatively fast, and mean ratings were centered on the 7-point familiarity rating scale. Ratings systematically reflected lexical frequency, and split-half reliability for the WordFAM was extremely high ($r = 0.95$). There was a strong association between the normative data gathered in the Prolific sample and existing normative data from the Hoosier mental lexicon corpus. Capitalizing on the high split-half reliability observed for each assessment, Experiment 3 tested two brief versions of each assessment separated in time. The results of each brief version patterned in line with the full versions tested in Experiments 1 and 2. Test-retest reliability of each assessment was strong

(VST, $r = 0.69$; WordFAM, $r = 0.82$) and the two assessments showed moderate-to-strong convergent validity (ranging from $r = 0.38$ to $r = 0.59$).

Collectively, these results indicate that the web-based vocabulary knowledge assessments developed here are suitable for use in remote research. We note that though the full versions of each assessment are on their own relatively brief, the brief versions are even more efficient. Assessment administration time is a critical consideration for both research and clinical practice. An increase in administration time is associated with increased research costs due to higher participant compensation. From a clinical perspective, an increase in testing time is also associated with an increase in cost for the individual receiving services. Measures with a shorter administration time therefore contribute to removing financial barriers associated with research and clinical practice. We also aimed to create an assessment that does not require any monetary investment or additional training to administer. All versions of the VST and WordFAM tests described here are freely available on Gorilla Experiment Builder as Open Materials (https://app.gorilla.sc/openmaterials/245615); moreover, the item lists are also provided on the OSF repository for this manuscript and thus available for use in other web-based platforms. The assessments on Gorilla are automated in their administration and do not require physical, synchronous interaction between the researcher and the participant.

The current results also provide normative data for the web-based versions of the VST and WordFAM measures developed in the current work. In addition to proving trial-level data and analysis code for all experiments presented here, we have also provided a comprehensive Supplemental Material document that provides detailed normative data for each item on each assessment, a complete reporting of participant demographics (e.g., age, gender, race, ethnicity, dialect), and figures showing individual subject functions for the three samples presented here.

We note that the normative data provided by the current work may be particularly useful for researchers because they extend beyond the typical age range included in existing standardized assessments. The current experiments were not designed to provide age-related norms; instead, we considered performance across the 18 – 35 years of age range in the aggregate. However, the open data set and open materials set a foundation for future research that could, for example, extend the existing data set to formally establish age-related norms across a wide age range.

Like any assessment, the measures presented in this paper are not without their limitations. For example, the use of multiple-choice questions to measure vocabulary competency on the VST may raise concern. While there is some evidence that multiple-choice assessments can be reliable measures due to their correlation to performance on assessments using other answer strategies and their high test-retest reliability (McCoubrie, 2004; Roediger & Marsh, 2005), substantive criticisms of multiple-choice questions have been raised due to their prioritization of "factual regurgitation" over more nuanced, higher-order thinking (Stewart, 2014). Prior work examining the VST also argues that the multiple-choice format used for this assessment results in a test of vocabulary recognition, rather than vocabulary knowledge, and may inflate the estimate of vocabulary knowledge (Stewart, 2014). It is also possible that performance on a vocabulary assessment is higher when the examinee is asked to recognize a vocabulary item from a closed set of options, as is required on the VST, rather than to recall a vocabulary word from memory (Laufer & Goldstein, 2004). However, gold-standard measures for vocabulary assessment, such as the PPVT-3, continue to rely on multiple-choice responses to assess vocabulary knowledge.

Likewise, results from the WordFAM test should not be interpreted without consideration of potential limitations. As stated previously, the WordFAM is a subjective measure of

vocabulary knowledge, and therefore cannot be interpreted as a definitive measure of vocabulary competence. However, extant research demonstrates that word familiarity ratings are strongly associated with behavioral measures of lexical access, providing some assurance that the subjective rating scale used on the WordFAM does not hinder its ability to measure vocabulary competency (Gernsbacher, 1984; Lewellen et al., 1993; Tamati et al., 2013; Tamati & Pisoni, 2014)

Finally, it is important to highlight that the normative data gathered for the current vocabulary assessments were obtained from (self-reported) monolingual speakers of American English from a single participant pool (Prolific; Palan & Schitter, 2018); accordingly, the utility of the normative data may be limited to this population. That is, though we specifically recruited monolingual English speakers via the Prolific participant pool and have no evidence to indicate that participants were dishonest in their self-reported language background, we do not have "ground truth" of language background that might be more obtainable in a traditional laboratory-based environment. However, the striking similarity in normative data between with the current Prolific sample and the existing Hoosier norms for the WordFAM assessments provides some assurance of integrity in participants' self-reported language background.

Despite these limitations, the current results suggest that the web-based vocabulary knowledge assessments developed here can be used to reliably, and validly, assess English vocabulary knowledge in individuals ages 18 – 35 years on web-based testing platforms. Accordingly, these assessments, which are freely available for reuse, provide a new tool for screening based on vocabulary knowledge, confirming self-reported language proficiency, or for investigating the relationship between vocabulary and other constructs of interest. Each measure is suitable to stand alone, though joint administration is also possible given the brief completion

times. Future research should examine the relationship between performance on the current open-source measures and existing for-fee conventional, in-person standardized assessments (e.g., CELF-5, EVT-2, PPVT-3). Future research should also aim to adapt these measures to increase the dialectal and multicultural sensitivity of the stimuli to capture vocabulary competency across the diversity of English-speaking individuals.

Acknowledgments

This work was supported by NIH NIDCD grant R21DC016141 to RMT, NSF grants DGE-1747486 and DGE-1144399 to the University of Connecticut, NIH NIDCD grant R01DC015257 to Indiana University, and by the Jorgensen Fellowship (University of Connecticut) to NG. LD was supported by NIH NIDCD grant T32DC017703. The views expressed here reflect those of the authors and not the NIH, the NIDCD, or the NSF. Portions of these data were presented at the 2021 convention of the American Speech-Language-Hearing Association.

Open Practices Statement

Trial-level data, analysis code, and materials for all experiments are available at https://osf.io/pcsu6/. Ready-to-run tasks are provided as Open Materials for Gorilla Experiment Builder at https://app.gorilla.sc/openmaterials/245615. The experiments were not preregistered.

References

American Speech-Language-Hearing-Association. (n.d.). *Telepractice*. American Speech-Language-Hearing Association; American Speech-Language-Hearing Association. Retrieved June 11, 2022, from https://www.asha.org/practice-portal/professional-issues/telepractice/

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Prentice Hall/Pearson Education.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388-4071–20. https://doi.org/10.3758/s13428-019-01237-x

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101–118.

Beglar, D., & Nation, P. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13.

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*, *37*(6), 1461–1476. https://doi.org/10.1017/S0142716416000060

Bloom, P. (2002). *How children learn the meanings of words*. MIT press.

Colby, S., Clayards, M., & Baum, S. (2018). The role of lexical status and individual differences for perceptual learning in younger and older adults. *Journal of Speech, Language, and Hearing Research*, *61*(8), 1855–1874.

Coxhead, A. (2016). Dealing with low response rates in quantitative studies. In *Doing Research in Applied Linguistics* (pp. 81–90). Routledge.

Coxhead, A., Nation, P., & Sim, D. (2014). Creating and trialling six versions of the vocabulary size test. *The Tesolanz Journal*, *22*, 13–27.

Coxhead, A., Nation, P., & Sim, D. (2015). Measuring the vocabulary size of native speakers of English in New Zealand secondary schools. *New Zealand Journal of Educational Studies*, *50*(1), 121–135.

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test–retest reliability in infant speech perception tasks. *Infancy*, *21*(5), 648–667.

Dunn, L. M., & Dunn, L. M. (1997). *PPVT-III: Peabody Picture Vocabulary Test*. American Guidance Service.

Gathercole, S. E., & Baddeley, A. D. (1993). Phonological working memory: A critical building block for reading development and vocabulary acquisition? *European Journal of Psychology of Education*, *8*(3), 259–272.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256–281.

Giovannone, N., & Theodore, R. M. (2021). Individual differences in lexical contributions to speech perception. *Journal of Speech, Language, and Hearing Research*, *64*(3), 707–724.

Godinho, A., Schell, C., & Cunningham, J. A. (2020). Out damn bot, out: Recruiting real people into substance use studies on the internet. *Substance Abuse*, *41*(1), 3–5.

Griffin, M., Martino, R. J., LoSchiavo, C., Comer-Carruthers, C., Krause, K. D., Stults, C. B., & Halkitis, P. N. (2021). Ensuring survey research data integrity in the era of internet bots. *Quality & Quantity*, 1–12.

Heffner, C. C., Fuhrmeister, P., Luthra, S., Mechtenberg, H., Saltzman, D., & Myers, E. B. (2022). Reliability and validity for perceptual flexibility in speech. *Brain and Language*, *226*, 105070.

Irwin, J. R., Carter, A. S., & Briggs-Gowan, M. J. (2002). The social-emotional development of "late-talking" toddlers. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(11), 1324–1332.

Kavé, G., Knafo, A., & Gilboa, A. (2010). The rise and fall of word retrieval across the lifespan. *Psychology and Aging*, *25*(3), 719.

Landi, N. (2010). An examination of the relationship between reading comprehension, higher-level and lower-level reading sub-skills in adults. *Reading and Writing*, *23*(6), 701–717.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436.

Lewellen, M. J., Goldinger, S. D., Pisoni, D. B., & Greene, B. G. (1993). Lexical familiarity and processing efficiency: Individual differences in naming, lexical decision, and semantic categorization. *Journal of Experimental Psychology: General*, *122*(3), 316.

Mancilla-Martinez, J., Christodoulou, J. A., & Shabaker, M. M. (2014). Preschoolers' English vocabulary development: The influence of language proficiency and at-risk factors. *Learning and Individual Differences*, *35*, 79–86. https://doi.org/10.1016/j.lindif.2014.06.008

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, *26*(8), 709–712.

Nation, P. (2012). *The Vocabulary Size Test*. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-information-and-specifications.pdf

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon. *Research on Spoken Language Processing Report*, *10*(3), 357–376.

Palan, S., & Schitter, C. (2018). Prolific. Ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Pisoni, D. B. (2007). *WordFam: Rating word familiarity in English*. Indiana University.

Rodd, J. (2019). How to maintain data quality when you can't see your participants. *APS Observer*, *32*(3).

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155–1159.

Rotman, T., Lavie, L., & Banai, K. (2020). Rapid perceptual learning: A potential source of individual differences in speech perception under adverse conditions? *Trends in Hearing*, *24*, 2331216520930541.

Snow, C. E., & Kim, Y.-S. (2007). Large problem spaces: The challenge of vocabulary for English language learners. In *Vocabulary acquisition: Implications for reading comprehension* (pp. 123–139). Guilford Press.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*(3), 271–282.

Storozuk, A., Ashley, M., Delage, V., & Maloney, E. A. (2020). Got bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology*, *16*(5), 472–481.

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research*, *61*(6), 1463–1486.

Tamati, T. N., Gilbert, J. L., & Pisoni, D. B. (2013). Some factors underlying individual differences in speech recognition on PRESTO: A first report. *Journal of the American Academy of Audiology*, *24*(07), 616–634.

Tamati, T. N., & Pisoni, D. B. (2014). Non-native listeners' recognition of high-variability speech using PRESTO. *Journal of the American Academy of Audiology*, *25*(09), 869–892.

Theodore, R. M., Monto, N. R., & Graham, S. (2019). Individual differences in distributional learning for speech: What's ideal for ideal observers? *Journal of Speech, Language, and Hearing Research*, 1–13.

Wasik, B. A., Hindman, A. H., & Snell, E. K. (2016). Book reading and vocabulary development: A systematic review. *Early Childhood Research Quarterly*, *37*, 39–57.

Wiig, E. H., Semel, E., & Secord, W. (2013). Clinical Evaluation of Language Fundamentals—Fifth Edition. *Bloomington, MN: Pearson*.

Wilbiks, J. M., Brown, V. A., & Strand, J. F. (2022). Speech and non-speech measures of audiovisual integration are not correlated. *Attention, Perception, & Psychophysics*, 1–11.

Williams, K. T. (1997). Expressive vocabulary test second edition (EVT^{TM} 2). *Journal of the American Academy of Child Adolescent Psychiatry*, *42*, 864–872.

SUPPLEMENTARY MATERIAL

This document provides supplementary material for "Validation of two measures for assessing English vocabulary knowledge on web-based testing platforms," including detailed demographic reporting, individual subject performance, and normative values for the web-based versions of the Vocabulary Size Test (VST) and Word Familiarity Test (WordFAM). The list of contents is shown below.

Tables

Figures

Table S1. Self-reported gender and age characteristics of the participant sample in each experiment.

		Gender			Age (years)		
Experiment	*n*	Men	Women	Unreported	*Mean*	*SD*	*Range*
E1	100	47	53	0	25	6	18 – 35
E2	100	48	51	1	27	5	18 – 35
E3	85	46	39	0	26	5	18 – 35

[Continued on next page.]

Table S2. Self-reported dialect characteristics of the participant sample in each experiment.

Dialect	E1	E2	E3
American	–	1	–
Californian	–	–	1
Decline to state	2	1	1
Delmarva (between New England and Southern)	–	–	1
Do not know	11	17	12
Do not know, Mid Atlantic	1	–	–
Do not know, Mid Atlantic (Pennsylvania)	1	–	–
Do not know, New England	1	–	–
Do not know, Pacific Northwest	–	–	1
Do not know, Southern	–	–	2
Do not know, Southern, Southwestern	–	1	–
Eastern/Pittsburgh	1	–	–
Floridian	–	–	1
Inland Northern	1	–	–
Mid Atlantic	3	2	–
Midwestern	27	22	22
Midwestern, Pacific Southwest	1	–	–
Midwestern, Southwestern	1	–	1
New England	10	13	11
New England, Midwestern	1	–	–
New England, New York	1	–	–
New York	1	1	–
Northeastern	1	2	1
Northeastern (Philadelphia)	1	–	–
Pacific Northwest	5	9	7
Pacific Southwest	7	10	9
Southern	12	16	11
Southern Californian	1	–	–
Southern, Midwestern	2	–	–
Southern, Southwestern	–	–	1
Southern, Texan	1	–	–
Southwestern	5	4	3
Southwestern, Texan, Non-regional Diction	1	–	–
Spanglish	1	–	–
Tristate	–	1	–

Table S3. Self-reported race and ethnicity characteristics of the participant sample in each experiment.

Experiment	Race	Ethnicity		
		Hispanic or Latino	Not Hispanic or Latino	Unreported
E1	American Indian/Alaska Native	–	1	–
	American Indian/Alaska Native, White, More than One Race	–	1	–
	Asian	–	6	–
	Asian, Native Hawaiian or Other Pacific Islander, White	–	1	–
	Asian, White	–	2	–
	Black or African American	–	7	–
	Black or African American, White	1	–	–
	Black or African American, White, More than One Race	–	1	–
	Decline to state	–	–	2
	More than One Race	1	1	–
	White	3	73	–
E2	Asian	–	11	–
	Black or African American	–	8	–
	Black or African American, White	–	2	–
	More than One Race	1	–	–
	White	1	72	1
	American Indian/Alaska Native, White	–	1	–
	Asian, White, More than One Race	–	3	–
E3	American Indian/Alaska Native	–	1	–
	Asian	–	6	–
	Asian, White	–	1	–
	Black or African American	1	11	–
	Black or African American, White	1	–	–
	Black or African American, White, More than One Race	–	2	–
	Decline to state	–	–	1
	White	7	50	1
	American Indian/Alaska Native, White	1	1	–
	White, More than One Race	–	1	–

[Continued on next page.]

Table S4. Mean proportion correct (and *SD*) across the 100 participants in experiment 1 for each item on the VST. Number and group correspond to those used in the original VST Form A assessment, available at: https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/VST-version-A.pdf. Bin and version correspond to the divisions made in the current work as described in the main text.

Number	Group	Bin	Version	Item	*Mean*	*SD*
1	G.01000	High	A	see	1.00	0.00
2	G.01000	High	B	time	1.00	0.00
3	G.01000	High	A	period	0.99	0.10
4	G.01000	High	B	figure	0.85	0.36
5	G.01000	High	A	poor	1.00	0.00
6	G.02000	High	B	microphone	0.92	0.27
7	G.02000	High	A	nil	0.94	0.24
8	G.02000	High	B	pub	1.00	0.00
9	G.02000	High	A	circle	0.95	0.22
10	G.02000	High	B	dig	1.00	0.00
11	G.03000	High	A	soldier	1.00	0.00
12	G.03000	High	B	restore	0.99	0.10
13	G.03000	High	A	pro	0.97	0.17
14	G.03000	High	B	compound	0.94	0.24
15	G.03000	High	A	deficit	0.89	0.31
16	G.04000	High	B	strap	0.97	0.17
17	G.04000	High	A	weep	0.99	0.10
18	G.04000	High	B	haunt	0.99	0.10
19	G.04000	High	A	cube	1.00	0.00
20	G.04000	High	B	butler	0.99	0.10
21	G.05000	High	A	nun	1.00	0.00
22	G.05000	High	B	olive	1.00	0.00
23	G.05000	High	A	shudder	0.98	0.14
24	G.05000	High	B	threshold	0.97	0.17
25	G.05000	High	A	demography	0.86	0.35
26	G.06000	Mid-High	B	malign	0.85	0.36
27	G.06000	Mid-High	A	strangle	0.99	0.10
28	G.06000	Mid-High	B	dinosaur	0.98	0.14
29	G.06000	Mid-High	A	jug	1.00	0.00
30	G.06000	Mid-High	B	crab	0.98	0.14
31	G.07000	Mid-High	A	quilt	0.99	0.10
32	G.07000	Mid-High	B	tummy	1.00	0.00
33	G.07000	Mid-High	A	eclipse	1.00	0.00

34	G.07000	Mid-High	B	excrete	0.96	0.20
35	G.07000	Mid-High	A	ubiquitous	0.61	0.49
36	G.08000	Mid-High	B	marrow	1.00	0.00
37	G.08000	Mid-High	A	cabaret	0.92	0.27
38	G.08000	Mid-High	B	cavalier	0.49	0.50
39	G.08000	Mid-High	A	veer	0.74	0.44
40	G.08000	Mid-High	B	yogurt	1.00	0.00
41	G.09000	Mid-High	A	octopus	1.00	0.00
42	G.09000	Mid-High	B	monologue	0.96	0.20
43	G.09000	Mid-High	A	candid	0.83	0.38
44	G.09000	Mid-High	B	nozzle	0.98	0.14
45	G.09000	Mid-High	A	psychosis	0.97	0.17
46	G.10000	Mid-High	B	ruck	0.21	0.41
47	G.10000	Mid-High	A	rouble	0.63	0.49
48	G.10000	Mid-High	B	canonical	0.65	0.48
49	G.10000	Mid-High	A	puree	0.97	0.17
50	G.10000	Mid-High	B	vial	1.00	0.00
51	G.11000	Mid-Low	A	counterclaim	0.95	0.22
52	G.11000	Mid-Low	B	refectory	0.31	0.46
53	G.11000	Mid-Low	A	trill	0.70	0.46
54	G.11000	Mid-Low	B	talon	0.99	0.10
55	G.11000	Mid-Low	A	plankton	0.98	0.14
56	G.12000	Mid-Low	B	soliloquy	0.75	0.44
57	G.12000	Mid-Low	A	puma	0.98	0.14
58	G.12000	Mid-Low	B	augur	0.32	0.47
59	G.12000	Mid-Low	A	emir	0.34	0.48
60	G.12000	Mid-Low	B	didactic	0.43	0.50
61	G.13000	Mid-Low	A	cranny	0.75	0.44
62	G.13000	Mid-Low	B	lectern	0.49	0.50
63	G.13000	Mid-Low	A	azalea	0.91	0.29
64	G.13000	Mid-Low	B	marsupial	0.87	0.34
65	G.13000	Mid-Low	A	bawdy	0.71	0.46
66	G.14000	Mid-Low	B	crowbar	0.97	0.17
67	G.14000	Mid-Low	A	spangled	0.76	0.43
68	G.14000	Mid-Low	B	aver	0.47	0.50
69	G.14000	Mid-Low	A	retro	0.88	0.33
70	G.14000	Mid-Low	B	rascal	0.99	0.10
71	G.15000	Mid-Low	A	tweezers	0.93	0.26
72	G.15000	Mid-Low	B	bidet	0.97	0.17

73	G.15000	Mid-Low	A	sloop	0.56	0.50
74	G.15000	Mid-Low	B	swingeing	0.36	0.48
75	G.15000	Mid-Low	A	cenotaph	0.39	0.49
76	G.16000	Low	B	denouement	0.50	0.50
77	G.16000	Low	A	bittern	0.48	0.50
78	G.16000	Low	B	reconnoitre	0.37	0.49
79	G.16000	Low	A	magnanimity	0.54	0.50
80	G.16000	Low	B	effete	0.75	0.44
81	G.17000	Low	A	rollick	0.87	0.34
82	G.17000	Low	B	gobbet	0.59	0.49
83	G.17000	Low	A	rigmarole	0.76	0.43
84	G.17000	Low	B	alimony	0.88	0.33
85	G.17000	Low	A	roughshod	0.14	0.35
86	G.18000	Low	B	copra	0.41	0.49
87	G.18000	Low	A	bier	0.36	0.48
88	G.18000	Low	B	torpid	0.41	0.49
89	G.18000	Low	A	dachshund	0.96	0.20
90	G.18000	Low	B	cadenza	0.51	0.50
91	G.19000	Low	A	obtrude	0.59	0.49
92	G.19000	Low	B	panzer	0.47	0.50
93	G.19000	Low	A	cyborg	1.00	0.00
94	G.19000	Low	B	zygote	0.91	0.29
95	G.19000	Low	A	sylvan	0.50	0.50
96	G.20000	Low	B	sagacious	0.32	0.47
97	G.20000	Low	A	spatiotemporal	0.86	0.35
98	G.20000	Low	B	casuist	0.33	0.47
99	G.20000	Low	A	cyberpunk	0.88	0.33
100	G.20000	Low	B	pussyfoot	0.75	0.44

[Continued on next page.]

Table S5. Mean rating (and *SD*) across the 100 participants in experiment 2 for each word on the WordFAM. Number refers to the order in which items appear on the paper-and-pencil version of the WordFAM. The Hoosier Norm vector shows the mean rating for each item in the original Hoosier sample. Bin and version correspond to the divisions made in the current work as described in the main text.

Number	Hoosier Norm	Bin	Version	Word	*Mean*	*SD*
25	1.00	Low	A	inrush	2.65	1.87
46	1.00	Low	B	systolic	3.71	2.20
74	1.00	Low	A	oppress	6.82	0.70
75	1.00	Low	B	braggadocio	2.74	2.33
132	1.00	Low	A	seasonable	6.31	1.32
53	1.33	Low	B	jalousie	1.54	1.08
78	1.33	Low	A	dysprosium	1.89	1.43
105	1.33	Low	B	shibboleth	1.90	1.67
81	1.50	Low	A	chemurgic	1.53	0.96
62	1.67	Low	B	lek	1.45	1.14
24	1.75	Low	A	cenobitic	1.61	1.14
141	1.75	Low	B	jardiniere	1.93	1.47
131	1.83	Low	A	obi	2.59	2.19
26	1.91	Low	B	campanile	1.84	1.45
41	1.92	Low	A	malfeasance	3.97	2.26
35	2.00	Low	B	molybdenum	2.29	2.10
52	2.00	Low	A	bosh	2.86	1.81
56	2.00	Low	B	aileron	2.06	1.82
32	2.08	Low	A	batrachian	1.45	0.98
67	2.17	Low	B	encomium	2.11	1.87
87	2.17	Low	A	bemire	2.05	1.50
99	2.17	Low	B	tinct	2.35	1.80
9	2.18	Low	A	appanage	1.97	1.55
6	2.25	Low	B	palliation	2.58	2.08
80	2.25	Low	A	citify	2.52	1.74
150	2.25	Low	B	inchoate	1.93	1.60
51	2.33	Low	A	puttee	2.01	1.59
94	2.33	Low	B	parquetry	1.79	1.35
110	2.33	Low	A	arable	3.14	2.28
116	2.33	Low	B	duenna	1.63	1.26
18	2.42	Low	A	ferrule	2.01	1.61
21	2.42	Low	B	egregious	5.27	2.21
44	2.42	Low	A	meliorate	2.53	1.84
117	2.42	Low	B	equine	4.59	2.53

76	2.50	Low	A	capstan	1.79	1.39
113	2.50	Low	B	viceregal	1.93	1.42
137	2.50	Low	A	hidalgo	2.69	1.74
139	2.50	Low	B	alembic	1.85	1.55
45	2.67	Low	A	crosier	1.97	1.49
50	2.67	Low	B	aniline	1.91	1.58
65	2.67	Low	A	ennui	2.84	2.40
68	2.67	Low	B	gustatory	2.73	2.14
86	2.75	Low	A	exegesis	2.03	1.42
143	2.75	Low	B	sessile	2.21	1.72
8	2.80	Low	A	mullion	2.01	1.57
28	2.83	Low	B	flivver	1.62	1.44
134	2.83	Low	A	fief	3.00	2.26
27	2.91	Low	B	imprimatur	1.81	1.37
40	2.92	Low	A	triumvir	1.83	1.56
114	2.92	Low	B	torsion	4.14	2.22
123	3.00	Medium	A	mastoid	3.12	2.09
2	3.08	Medium	B	cacophony	4.32	2.51
120	3.08	Medium	A	hemolytic	2.79	2.12
135	3.17	Medium	B	fusillade	2.41	1.90
5	3.25	Medium	A	scintillate	3.36	2.36
57	3.25	Medium	B	undulant	3.28	2.00
63	3.25	Medium	A	czarina	2.37	2.12
133	3.25	Medium	B	transept	2.25	1.69
47	3.33	Medium	A	warder	3.55	2.16
13	3.42	Medium	B	overawe	2.91	1.97
16	3.42	Medium	A	darnel	2.00	1.39
70	3.42	Medium	B	diathermy	1.78	1.24
121	3.50	Medium	A	grandiloquence	2.34	1.84
72	3.55	Medium	B	titivate	1.97	1.59
85	3.67	Medium	A	coitus	4.88	2.43
88	3.67	Medium	B	expatriate	4.41	2.47
146	3.67	Medium	A	triennial	3.08	2.25
98	3.73	Medium	B	expletive	6.05	1.84
34	3.75	Medium	A	smirch	3.22	2.20
36	3.75	Medium	B	dactyl	3.28	2.02
55	3.75	Medium	A	concertina	2.67	2.00
66	3.75	Medium	B	mastodon	4.62	2.29
77	3.75	Medium	A	rasher	3.54	2.32

90	3.75	Medium	B	parboil	3.23	2.46
107	3.75	Medium	A	ruse	6.05	1.66
111	3.75	Medium	B	hullabaloo	5.05	2.29
122	3.75	Medium	A	audiophile	5.76	1.94
96	3.83	Medium	B	exonerate	6.13	1.80
11	3.92	Medium	A	perspicuous	3.57	2.08
12	3.92	Medium	B	philodendron	2.68	2.08
15	3.92	Medium	A	stolidly	2.51	1.91
92	4.00	Medium	B	vestry	2.57	1.81
43	4.17	Medium	A	Pullman	3.31	2.15
119	4.17	Medium	B	conjugal	5.42	1.89
125	4.17	Medium	A	vesture	3.68	1.97
95	4.25	Medium	B	crocus	2.84	2.17
30	4.33	Medium	A	underslung	3.03	2.08
129	4.33	Medium	B	mallow	4.01	2.14
147	4.42	Medium	A	refutation	4.38	2.30
14	4.50	Medium	B	histamine	5.22	1.94
3	4.58	Medium	A	briquette	3.34	2.41
84	4.58	Medium	B	gentry	4.43	2.20
48	4.67	Medium	A	pommel	4.53	2.34
101	4.75	Medium	B	assiduous	3.19	2.12
73	4.83	Medium	A	knobbed	4.84	1.92
130	4.83	Medium	B	tendril	4.97	2.24
10	4.92	Medium	A	affray	3.26	1.99
54	4.92	Medium	B	Pakistani	6.66	1.12
82	4.92	Medium	A	genic	2.89	1.91
142	4.92	Medium	B	radioisotope	4.43	2.12
102	5.00	High	A	freedman	4.52	2.26
115	5.00	High	B	ply	5.80	1.76
140	5.17	High	A	autumnal	4.71	2.39
97	5.33	High	B	denature	4.28	2.32
59	5.50	High	A	goodwife	4.84	1.99
103	5.50	High	B	cessation	4.62	2.29
4	5.83	High	A	deluge	4.70	2.19
17	5.92	High	B	authenticate	6.94	0.28
38	5.92	High	A	defy	6.68	1.09
128	5.92	High	B	morbidity	6.01	1.56
106	6.00	High	A	invigorate	6.12	1.83
20	6.08	High	B	mutt	6.55	1.13

29	6.25	High	A	euphemism	6.16	1.67
1	6.33	High	B	pox	5.35	1.88
39	6.33	High	A	dike	5.27	1.99
118	6.33	High	B	accusal	5.69	1.96
89	6.42	High	A	forked	6.09	1.40
61	6.45	High	B	immobility	6.88	0.41
124	6.50	High	A	drag	6.96	0.24
49	6.55	High	B	drab	6.26	1.47
104	6.58	High	A	handrail	6.11	1.84
109	6.58	High	B	municipal	5.99	1.55
69	6.67	High	A	gab	5.17	2.22
91	6.67	High	B	hedge	6.70	0.98
22	6.75	High	A	destitute	5.60	2.06
33	6.75	High	B	fabrication	6.85	0.44
148	6.75	High	A	antler	6.44	1.60
7	6.83	High	B	misapplication	6.02	1.63
58	6.83	High	A	cannibal	6.97	0.17
60	6.83	High	B	index	6.80	0.60
71	6.83	High	A	quit	6.97	0.30
42	6.92	High	B	inflexible	6.65	1.20
79	6.92	High	A	objectivity	6.61	0.94
127	6.92	High	B	impair	6.58	1.25
19	7.00	High	A	comforter	6.82	0.73
23	7.00	High	B	outcast	6.90	0.63
31	7.00	High	A	greed	6.90	0.63
37	7.00	High	B	ash	6.88	0.46
64	7.00	High	A	grab	6.95	0.33
83	7.00	High	B	educate	6.99	0.10
93	7.00	High	A	mother	7.00	0.00
100	7.00	High	B	cop	6.93	0.33
108	7.00	High	A	central	6.95	0.22
112	7.00	High	B	monologue	6.80	0.71
126	7.00	High	A	glory	6.87	0.66
136	7.00	High	B	classmate	6.93	0.61
138	7.00	High	A	battery	6.97	0.17
144	7.00	High	B	affect	6.88	0.41
145	7.00	High	A	leaves	6.94	0.28
149	7.00	High	B	cancel	7.00	0.00

Table S6. Mean proportion correct (and *SD*) across the 85 participants in experiment 3 for each item on the VST Brief-A. Number and group correspond to those used in the original VST Form A assessment, available at: https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/VST-version-A.pdf. Bin and version correspond to the divisions made in the current work as described in the main text.

Number	Group	Bin	Version	Item	*Mean*	*SD*
1	G.01000	High	Brief-A	see	1.00	0.00
3	G.01000	High	Brief-A	period	0.98	0.15
5	G.01000	High	Brief-A	poor	0.99	0.11
7	G.02000	High	Brief-A	nil	0.88	0.32
9	G.02000	High	Brief-A	circle	0.99	0.11
11	G.03000	High	Brief-A	soldier	1.00	0.00
13	G.03000	High	Brief-A	pro	0.96	0.19
17	G.04000	High	Brief-A	weep	0.98	0.15
19	G.04000	High	Brief-A	cube	0.99	0.11
21	G.05000	High	Brief-A	nun	0.99	0.11
23	G.05000	High	Brief-A	shudder	0.88	0.32
27	G.06000	Mid-High	Brief-A	strangle	0.98	0.15
31	G.07000	Mid-High	Brief-A	quilt	0.91	0.29
33	G.07000	Mid-High	Brief-A	eclipse	1.00	0.00
35	G.07000	Mid-High	Brief-A	ubiquitous	0.64	0.48
37	G.08000	Mid-High	Brief-A	cabaret	0.86	0.35
39	G.08000	Mid-High	Brief-A	veer	0.73	0.45
43	G.09000	Mid-High	Brief-A	candid	0.86	0.35
45	G.09000	Mid-High	Brief-A	psychosis	0.95	0.21
47	G.10000	Mid-High	Brief-A	rouble	0.64	0.48
49	G.10000	Mid-High	Brief-A	puree	0.86	0.35
51	G.11000	Mid-Low	Brief-A	counterclaim	0.93	0.26
53	G.11000	Mid-Low	Brief-A	trill	0.67	0.47
59	G.12000	Mid-Low	Brief-A	emir	0.40	0.49
61	G.13000	Mid-Low	Brief-A	cranny	0.75	0.43
63	G.13000	Mid-Low	Brief-A	azalea	0.84	0.37
65	G.13000	Mid-Low	Brief-A	bawdy	0.72	0.45
67	G.14000	Mid-Low	Brief-A	spangled	0.69	0.46
69	G.14000	Mid-Low	Brief-A	retro	0.88	0.32
71	G.15000	Mid-Low	Brief-A	tweezers	0.85	0.36
73	G.15000	Mid-Low	Brief-A	sloop	0.52	0.50
75	G.15000	Mid-Low	Brief-A	cenotaph	0.39	0.49
77	G.16000	Low	Brief-A	bittern	0.40	0.49

79	G.16000	Low	Brief-A	magnanimity	0.52	0.50
81	G.17000	Low	Brief-A	rollick	0.84	0.37
83	G.17000	Low	Brief-A	rigmarole	0.75	0.43
85	G.17000	Low	Brief-A	roughshod	0.20	0.40
87	G.18000	Low	Brief-A	bier	0.35	0.48
91	G.19000	Low	Brief-A	obtrude	0.46	0.50
95	G.19000	Low	Brief-A	sylvan	0.47	0.50
97	G.20000	Low	Brief-A	spatiotemporal	0.79	0.41
99	G.20000	Low	Brief-A	cyberpunk	0.76	0.43

[Continued on next page.]

Table S7. Mean proportion correct (and *SD*) across the 85 participants in experiment 3 for each item on the VST Brief-B. Number and group correspond to those used in the original VST Form A assessment, available at: https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/VST-version-A.pdf. Bin and version correspond to the divisions made in the current work as described in the main text.

Number	Group	Bin	Version	Item	*Mean*	*SD*
2	G.01000	High	Brief-B	time	0.98	0.15
8	G.02000	High	Brief-B	pub	0.98	0.15
10	G.02000	High	Brief-B	dig	0.99	0.11
12	G.03000	High	Brief-B	restore	1.00	0.00
14	G.03000	High	Brief-B	compound	0.89	0.31
16	G.04000	High	Brief-B	strap	0.93	0.26
18	G.04000	High	Brief-B	haunt	1.00	0.00
20	G.04000	High	Brief-B	butler	1.00	0.00
22	G.05000	High	Brief-B	olive	0.98	0.15
24	G.05000	High	Brief-B	threshold	0.92	0.28
26	G.06000	Mid-High	Brief-B	malign	0.85	0.36
28	G.06000	Mid-High	Brief-B	dinosaur	0.92	0.28
30	G.06000	Mid-High	Brief-B	crab	0.96	0.19
32	G.07000	Mid-High	Brief-B	tummy	0.99	0.11
34	G.07000	Mid-High	Brief-B	excrete	0.94	0.24
36	G.08000	Mid-High	Brief-B	marrow	0.95	0.21
40	G.08000	Mid-High	Brief-B	yogurt	0.99	0.11
42	G.09000	Mid-High	Brief-B	monologue	0.92	0.28
44	G.09000	Mid-High	Brief-B	nozzle	0.98	0.15
48	G.10000	Mid-High	Brief-B	canonical	0.51	0.50
50	G.10000	Mid-High	Brief-B	vial	0.96	0.19
52	G.11000	Mid-Low	Brief-B	refectory	0.31	0.46
56	G.12000	Mid-Low	Brief-B	soliloquy	0.68	0.47
58	G.12000	Mid-Low	Brief-B	augur	0.33	0.47
60	G.12000	Mid-Low	Brief-B	didactic	0.38	0.49
62	G.13000	Mid-Low	Brief-B	lectern	0.56	0.50
64	G.13000	Mid-Low	Brief-B	marsupial	0.74	0.44
66	G.14000	Mid-Low	Brief-B	crowbar	0.96	0.19
68	G.14000	Mid-Low	Brief-B	aver	0.48	0.50
72	G.15000	Mid-Low	Brief-B	bidet	0.91	0.29
74	G.15000	Mid-Low	Brief-B	swingeing	0.31	0.46
76	G.16000	Low	Brief-B	denouement	0.47	0.50
78	G.16000	Low	Brief-B	reconnoitre	0.42	0.50

80	G.16000	Low	Brief-B	effete	0.73	0.45
82	G.17000	Low	Brief-B	gobbet	0.71	0.46
86	G.18000	Low	Brief-B	copra	0.44	0.50
88	G.18000	Low	Brief-B	torpid	0.31	0.46
90	G.18000	Low	Brief-B	cadenza	0.47	0.50
92	G.19000	Low	Brief-B	panzer	0.52	0.50
96	G.20000	Low	Brief-B	sagacious	0.38	0.49
98	G.20000	Low	Brief-B	casuist	0.32	0.47
100	G.20000	Low	Brief-B	pussyfoot	0.65	0.48

[Continued on next page.]

Table S8. Mean rating (and *SD*) across the 85 participants in experiment 3 for each word on the WordFAM Brief-A. Number refers to the order in which items are listed on the paper-and-pencil version of the WordFAM. The Hoosier Norm vector shows the mean rating for each item in the original Hoosier sample. Bin and version correspond to the divisions made in the current work as described in the main text.

Number	Hoosier Norm	Bin	Version	Word	*Mean*	*SD*
25	1.00	Low	Brief-A	inrush	2.87	2.00
78	1.33	Low	Brief-A	dysprosium	2.05	1.51
81	1.50	Low	Brief-A	chemurgic	1.59	1.05
24	1.75	Low	Brief-A	cenobitic	1.86	1.43
131	1.83	Low	Brief-A	obi	2.40	1.73
52	2.00	Low	Brief-A	bosh	2.62	1.72
32	2.08	Low	Brief-A	batrachian	1.52	0.95
87	2.17	Low	Brief-A	bemire	2.24	1.45
9	2.18	Low	Brief-A	appanage	2.02	1.35
80	2.25	Low	Brief-A	citify	2.61	1.81
51	2.33	Low	Brief-A	puttee	2.53	1.78
110	2.33	Low	Brief-A	arable	3.26	2.32
18	2.42	Low	Brief-A	ferrule	2.08	1.49
44	2.42	Low	Brief-A	meliorate	2.32	1.77
76	2.50	Low	Brief-A	capstan	2.04	1.51
137	2.50	Low	Brief-A	hidalgo	2.75	1.82
45	2.67	Low	Brief-A	crosier	2.11	1.39
65	2.67	Low	Brief-A	ennui	2.71	2.28
86	2.75	Low	Brief-A	exegesis	2.44	1.88
8	2.80	Low	Brief-A	mullion	2.14	1.39
134	2.83	Low	Brief-A	fief	3.25	2.31
40	2.92	Low	Brief-A	triumvir	2.18	1.81
123	3.00	Medium	Brief-A	mastoid	2.80	1.90
120	3.08	Medium	Brief-A	hemolytic	2.74	2.04
5	3.25	Medium	Brief-A	scintillate	3.41	2.43
63	3.25	Medium	Brief-A	czarina	2.24	2.09
47	3.33	Medium	Brief-A	warder	3.67	2.16
16	3.42	Medium	Brief-A	darnel	2.06	1.38
121	3.50	Medium	Brief-A	grandiloquence	2.44	1.84
85	3.67	Medium	Brief-A	coitus	4.96	2.41
146	3.67	Medium	Brief-A	triennial	3.29	2.31
34	3.75	Medium	Brief-A	smirch	3.46	2.00
55	3.75	Medium	Brief-A	concertina	2.66	1.93
77	3.75	Medium	Brief-A	rasher	3.36	2.05

107	3.75	Medium	Brief-A	ruse	5.93	1.82
122	3.75	Medium	Brief-A	audiophile	5.58	1.92
11	3.92	Medium	Brief-A	perspicuous	3.98	1.99
15	3.92	Medium	Brief-A	stolidly	2.66	1.82
43	4.17	Medium	Brief-A	Pullman	3.33	2.12
125	4.17	Medium	Brief-A	vesture	3.88	1.89
30	4.33	Medium	Brief-A	underslung	3.85	2.16
147	4.42	Medium	Brief-A	refutation	4.68	2.21
3	4.58	Medium	Brief-A	briquette	3.42	2.36
48	4.67	Medium	Brief-A	pommel	4.29	2.46
73	4.83	Medium	Brief-A	knobbed	5.04	1.98
10	4.92	Medium	Brief-A	affray	3.15	1.96
82	4.92	Medium	Brief-A	genic	3.44	2.02
102	5.00	High	Brief-A	freedman	4.82	2.04
140	5.17	High	Brief-A	autumnal	4.56	2.17
59	5.50	High	Brief-A	goodwife	5.52	1.65
4	5.83	High	Brief-A	deluge	5.15	1.89
38	5.92	High	Brief-A	defy	6.66	1.24
106	6.00	High	Brief-A	invigorate	6.29	1.58
29	6.25	High	Brief-A	euphemism	6.05	1.79
39	6.33	High	Brief-A	dike	5.67	1.86
89	6.42	High	Brief-A	forked	6.46	1.13
124	6.50	High	Brief-A	drag	6.91	0.50
104	6.58	High	Brief-A	handrail	6.40	1.68
69	6.67	High	Brief-A	gab	5.27	2.12
22	6.75	High	Brief-A	destitute	5.82	1.95
148	6.75	High	Brief-A	antler	6.55	1.35
58	6.83	High	Brief-A	cannibal	6.89	0.56
71	6.83	High	Brief-A	quit	6.93	0.65
79	6.92	High	Brief-A	objectivity	6.76	0.85
19	7.00	High	Brief-A	comforter	6.85	0.61
31	7.00	High	Brief-A	greed	6.92	0.38
64	7.00	High	Brief-A	grab	6.99	0.11
93	7.00	High	Brief-A	mother	6.94	0.54
108	7.00	High	Brief-A	central	6.84	0.87
126	7.00	High	Brief-A	glory	6.98	0.22
138	7.00	High	Brief-A	battery	6.89	0.51
145	7.00	High	Brief-A	leaves	7.00	0.00

Table S9. Mean rating (and *SD*) across the 85 participants in experiment 3 for each word on the WordFAM Brief-B. Number refers to the order in which items are listed on the paper-and-pencil version of the WordFAM. The Hoosier Norm vector shows the mean rating for each item in the original Hoosier sample. Bin and version correspond to the divisions made in the current work as described in the main text.

Number	Hoosier Norm	Bin	Version	Word	*Mean*	*SD*
75	1.00	Low	Brief-B	braggadocio	2.95	2.22
53	1.33	Low	Brief-B	jalousie	1.52	0.85
105	1.33	Low	Brief-B	shibboleth	2.02	1.65
62	1.67	Low	Brief-B	lek	1.56	1.15
141	1.75	Low	Brief-B	jardiniere	1.86	1.41
26	1.91	Low	Brief-B	campanile	1.79	1.27
35	2.00	Low	Brief-B	molybdenum	2.39	2.12
56	2.00	Low	Brief-B	aileron	1.98	1.61
67	2.17	Low	Brief-B	encomium	2.19	1.64
99	2.17	Low	Brief-B	tinct	2.65	1.81
6	2.25	Low	Brief-B	palliation	2.68	1.91
150	2.25	Low	Brief-B	inchoate	1.85	1.43
94	2.33	Low	Brief-B	parquetry	1.67	1.20
116	2.33	Low	Brief-B	duenna	1.66	0.99
113	2.50	Low	Brief-B	viceregal	2.20	1.61
139	2.50	Low	Brief-B	alembic	1.95	1.41
50	2.67	Low	Brief-B	aniline	2.09	1.58
68	2.67	Low	Brief-B	gustatory	2.82	1.98
143	2.75	Low	Brief-B	sessile	2.36	1.70
28	2.83	Low	Brief-B	flivver	1.65	1.31
27	2.91	Low	Brief-B	imprimatur	2.27	1.81
114	2.92	Low	Brief-B	torsion	4.11	2.25
2	3.08	Medium	Brief-B	cacophony	4.67	2.53
135	3.17	Medium	Brief-B	fusillade	2.60	1.92
57	3.25	Medium	Brief-B	undulant	3.34	2.10
133	3.25	Medium	Brief-B	transept	2.72	1.87
13	3.42	Medium	Brief-B	overawe	3.75	2.27
70	3.42	Medium	Brief-B	diathermy	2.00	1.50
72	3.55	Medium	Brief-B	titivate	1.96	1.48
88	3.67	Medium	Brief-B	expatriate	4.53	2.36
98	3.73	Medium	Brief-B	expletive	5.86	1.88
36	3.75	Medium	Brief-B	dactyl	3.31	1.96
66	3.75	Medium	Brief-B	mastodon	4.60	2.42
90	3.75	Medium	Brief-B	parboil	3.22	2.37

111	3.75	Medium	Brief-B	hullabaloo	5.13	2.25
96	3.83	Medium	Brief-B	exonerate	6.29	1.57
12	3.92	Medium	Brief-B	philodendron	3.07	2.13
92	4.00	Medium	Brief-B	vestry	3.07	1.97
119	4.17	Medium	Brief-B	conjugal	5.16	2.24
95	4.25	Medium	Brief-B	crocus	3.02	2.13
129	4.33	Medium	Brief-B	mallow	4.44	1.86
14	4.50	Medium	Brief-B	histamine	5.01	2.24
84	4.58	Medium	Brief-B	gentry	4.69	1.99
101	4.75	Medium	Brief-B	assiduous	3.52	1.96
130	4.83	Medium	Brief-B	tendril	5.19	2.25
54	4.92	Medium	Brief-B	Pakistani	6.49	1.33
142	4.92	Medium	Brief-B	radioisotope	4.42	2.17
115	5.00	High	Brief-B	ply	5.71	1.88
97	5.33	High	Brief-B	denature	4.48	2.33
103	5.50	High	Brief-B	cessation	5.02	2.08
17	5.92	High	Brief-B	authenticate	6.80	0.78
128	5.92	High	Brief-B	morbidity	6.26	1.42
20	6.08	High	Brief-B	mutt	6.36	1.42
1	6.33	High	Brief-B	pox	5.66	1.91
118	6.33	High	Brief-B	accusal	5.61	1.87
61	6.45	High	Brief-B	immobility	6.80	0.75
49	6.55	High	Brief-B	drab	5.82	1.96
109	6.58	High	Brief-B	municipal	6.34	1.12
91	6.67	High	Brief-B	hedge	6.85	0.63
33	6.75	High	Brief-B	fabrication	6.75	0.99
7	6.83	High	Brief-B	misapplication	6.46	1.28
60	6.83	High	Brief-B	index	6.79	0.79
42	6.92	High	Brief-B	inflexible	6.76	0.67
127	6.92	High	Brief-B	impair	6.47	1.41
23	7.00	High	Brief-B	outcast	6.89	0.60
37	7.00	High	Brief-B	ash	6.85	0.72
83	7.00	High	Brief-B	educate	6.88	0.59
100	7.00	High	Brief-B	cop	6.96	0.19
112	7.00	High	Brief-B	monologue	6.82	0.73
136	7.00	High	Brief-B	classmate	6.96	0.24
144	7.00	High	Brief-B	affect	6.76	1.08
149	7.00	High	Brief-B	cancel	6.92	0.66

Figure S1. Mean accuracy (proportion correct) by frequency bin on the VST for each subject in experiment 1.

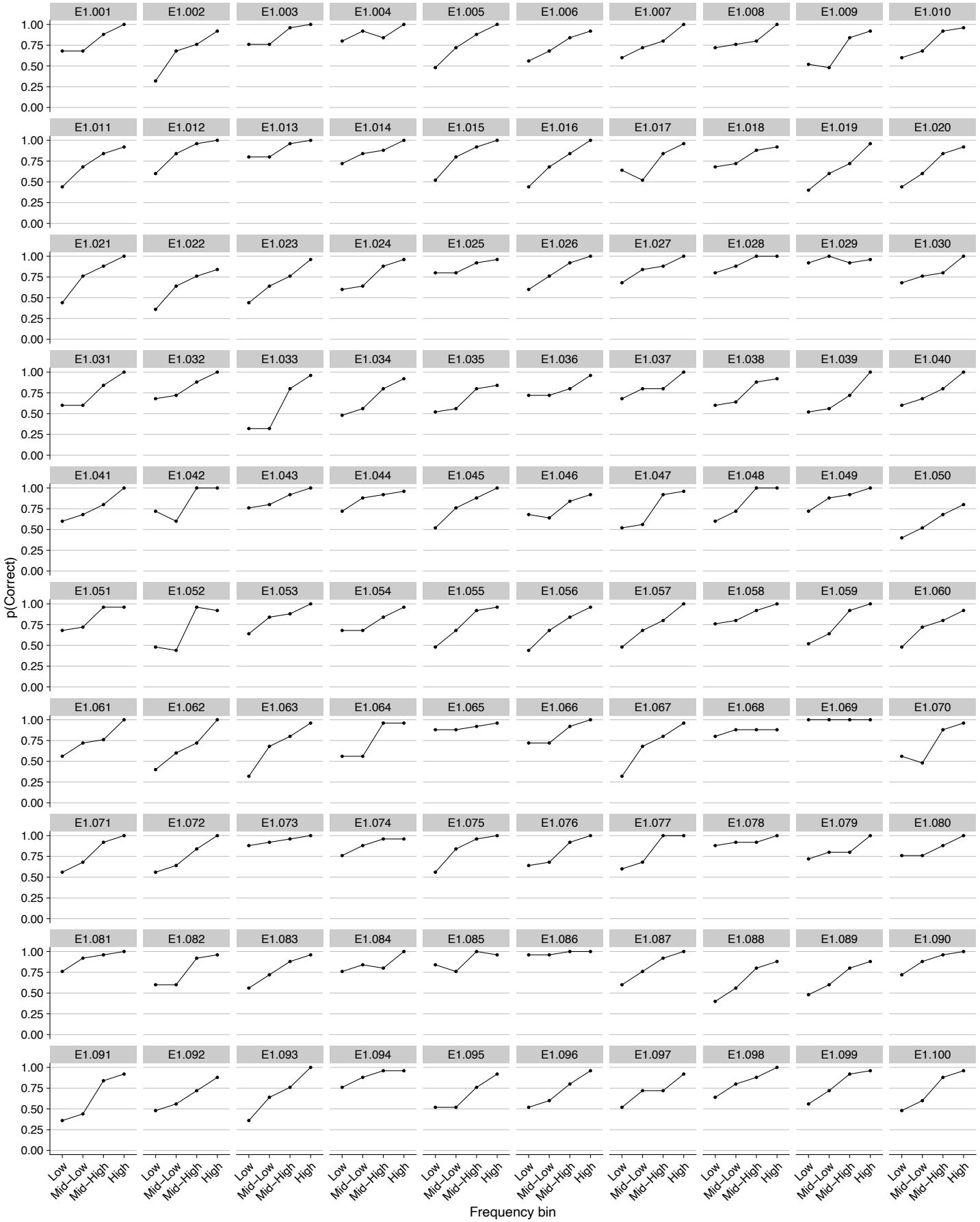


Figure S2. Mean familiarity rating by frequency bin on the WordFAM for each subject in experiment 2.

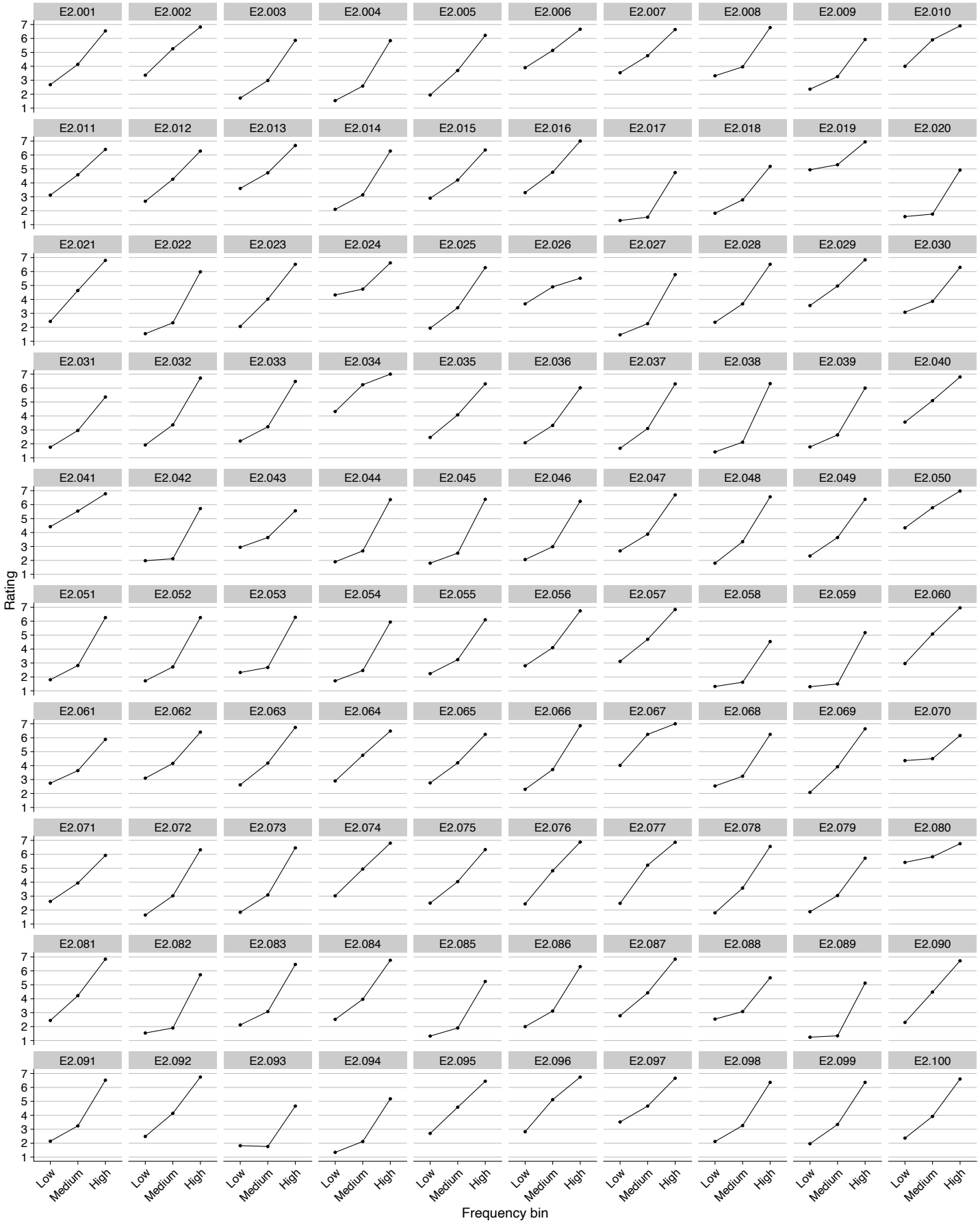


Figure S3. Mean accuracy (proportion correct) by frequency bin on the VST Brief-A and the VST Brief-B for each subject in experiment 3.

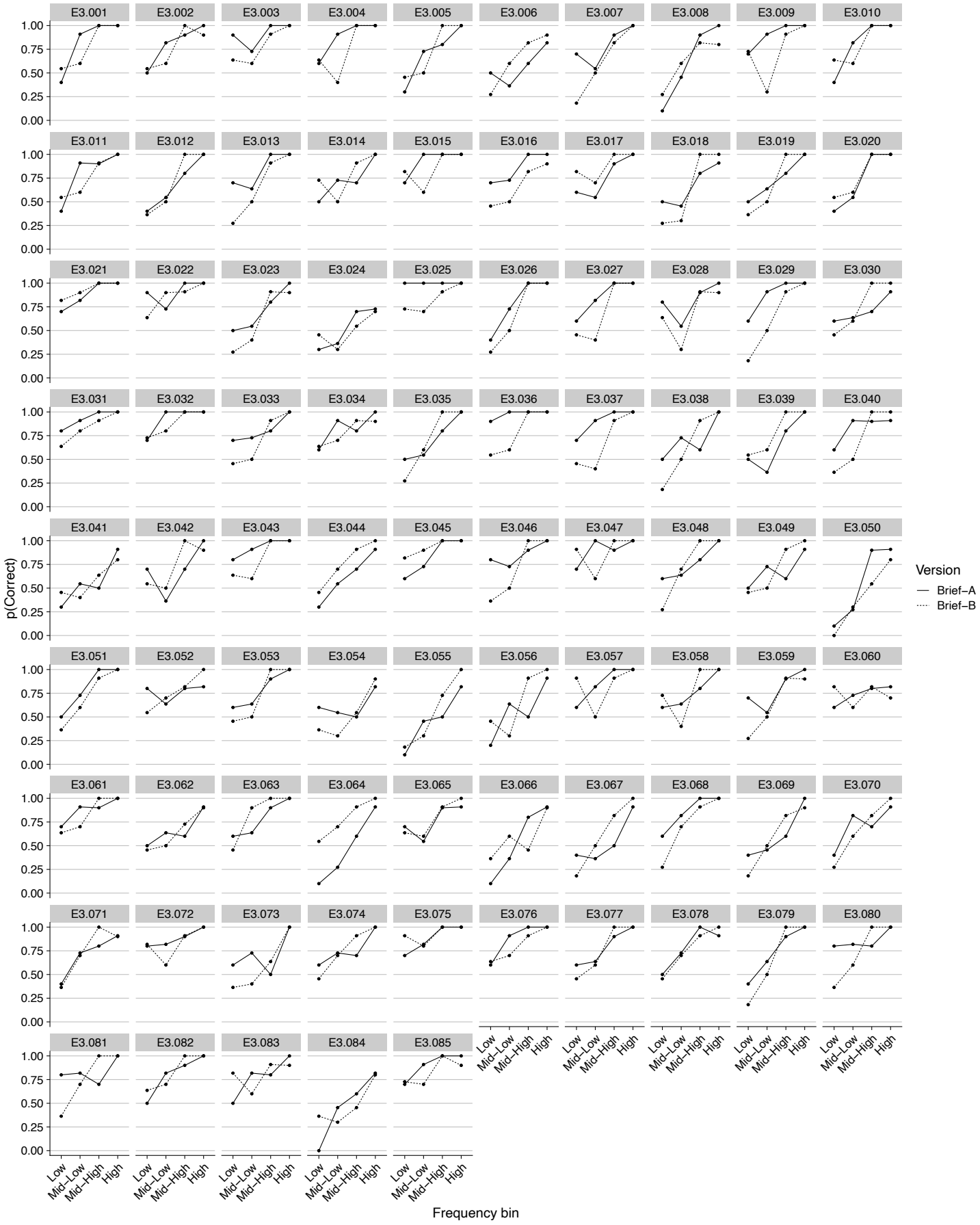


Figure S4. Mean familiarity rating by frequency bin on the WordFAM Brief-A and the WordFAM Brief-B for each subject in experiment 3.