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

Abstract

Two measures for assessing English vocabulary knowledge, the Vocabulary Size Test (VST) and Word Familiarity Test (WordFAM), were recently validated for web-based administration. An analysis of the psychometric properties of these assessments revealed high internal consistency, suggesting that stable assessment could be achieved with fewer test items. Because researchers may use these assessments in conjunction with other experimental tasks, the utility may be enhanced if they are shorter in duration. To this end, two "brief" versions of the VST and WordFAM were developed and submitted to validation testing. Each version consisted of approximately half of the items from the full assessment, with novel items across each brief version. Participants (n = 85) completed one brief version of both the VST and WordFAM at session one, followed by the other brief version of each assessment at session two. The results showed high test-retest reliability for both the VST (r = 0.68) and WordFAM (r = 0.82). The assessments also showed moderate convergent validity (ranging from r = 0.38 to r = 0.59) indicative of assessment validity. This work provides open-source English vocabulary knowledge assessments with normative data that researchers can use to foster high quality data collection in web-based environments.

1 Introduction

Reliable, valid measures of language proficiency can be useful research tools. Such measures could serve to describe a research sample or examine the relationship between broad language proficiency phenotype and specific constructs of interest. Vocabulary is one aspect of linguistic knowledge that contributes to language proficiency (e.g., Bleses et al., 2016; Bloom, 2002; Colby et al., 2018; Gathercole & Baddeley, 1993; Giovannone & Theodore, 2021; Irwin et al., 2002; Landi, 2010; Lewellen et al., 1993; Mancilla-Martinez et al., 2014; Rotman et al., 2020; Snow & Kim, 2007; Tamati & Pisoni, 2014; Theodore et al., 2019; Wasik et al., 2016). Though many standardized assessments of vocabulary exist (e.g., Dunn & Dunn, 1997; Wiig et al., 2013; Williams, 1997), their use in the research domain is challenged due to barriers that include substantial training or specialized degrees required for administration, long administration times, and steep licensing costs. More recently, required in-person administration is viewed as a potential limitation of existing standardized assessments due to safety concerns resulting from the COVID-19 pandemic and the rising adoption of web-based research methodologies.

To address some of these concerns, Drown and colleagues (Under review) recently validated two existing paper-and-pencil assessments for web-based administration, the Vocabulary Size Test and the Word Familiarity Test. 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). Drown et al. adapted Form A of the 20,000 word families VST (Nation, 2012; available at https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/VST-version-A.pdf), which consists of 100 multiple-choice items that sample

vocabulary knowledge across a wide range of lexical frequencies. The Word Familiarity Test (WordFAM; Lewellen et al., 1993; Pisoni, 2007) is a subjective word familiarity rating questionnaire that was developed based on normative data from the Hoosier mental lexicon corpus (Nusbaum et al., 1984). Specifically, the WordFAM requires participants to provide a familiarity rating for 150 English words that sample a wide range of lexical frequencies. In Drown et al. (Under review), a sample of 100 participants completed a web-based administration of the VST and a separate sample of 100 participants completed a web-based administration of the WordFAM. The results demonstrated that the VST and WordFAM long-form assessments are well-suited for web-based administration and provided normative data for their interpretation. For example, both assessments were relatively brief and showed high internal consistency as indicated by Cronbach's alpha and split-half reliability. In addition, both assessments yielded the expected lexical frequency effect, which was stable in the aggregate and at the level of individual subjects. Furthermore, both assessments showed a wide range of item discrimination scores, with lower frequency items showing higher item discrimination scores compared to higher frequency items.

Though the results of Drown et al. (Under review) indicated that administration of the long-form assessments was relatively brief, the high internal consistency for each assessment suggests that stable assessment may be achieved in each task with half the number of trials. Task duration is associated with data quality in web-based research, with shorter tasks expected to yield higher data quality compared to longer tasks (Rodd, 2019). Given that researchers may choose to use vocabulary assessments in conjunction with other experimental tasks, perhaps to screen for the enrollment of "bots" or other low-effort respondents (Godinho et al., 2020; Griffin et al., 2021; Rodd, 2019; Storozuk et al., 2020), the utility of the web-based VST and WordFAM

assessments may be enhanced if they tasks were even more brief. In addition, the normative data gathered for the long-form VST and WordFAM assessments revealed some outlier items for a given lexical frequency bin, consistent with word usage changing over time.

Moreover, the design of Drown et al. (Under review) did not allow for assessment of test-retest reliability, nor did it afford assessment of convergent validity across tasks. 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 (Anastasi & Urbina, 1997). Convergent validity is a subtype of construct validity, which reflects the degree to which a test measures what it is intended to measure, and is necessary to ensure valid interpretation of performance (Anastasi & Urbina, 1997). Convergent validity can be measured, at least in part, by assessing convergence between individuals' performance on two tasks intended to measure the same construct.

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 (e.g., Hedge et al., 2018; Heffner et al., 2022). Unknown task reliability and validity 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 (e.g., Giovannone & Theodore, Under review; Heffner et al., 2022). For example, a recent study of commonly used infant speech perception tasks assessed their test-retest reliability across 13 separate samples and found that only three samples 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. In addition, 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), listening effort (Strand et al., 2018), and lexical reliance (Giovannone & Theodore, Under review) 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. 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.

In this context, the goal of the current work was to develop and validate brief versions of the web-based VST and WordFAM measures for assessing English vocabulary knowledge to supplement the resources developed by Drown et al. (Under review). We aimed to meet the same criteria of Drown et al., which included that the assessments should (1) be openly available for free, public re-use in the research domain, (2) be brief and easy to complete without real-time interaction between the researcher and the participant, and (3) yield acceptable psychometric properties. We also aimed to meet a fourth criterion, which is that the assessments should show high test-retest reliability and convergent validity. To meet this goal, we developed two brief versions of the web-based VST (Beglar & Nation, 2007; Drown et al., Under review) and WordFAM assessments (Drown et al., Under review; Lewellen et al., 1993; Pisoni, 2007) and then submitted the brief, web-based versions to validation testing. Specifically, participants (n = 85) completed a brief version of both the VST and WordFAM at two points in time. Analyses were conducted to determine the suitability of each brief version as an independent assessment of

English vocabulary knowledge, to examine the stability of performance at the individual subject level over time, and to determine the degree to which performance on the two measures were associated.

2 Description of Supplementary Materials

Four Supplementary Materials are provided for this manuscript. First, all experimental tasks described below are available to preview and clone for re-use in Gorilla Open Materials (https://app.gorilla.sc/openmaterials/245615). Second, additional methodological information and analysis results are available in the "SupplementaryMaterials-MethodsResults-Brief.pdf" document. Third, the "SupplementaryMaterials-NormativeData-Brief.pdf" document provides (1) comprehensive demographic characteristics of all participant samples including race, ethnicity, and self-reported dialect, (2) figures illustrating performance for each individual participant, and (3) a complete report of normative data for each item in each assessment. Fourth, a repository that contains trial-level data, analysis code, and materials for all experiments is available at https://osf.io/pcsu6/.

3 Methods

3.1 Participants

Participants (n = 85; 46 men, 39 women) were recruited from the Prolific participant pool (https://www.prolific.co; Palan & Schitter, 2018). The inclusion criteria 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. We note that 15 additional participants completed session one but declined the invitation to also participate in session two and were thus excluded from the study.

3.2. Stimuli

Two versions of each assessment, which we refer to as the Brief-A and Brief-B versions, respectively, were created that (1) contained equal numbers of items in each version of a given assessment (42 items for each VST brief version, 72 items for each WordFAM brief version), (2) reflected unique items across the two brief versions of each assessment, (3) equivalently sampled across frequency bins, and (4) removed items that deviated most substantially from the median item accuracy or median rating based on the results of Drown et al. (Under review). Stimulus details unique to each assessment are provided below; we note that additional details on stimuli including complete stimulus lists are available in the Supplementary Materials.

3.2.1 VST

The 42 items for each brief version were unique subset of the 100 items on Form A of the monolingual (20,000) version of the VST (Nation, 2012; available at https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/VST-version-A.pdf). Each item consists 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*). Items on Form A sample five English words from each of 20 frequency categories that range from extremely high frequency items (e.g., *see*) to extremely low frequency items (e.g., *sagacious*). The 20 frequency categories of the VST are coded as frequency groups that range from 1,000 (lowest frequency items) to 20,000 (highest frequency items) in 1,000 unit bins. As in Drown et al. (Under review), we assigned items to one of four frequency bins (low, mid-low, mid-high, high; each consisting of successive groupings of five consecutive frequency groups) to promote more direct comparison to the WordFAM assessment, which is arranged into low, medium, and high frequency bins.

3.2.2 WordFAM

The 72 items for each brief version were a unique subset of the 150 items on the long-form Word Familiarity Test (WordFAM; Lewellen et al., 1993; Pisoni, 2007). Each item on the long-form WordFAM is a single word, with 50 items in each of three frequency categories: low (e.g., *inrush*), medium (e.g., *undulant*), and high (e.g., *mother*).

3.3. Procedure

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 high rate of return for session two, participants were given a 60-day window in which to complete both sessions. The mean time between sessions was 16 days (*SD* = 10 days, *range* = 1 – 53 days). As described below, the time between sessions did not predict the difference in performance across sessions. Participants were compensated $1.67, reflecting an estimated completion time of 10 minutes.

Procedural details of each task were identical to those described in Drown et al. (Under review), to which the reader is referred for a comprehensive report. In brief, each trial of the VST consisted of a visual array with the prompt displayed at the top of the screen and the four response options displayed below the prompt. On each trial, participants selected which of the four response options best defined the word shown in the item prompt. Each trial of the WordFAM consisted of a visual array, with the Likert rating scale (shown in Table 1) presented at the top of the display. The to-be-rated word appeared below the scale, and the response options were displaced as clickable buttons beneath the word. Participants were directed to rate

their familiarity with the word according to the provided scale.

Table 1: Likert scale used to elicit familiarity ratings for the WordFAM assessment.

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.

4 Results

4.1 Vocabulary Size Test (VST)

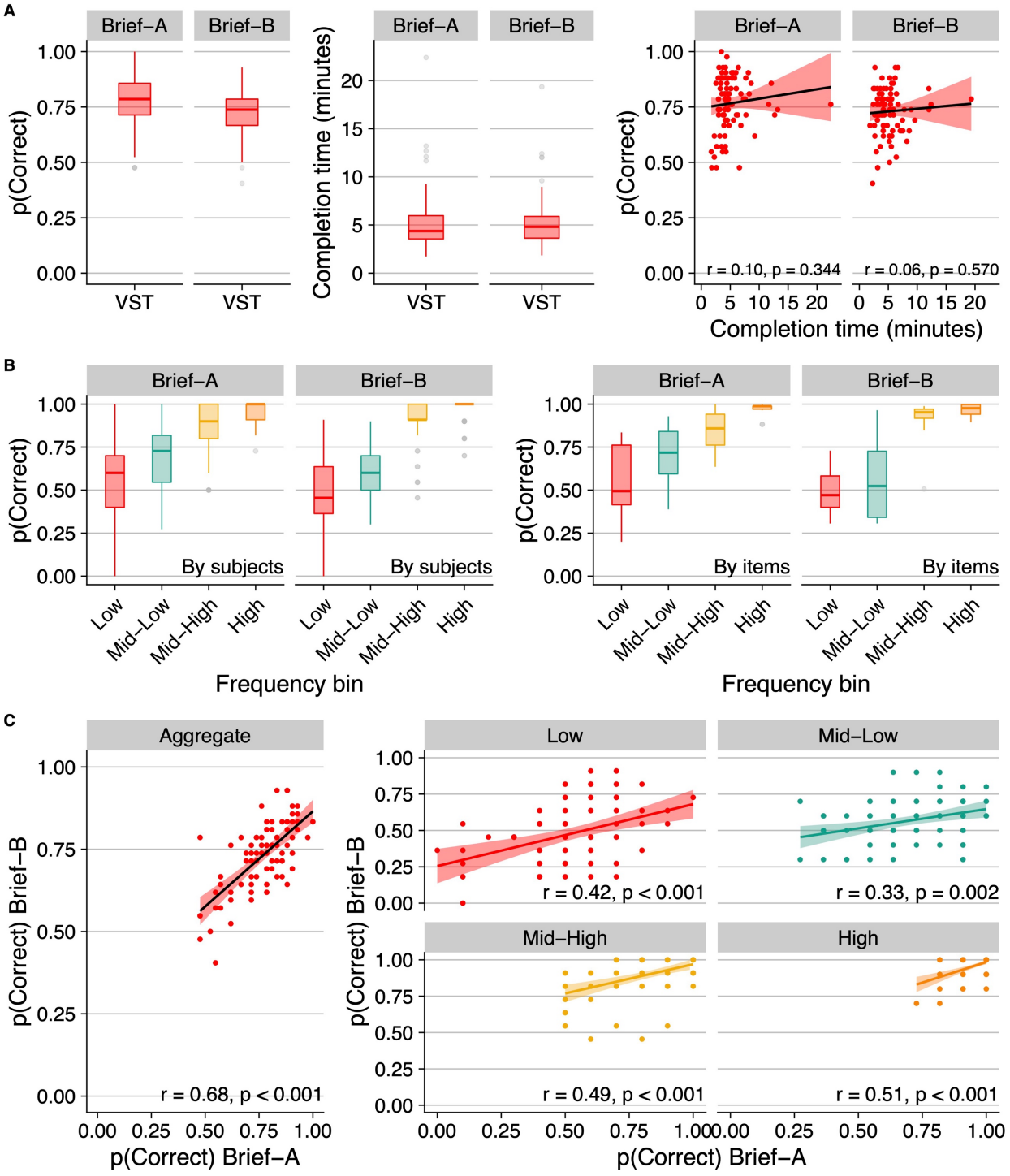
4.1.1 Accuracy and completion time

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 1A, 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 version (Brief-A: $r = 0.10$, $p = 0.344$; Brief-B: $r = 0.06$, $p = 0.570$).

4.1.2 Accuracy by frequency bin

For each VST version, the boxplot distribution of accuracy scores for each frequency bin is shown in Figure 1B by subjects and by items. Statistical analysis (reported in full in the Supplementary Material) showed no significant difference in accuracy between the low and mid-

Figure 1: Results of the Brief-A and Brief-B versions of the VST. 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.



low frequency bins, and a monotonic increase in accuracy between the mid-low and mid-high frequency bins and the mid-high and high frequency bins. Moreover, there was no effect of version, nor did version interact with frequency bin.

4.1.3 Internal consistency

Cronbach's alpha was high for both the Brief-A ($\alpha = 0.80$, *95% CI* = 0.74 – 0.84) and Brief-B ($\alpha = 0.72$, *95% CI* = 0.60 – 0.79) versions of the VST.

4.1.4 Test-retest reliability

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 1C shows the association between accuracy on the brief assessments in the aggregate and separately for each frequency bin. In the aggregate, the brief assessments yielded high test-retest reliability ($r = 0.68$, $p < 0.001$). Test-retest reliability was comparable across frequency bins, which all 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: $r = 0.51$, $p < 0.001$). To examine whether the time between sessions influenced participants' performance, we examined the association between the time interval between sessions (in days) and the difference in mean accuracy between the two versions. As shown in Figure 2A, there was no significant association between the time between sessions and the difference in mean accuracy of the two test versions of the VST ($r = -0.14$, $p = 0.192$).

4.1.5 Item discrimination analysis

The point-biserial coefficient was calculated for each item to determine the association between performance on each individual item (binary; correct vs. incorrect) and performance on all other items (continuous; sum of correct responses). The distribution of item discrimination

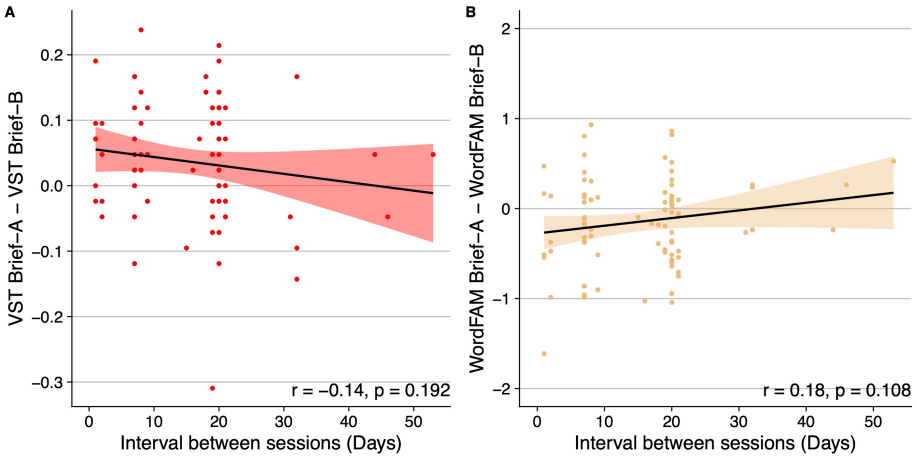


Figure 2: Relationship between time between sessions and difference in performance between sessions for the two brief versions of the VST (panel A) and the two brief versions of the WordFAM (panel B). Individual points show mean by-subject familiarity ratings; the black function indicates the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit.

scores on each brief version are shown in Figure 3. Three items on each version showed uniform ceiling performance across all 85 participants and thus the point-biserial correlation could not be calculated. For the remaining items on the Brief-A version of the VST, the mean point-biserial correlation across items was 0.28 (*SD* = 0.15, *median* = 0.26), with 27 items showing $r \geq 0.20$ (a common criterion for acceptable item discrimination; e.g., McGahee & Ball, 2009). For the remaining items on the Brief-B version of the VST, the mean point-biserial correlation across items was 0.24 (*SD* = 0.15, *median* = 0.24), with 26 items showing $r \geq 0.20$.

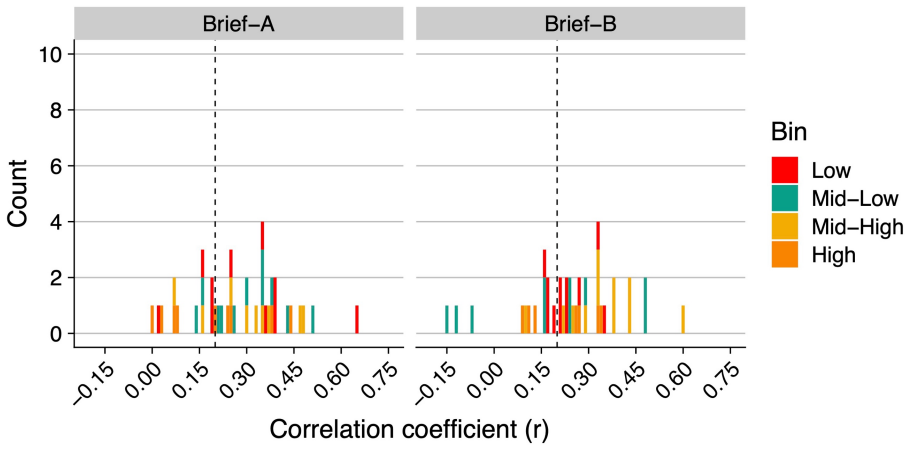


Figure 3: Results of the item discrimination analysis for the brief versions of the WordFAM. The plot shows the distribution of item correlations obtained across WordFAM items, with color used to mark the lexical frequency bin of each item. The vertical dashed line marks $r = 0.20$.

4.2 Word Familiarity Test (WordFAM)

4.2.1 Mean rating and completion time

For each test version, mean familiarity rating was calculated for each participant in addition to completion time. As shown in Figure 4A, 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). There was a small but statistically reliable 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 ($r = 0.09$, $p = 0.392$).

4.2.2 Ratings by frequency bin

The boxplot distribution of mean ratings across subjects for each frequency bin are shown in Figure 4B. In both versions, visual inspection suggests a monotonic increase across the three frequency bins for both the by-subject and by-item rating distributions. This pattern was confirmed by statistical analysis (presented in the Supplementary Materials), which showed a monotonic increase in familiarity ratings from the low to middle frequency bin and from the middle to high frequency bin. The main effect of version was not significant, nor did version interact with frequency bin.

4.2.3 Internal consistency

Cronbach's alpha was high for both the Brief-A ($\alpha = 0.95$, *95% CI* = 0.94 – 0.96) and Brief-B ($\alpha = 0.96$, *95% CI* = 0.94 – 0.97) versions of the WordFAM.

4.2.4 Test-retest reliability

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 4: Results of the Brief-A and Brief-B versions of the WordFAM. 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.

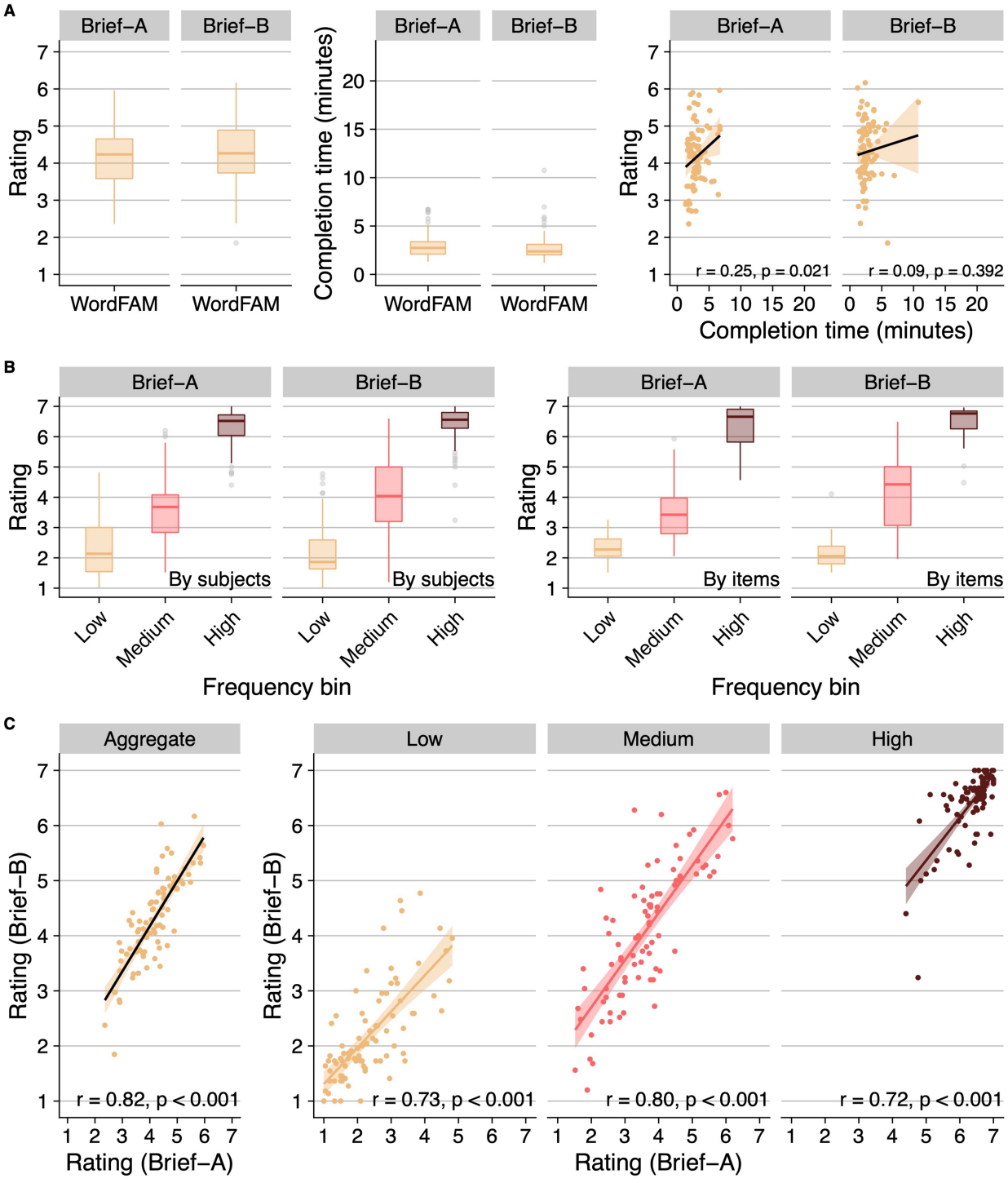


Figure 4C shows the association between familiarity ratings on the brief assessments in the aggregate (at left) and by each of the three frequency bins (at right). Test-retest reliability for the brief assessments was extremely high in the aggregate ($r = 0.82$, $p < 0.001$) and for each of the low ($r = 0.73$, $p < 0.001$), middle ($r = 0.80$, $p < 0.001$), and high ($r = 0.72$, $p < 0.001$) frequency bins. As shown in Figure 2B, there was no significant association between the time between sessions and the difference in mean familiarity ratings of the two test versions ($r = 0.18$, $p = 0.108$).

4.2.5 Item discrimination analysis

The correlation coefficient was calculated for each item to determine the association between performance on each individual item (i.e., the rating on a given item) and performance on all other items (i.e., the mean rating across all other items). The distribution of item discrimination scores for each test version is shown in Figure 5. One item on the Brief-A version of the WordFAM showed ceiling ratings across all participants and thus the item discrimination correlation coefficient could not be calculated. For the Brief-A version of the WordFAM, the mean correlation across items was 0.44 (*SD* = 0.18, *median* = 0.48), with 64 items showing $r \geq 0.20$. For the Brief-B version of the WordFAM, the mean correlation across items was 0.46 (*SD* = 0.13, *median* = 0.47), with 69 items showing $r \geq 0.20$.

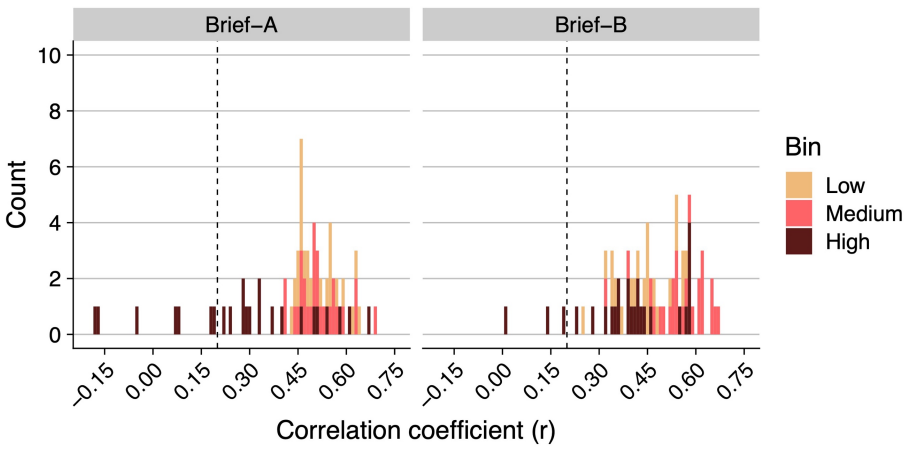
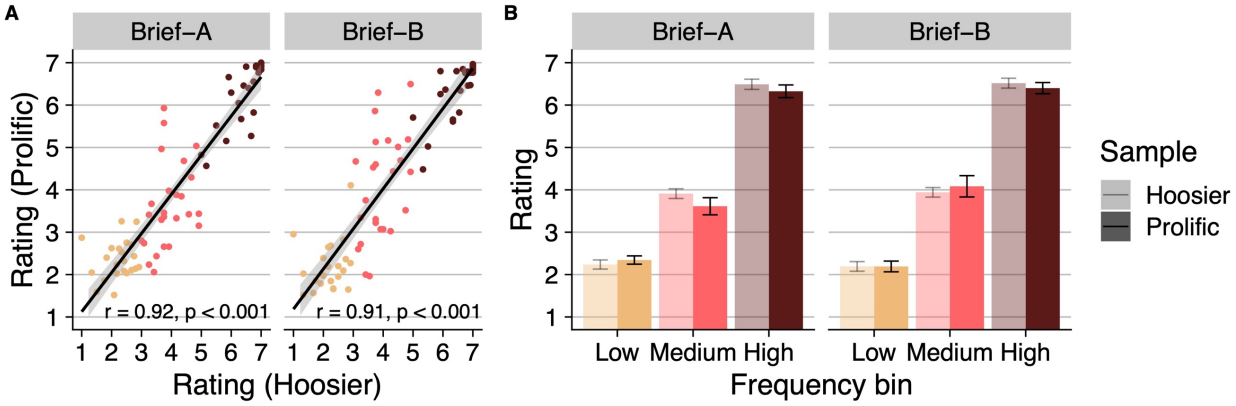


Figure 5: Results of the item discrimination analysis for the brief versions of the WordFAM. The plot shows the distribution of item correlations across items; color marks the lexical frequency bin of each item. The vertical dashed line marks $r = 0.20$.

4.2.5 Comparison between the Prolific sample and existing norms

Performance of the current sample was compared to the existing normative data for the WordFAM (Nusbaum et al., 1984), which were collected in the late 1990s from participants in the Indiana University community (i.e., the Hoosier sample). As shown in Figure 6A, there was 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 6B shows the mean item rating for each frequency bin for each sample, which reveals similar ratings between the two samples for each test version.

Figure 6: Comparison between results of the Brief-A and Brief-B versions of the WordFAM test 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.



4.3 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; these data are shown in Figure 7. 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); $r = 0.47$, $p < 0.001$) and between the VST Brief-B (completed at session two) and the WordFAM Brief-A (completed at session 1; $r = 0.45$, $p < 0.001$).

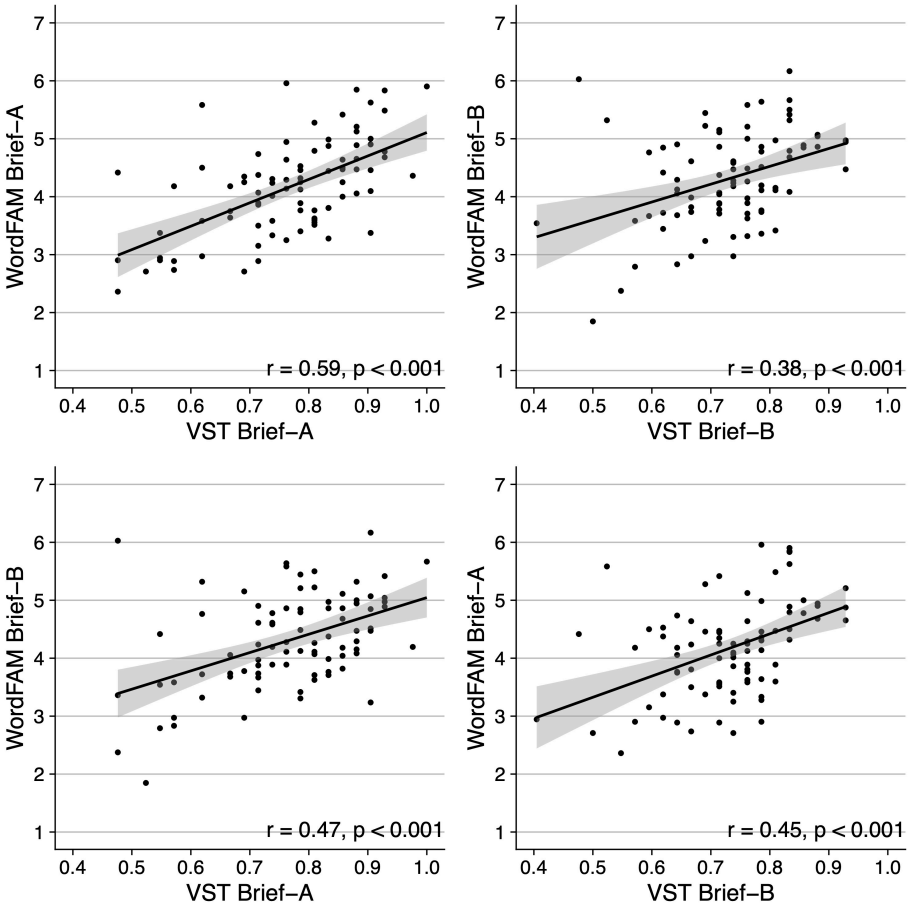


Figure 7: Relationship between performance on the VST and WordFAM assessments. Individual points show mean accuracy (VST) and mean rating (WordFAM) for individual subjects. In all plots, the black function indicates the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit.

5 Discussion

The goal of the current study was to develop and validate two brief web-based measures for assessing English vocabulary knowledge, building on the long-form assessments reported in Drown et al. (Under review). Capitalizing on the high split-half reliability observed for the VST

and WordFAM assessments in Drown et al., here we 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 Drown et al. (Under review). Critically, the current results showed that test-retest reliability of each assessment was strong (VST, $r = 0.69$; WordFAM, $r = 0.82$). Moreover, the two assessments showed moderate convergent validity (ranging from $r = 0.38$ to $r = 0.59$). These results indicate that the web-based vocabulary knowledge assessments developed here are suitable for use in remote research. 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 provided on the OSF repository for this manuscript and thus available for use on other platforms.

Here we outline two avenues for future research that would extend the utility of the assessments developed here. First, as described in the introduction, the VST provides an objective measure of vocabulary competence given that the task requires participants to pick which of four definitions best defines the target word. In contrast, the WordFAM is a subjective measure of vocabulary competency given that it draws on word familiarity ratings. It is not yet known whether the lack of perfect convergent validity between the two assessments reflects the different items used in each assessment or the different way of eliciting vocabulary knowledge in each assessment. In their current forms, there is only one lexical item that is shared between the VST and WordFAM, and thus it was not possible to examine whether familiarity ratings (on the WordFAM) predict definition selection (on the VST) for a shared set of lexical items. Future research that examines performance in the two tasks for the same lexical items could elucidate this relationship.

Second, we note that both assessments provide a model that could be used to develop parallel assessments for additional languages. This is perhaps most straightforward for the WordFAM. The process would entail the following. First, a set of lexical items could be curated to model the range of lexical frequencies sampled in the current WordFAM. The Supplementary Materials provide current frequency values for the English WordFAM items (as obtained from the SubtlexUS corpus, available at http://www.lexique.org/?page_id=241; Brysbaert & New, 2009), which could be used to constrain the selection of lexical items from additional languages. Second, the rating scale should be translated into the language of interest. Finally, validation testing could be conducted to parallel that presented in the current study and in Drown et al. (Under review). To facilitate this charge, the Supplementary Materials include the current Gorilla programs, which could be cloned as a starting point for developing a program to execute validation testing of a novel WordFAM assessment. Moreover, all analysis code for the current work (and Drown et al., Under review) is publicly available, providing an extensive resource for data analysis.

Finally, we encourage the reader to consider the limitations of the VST and WordFAM assessments that were outlined in Drown et al. (Under review), as they also apply to the brief versions of these assessments developed in the current work. Specifically, these measures are not intended to replace existing standardized assessments, nor to provide direct comparisons to these measures, though that is a fruitful avenue for future research. Despite these limitations, the current work suggests that the brief versions of the VST and WordFAM developed here are psychometrically sound assessments that can be used to foster high quality data collection in web-based testing environments.

References

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

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.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

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. (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.

Drown, L., Giovannone, N., Pisoni, D. B., & Theodore, R. M. (Under review). *Validation of two measures for assessing English vocabulary knowledge on web-based testing platforms: Long-form assessments*. https://osf.io/pcsu6/

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.

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.

Giovannone, N., & Theodore, R. M. (Under review). *Do individual differences in lexical reliance reflect states or traits?* https://osf.io/dhybk/

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.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186.

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.

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.

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

McGahee, T. W., & Ball, J. (2009). How to read and really use an item analysis. *Nurse Educator*, *34*(4), 166–171.

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).

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.

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., & 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™ 2). *Journal of the American Academy of Child Adolescent Psychiatry*, *42*, 864–872.