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

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, (3) exhibited high internal consistency, and (4) showed an expected range of item discrimination scores, with low frequency items exhibiting higher item discrimination scores compared to high frequency items. 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 are useful in both the clinical and research domains. In the research domain, measures of language proficiency can serve to describe a sample, group participants based on a given proficiency, or be used to examine individual differences in performance. Vocabulary is one aspect of language proficiency (Bleses et al., 2016; Bloom, 2002; Irwin et al., 2002; Landi, 2010; Mancilla-Martinez et al., 2014; Snow & Kim, 2007; Wasik et al., 2016) that has been associated with other cognitive skills, including phonological working memory, lexical access, language comprehension, and perceptual learning (Colby et al., 2018; Gathercole & Baddeley, 1993; Giovannone & Theodore, 2021; Lewellen et al., 1993; Rotman et al., 2020; Tamati & Pisoni, 2014; Theodore et al., 2019).

Standardized assessments exist to measure vocabulary proficiency (e.g., Dunn & Dunn, 1997; Wiig et al., 2013; Williams, 1997). These assessments provide critical tools for clinicians and researchers alike; however, they are not without limitations. For example, standardized assessments often require substantial training and/or a specialized degree for administration (Wiig et al., 2013), they can be long in duration (Dunn & Dunn, 1997; Wiig et al., 2013; Williams, 1997), and most are licensed by for-profit companies, which introduces a financial barrier to their use. In addition, most standardized assessments are designed to be administered in-person, with the administrator and participant in a shared physical space, which may be viewed as a limitation due to safety concerns stemming from the COVID-19 pandemic and geographical considerations that potentially limit access to research participation for individuals who reside in underserved areas.

Web-based technologies have the potential to address some of these limitations (e.g., Anwyl-Irvine et al., 2020; Palan & Schitter, 2018). However, remote administration of existing

standardized vocabulary assessments is often not possible 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 use these assessments for non-clinical purposes. Though new tools for web-based research show strong promise, 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). For example, web-based research methods afford the possibility of automated enrollment in online studies by software applications, known as "bots," which pose a threat to data integrity (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 may remain regarding whether self-reported demographic information is accurate. For psycholinguistics research, language experience and proficiency are often foundational characteristics of the participant sample that are needed to interpret research findings. In principle, standardized assessments of vocabulary knowledge could provide researchers with a means to verify self-reported language proficiency. 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. To be explicit, we did not aim to develop a comprehensive replacement for existing standardized assessments. Instead, we aimed to meet three 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 indicative of reliable and valid vocabulary assessment. 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 assessment 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. 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, medium, and high frequency 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 provided rating scale.

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, lend themselves well to self-guided completion, and use a lexical frequency manipulation to assess breadth of vocabulary knowledge. Critically, past research provides some evidence to suggest that these assessments are reliable and valid measures of vocabulary knowledge (Beglar, 2010; Coxhead et al., 2014; Lewellen et al., 1993; Nusbaum et al., 1984; Tamati & Pisoni, 2014). For example, a Rasch analysis of the 14,000 word families VST showed that most assessment items showed strong measurement invariance and a good fit to the Rasch model (Beglar, 2010). In addition, Lewellen and colleagues (1993) observed a strong association ($r = 0.72$) between performance on the WordFAM and a standardized

assessment of vocabulary (the vocabulary subtest from the Nelson-Denny Reading Test; Nelson & Denny, 1960) in a sample of 70 participants, suggesting high construct validity for WordFAM. Moreover, the differences between the VST and WordFAM 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 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 in addition to a formal validation of web-based administration. To this end, two 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 the WordFAM. In both experiments, analyses were conducted to characterize select psychometric characteristics of each assessment to gauge the suitability of each measure for web-based testing platforms.

2 Description of Supplementary Materials

Four Supplementary Materials are provided. 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-Longform.pdf" document. Third, the "SupplementaryMaterials-NormativeData-Longform.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 Experiment 1

3.1 Methods

3.1.1 Participants

Participants (n = 100; 47 men, 53 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.

3.1.2. 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/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/VST-version-A.pdf; Nation, 2012). 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*). The items 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 groups that range from 1,000 (lowest frequency items) to 20,000 (highest frequency items) in 1,000 unit bins.

3.1.3 Procedure

All experiments reported in this manuscript were programmed using Gorilla Experiment

Builder (https://gorilla.sc; Anwyl-Irvine et al., 2020), which was also used to control online data collection. A visual display was presented on each trial. The item prompt appeared at the top of the display, the four response options appeared as clickable buttons in the middle of the display, and a progress bar appeared at the bottom of the display. On each trial, participants selected 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 each completed 100 trials, reflecting one unique randomization of the 100 test items. The ISI was 500 ms, timed from the participant's response. Participants were compensated $1.67, reflecting an estimated completion time of 10 minutes.

3.2 Results

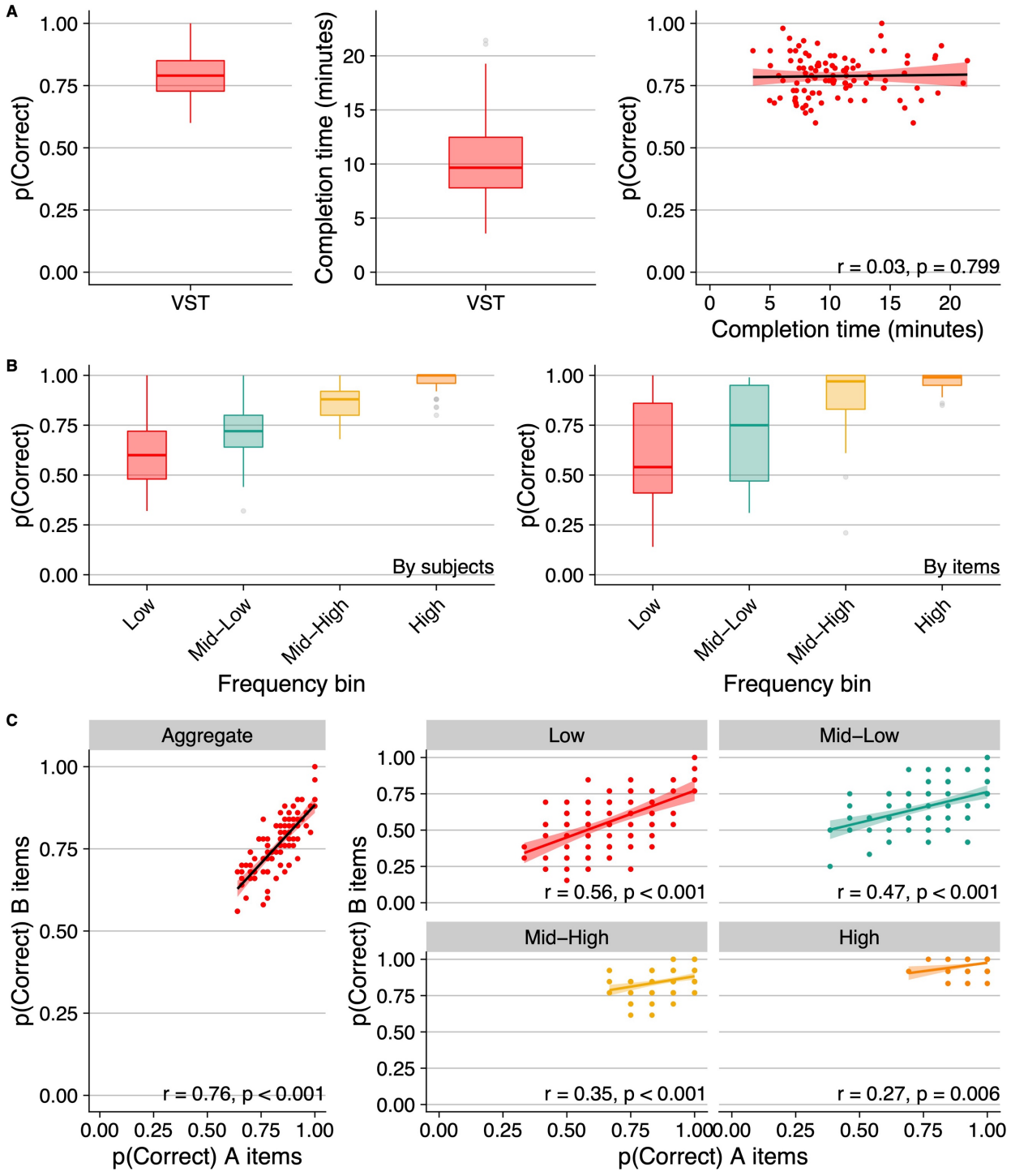
3.2.1 Accuracy and completion time

Accuracy (mean proportion correct) and completion time were calculated for each participant; which is shown in Figure 1A. Mean accuracy across participants was relatively high (0.79, *SD* = 0.08; *range* = 0.60 – 1.00), mean completion time was 11 minutes (*SD* = 4 minutes), and there was no evidence to suggest a speed-accuracy tradeoff ($r = 0.03$, $p = 0.799$).

3.2.2 Accuracy by frequency bin

Recall that the VST was designed to present five items from each of 20 frequency groups. 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, medium, and high frequency bins, the 20 frequency groups were each assigned to one of four frequency bins that consisted of successive groupings of five consecutive frequency groups. Accuracy scores for each frequency bin are shown in Figure 1B both by subjects and by items. Though visual

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.



inspection suggests a monotonic increase in accuracy across the four frequency bins, statistical analysis (presented in the Supplementary Materials) showed no significant change in accuracy between the low and mid-low bins, with monotonic improvement in accuracy from the mid-low bin to the mid-high bin and from the mid-high bin to the high frequency bin.

3.2.3 Internal consistency

Cronbach's alpha was high ($\alpha = 0.85$, *95% CI* = 0.81 – 0.88). As a second measure of internal consistency, split-half reliability was determined by first calculating mean proportion correct for each participant separately for odd- and even-numbered items, which we refer to as 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 1C shows the association between accuracy on the A and B items in the aggregate and by frequency bin. 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.

3.2.4 Item discrimination analysis

The point-biserial coefficient was calculated for each item to determine the association between performance on each individual item and performance on all other items. For example, for item 1, the point-biserial correlation was calculated to determine the association between binary performance on item 1 (i.e., 0 = incorrect, 1 = correct) and the sum of correct responses across items 2 – 99. Seventeen (of 100) items showed uniform ceiling performance across all 100 participants and thus the point-biserial correlation could not be calculated. For the remaining 83

items, the mean point-biserial correlation across items was 0.23 (*SD* = 0.12, *median* = 0.24), with 49 items showing $r \geq 0.20$ (a common criterion for acceptable item discrimination; e.g., McGahee & Ball, 2009). As shown in Figure 2, items in lower frequency bins tended to have higher point-biserial correlations than items in higher frequency bins.

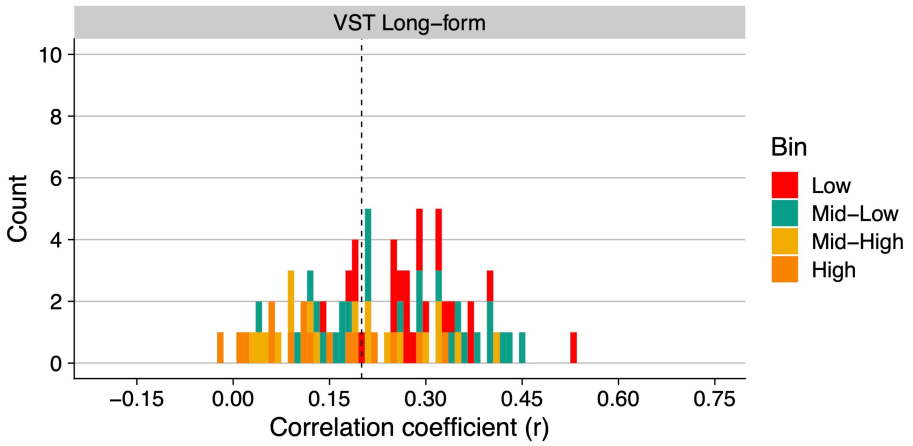


Figure 2: Results of the item discrimination analysis for the VST examined in Experiment 1. The plot shows the distribution of point-biserial correlations obtained across VST items, with color used to mark the lexical frequency bin of each item. The vertical dashed line marks $r = 0.20$, which is one criterion used to indicate an acceptable item discrimination coefficient.

4 Experiment 2

4.1 Methods

4.1.1 Participants

A different sample of participants (n = 100; 48 men, 51 women, one participant who declined to report gender) was recruited from the Prolific participant pool following the inclusion criteria described for Experiment 1.

4.1.2 Stimuli

Stimuli consisted of the 150 items on the Word Familiarity Test (WordFAM; Lewellen et al., 1993; Pisoni, 2007). Each item is a single word. Items sample a wide range of English lexical frequencies, with 50 items in each of three frequency categories: low (e.g., *inrush*), medium (e.g., *undulant*), and high (e.g., *mother*).

4.1.3 Procedure

A visual array was presented on each trial. The Likert scale (shown in Table 1) was presented at the top of the display, 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. On each trial, participants rated their familiarity with the word according to the provided scale. A response was required on every trial and participants were encouraged to guess if they were unsure. Participants completed 150 trials, 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, reflecting an estimated completion time of 15 minutes.

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

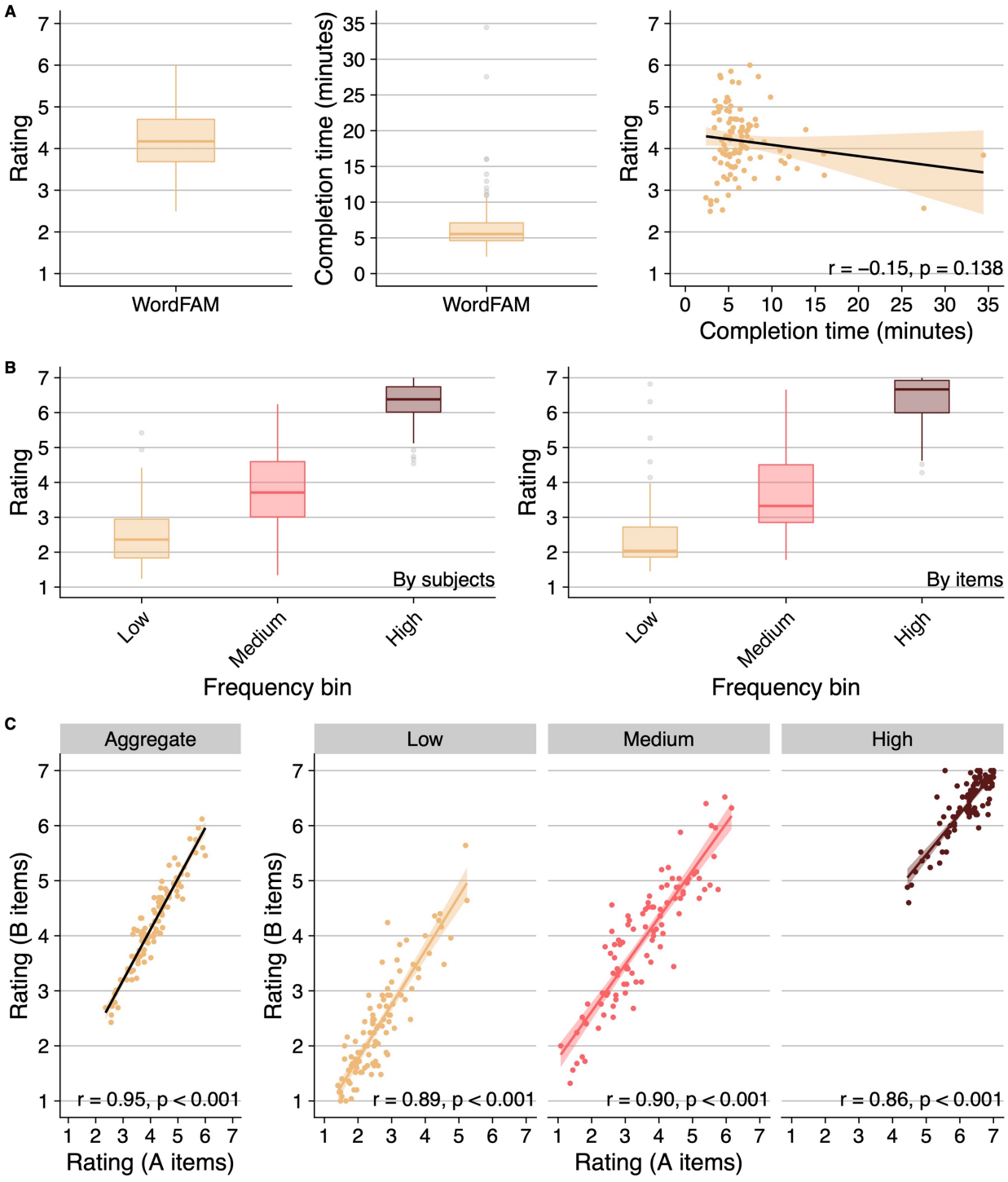
Rating	Reference
1	You have never seen or heard the word before.
2	You think that you might have seen or heard the word before.
3	You are pretty sure that you have seen or heard the word but you are not positive.
4	You recognize the word as one you have seen or heard before, but you don't know the meaning 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.

4.2 Results

4.2.1 Mean rating and completion time

Mean familiarity rating and completion time were calculated for each participant. As shown in Figure 3A, the mean rating across participants was at the center of the Likert scale (4.2,

Figure 3: 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.



SD = 0.8; range = 2.5 – 6.0) and mean completion time was 7 minutes (*SD* = 4 minutes). There was no evidence to suggest an association between participants' mean ratings and completion times (r = -0.15, p = 0.138); this relationship was further attenuated when the two participants exceeding completion times of 20 minutes were excluded (r = -0.04, p = 0.714).

4.2.2 Ratings by frequency bin

The boxplot distribution of mean ratings for each frequency bin is shown in Figure 3B both by subjects and by items. Visual inspection suggests a monotonic increase in accuracy across the three frequency bins for both the by-subject and by-item rating distributions. As described in the Supplementary Materials, this pattern was confirmed by statistical analysis.

4.2.3 Internal consistency

Cronbach's alpha was high (α = 0.98, *95% CI* = 0.97 – 0.98). Split-half reliability was calculated 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 score. This procedure yielded 25 items in each frequency bin for each of the A and B versions, with equivalent normed scores between the two versions for each frequency bin. Figure 3C shows the association between familiarity ratings on the A and B items in the aggregate and separately for each frequency bin. Split-half reliability for the WordFAM was high in the aggregate (r = 0.95, p < 0.001) and for each of the low (r = 0.89, p < 0.001), mid (r = 0.90, p < 0.001), and high (r = 0.86, p < 0.001) frequency bins.

4.2.4 Item discrimination analysis

The correlation coefficient was calculated for each item to determine the association between performance on each individual item and performance on all other items. For example, for item 1, the correlation was calculated to determine the association between the rating

provided for item 1 and the mean rating provided across items 2 – 99. Two (of 150) items showed a uniform ceiling rating across all 100 participants and thus the correlation could not be calculated. For the remaining 148 items, the mean correlation across items was 0.43 (*SD* = 0.16, *median* = 0.48), with 133 items showing $r \geq 0.20$. As shown in Figure 4, items in lower frequency bins tended to have higher correlations than items in higher frequency bins.

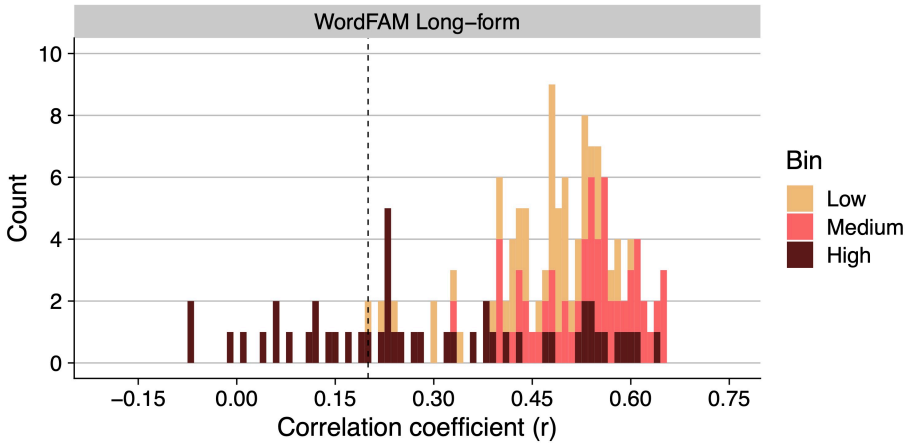
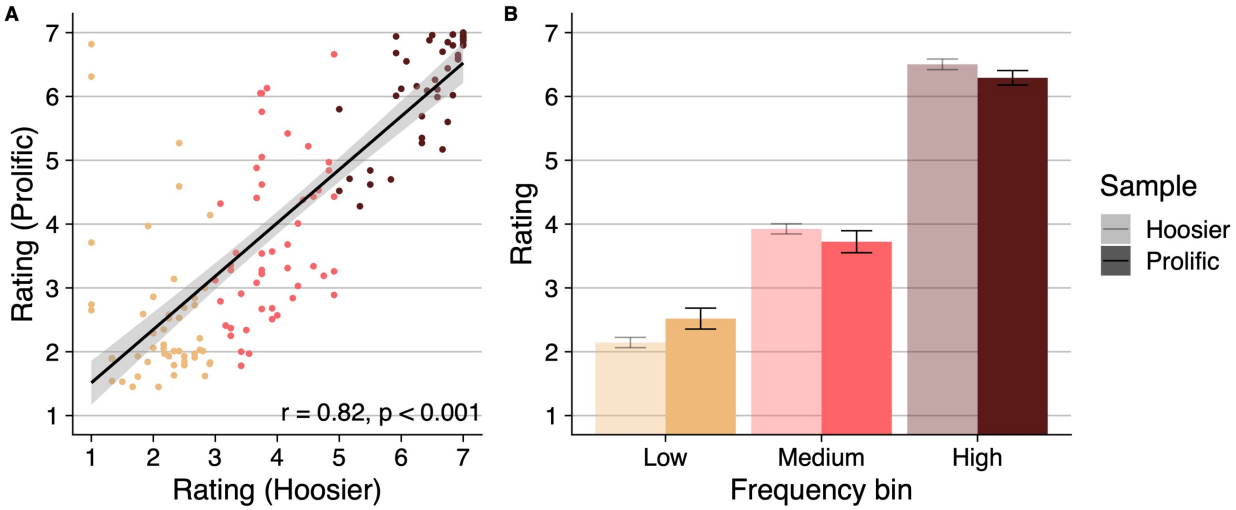


Figure 4: Results of the item discrimination analysis for the WordFAM examined in Experiment 1. The plot shows the distribution of item correlations obtained across WordFAM items, with color used to mark the lexical frequency bin of each item. The vertical dashed line marks $r = 0.20$, which is one criterion used to indicate an acceptable item discrimination coefficient.

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. The existing norms 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; Nusbaum et al., 1984). There was 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 5A). Moreover, the mean familiarity rating for each frequency bin was similar between the two samples (Figure 5B), as confirmed by a statistical analysis that is presented in the Supplementary Materials.

Figure 5: 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.



5 Discussion

The purpose of this study was to develop and validate two web-based measures for assessing English vocabulary knowledge. 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 relatively fast mean completion times and relatively high accuracy. Accuracy was linked to lexical frequency, indicative of assessment validity, and internal consistency (e.g., Cronbach's alpha, split-half reliability) was high, indicative of assessment reliability. In Experiment 2, participants completed a web-based version of the existing long-form WordFAM assessment. We observed minimal variability in completion times, which was relatively fast, and mean ratings were centered on the 7-point familiarity rating scale. Ratings systematically reflected lexical frequency, internal consistency was high, and there was a strong association between the normative data gathered in the Prolific sample and existing

normative data from the Hoosier mental lexicon corpus. Collectively, 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 in the Supplementary Materials to support the use of these assessments on other platforms.

Like any assessment, the measures presented here 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), prior work examining the VST has argued that the multiple-choice format used for this assessment may yield a test of vocabulary recognition rather than vocabulary knowledge, which 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 continue to rely on multiple-choice responses to assess vocabulary knowledge (e.g., Dunn & Dunn, 1997). 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 and at least one

standardized vocabulary assessment, 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).

Furthermore, it is important to highlight that the normative data (see Supplementary Materials) 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 between the current Prolific sample and the existing Hoosier norms for the WordFAM provides some assurance of integrity in participants' self-reported language background.

A final limitation to note is that the current experiments did not measure test-retest reliability or convergent validity (as a subtype of construct validity) of the web-based VST and WordFAM assessments. That is, because each participant only completed one of the two assessments at a single time point, it was not possible to examine whether an individual's performance on a given assessment is stable over time or whether an individual's performance on one assessment is associated with performance on the other assessment. To address this concern, Drown et al. (Under review) developed two brief versions of each assessment, capitalizing on the high split-half reliability that was observed for the long-form assessments. A large sample of participants (n = 85) completed the two brief versions of each assessment at separate timepoints. The results showed high test-retest reliability for both the VST (r = 0.68)

and WordFAM (r = 0.82) and moderate convergent validity between the two assessments (r = 0.38 – 0.59).

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 adults 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 that examines the relationship between performance on the current open-source measures and conventional for-fee, in-person standardized assessments would be fruitful for better understanding the validity of the web-based measures developed here. 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.

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/

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.

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: Brief 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.

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.

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.

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.

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

Nelson, M. J., & Denny, E. C. (1960). *The Nelson-Denny Reading Test: Forms A & B*. Houghton Mifflin.

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.

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.

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*.

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