ASHA American Speech-Language-Hearing Association

JSLHR

Research Note

Examining the Relationship Between Multiple Tests of Receptive Vocabulary

Daphna Harel,[a] Deanna Goudelias,[b] Hung-Shao Cheng,[b] Melissa M. Baese-Berk,[c] Rachel M. Theodore,[d] and Susannah V. Levi[b]

[a] Department of Applied Statistics, Social Science, and Humanities, New York University, New York City [b] Department of Communicative Sciences and Disorders, New York University, New York City [c] Department of Linguistics, The University of Chicago, IL [d] Department of Speech, Language, and Hearing Sciences, University of Connecticut, Storrs

ARTICLE INFO

Article History:
Received October 28, 2022
Revision received May 31, 2023
Accepted November 8, 2023

Editor-in-Chief: Stephen M. Camarata
Editor: Ryan Lee-James

https://doi.org/10.1044/2023_JSLHR-22-00617

ABSTRACT

Purpose: Numerous tasks have been developed to measure receptive vocabulary, many of which were designed to be administered in person with a trained researcher or clinician. The purpose of the current study is to compare a common, in-person test of vocabulary with other vocabulary assessments that can be self-administered.
Method: Fifty-three participants completed the Peabody Picture Vocabulary Test (PPVT) via online video call to mimic in-person administration, as well as four additional fully automated, self-administered measures of receptive vocabulary. Participants also completed three control tasks that do not measure receptive vocabulary.
Results: Pearson correlations indicated moderate correlations among most of the receptive vocabulary measures (approximately r = .50–.70). As expected, the control tasks revealed only weak correlations to the vocabulary measures. However, subsets of items of the four self-administered measures of receptive vocabulary achieved high correlations with the PPVT (r > .80). These subsets were found through a repeated resampling approach.
Conclusions: Measures of receptive vocabulary differ in which items are included and in the assessment task (e.g., lexical decision, picture matching, synonym matching). The results of the current study suggest that several self-administered tasks are able to achieve high correlations with the PPVT when a subset of items are scored, rather than the full set of items. These data provide evidence that subsets of items on one behavioral assessment can more highly correlate to another measure. In practical terms, these data demonstrate that self-administered, automated measures of receptive vocabulary can be used as reasonable substitutes of at least one test (PPVT) that requires human interaction. That several of the fully automated measures resulted in high correlations with the PPVT suggests that different tasks could be selected depending on the needs of the researcher. It is important to note the aim was not to establish clinical relevance of these measures, but establish whether researchers could use an experimental task of receptive vocabulary that probes a similar construct to what is captured by the PPVT, and use these measures of individual differences.

Many people who study speech, language, and hearing use tests of vocabulary as one of the assessments to get information about an individual's language ability. Most vocabulary tests are divided along the dimension of receptive versus expressive vocabulary, with the former being what a person understands and the latter being what a person uses and produces. There are several tests of receptive vocabulary that are commonly used, depending on the age of the participant, including the MacArthur–Bates Communicative Development Inventories (Fenson et al., 2007), the Receptive and Expressive One-Word Picture Vocabulary Tests (Brownell, 2010), and the Peabody

Correspondence to Susannah V. Levi: svlevi@nyu.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

Picture Vocabulary Test (PPVT; Dunn & Dunn, 2007). The PPVT is used widely within the field[1] because it is normed for a wide age range, is easy to administer, and is robust to administration difficulties (e.g., an item can be repeated if attention is briefly disrupted for a client or research participant or if the administrator reads the wrong word). A drawback to these types of vocabulary tests, however, is that they are time consuming and require a clinician or trained researcher to administer the test.

With the onset of the COVID-19 pandemic, many researchers turned to web-based, remote experimental methods that could be executed with minimal-to-no experimenter–participant interaction (e.g., Freeman & De Decker, 2021; Sanchez et al., 2022; Zhang et al., 2021, and a special issue of the journal *Laboratory* Phonology for evidence of researchers turning to remote data collection). Reducing experimenter interaction has several benefits including easing logistical challenges by not requiring scheduling between a participant and an experimenter. Fully automated assessments have the additional benefit of saving time; for example, time that would normally be spent in test administration can instead be spent on other research tasks such as data coding or analysis. In addition, several of the vocabulary tasks below have short versions that take only a small amount of time on the part of the participant. However, it is unclear the extent to which self-administered, automated measures of vocabulary compare to traditional, in-person assessments.

In this study, we examined the relationship between performance on multiple fully automated measures of receptive vocabulary and the PPVT (which requires a trained administrator) as a way to explore the extent to which the fully automated, self-administered measures of vocabulary capture the same construct. In addition to the vocabulary tasks, we also included three control tasks, which were not expected to be correlated with the measures of receptive vocabulary as they measure different aspects of cognition.

Our study focuses on determining and attempting to improve the concurrent validity of multiple fully automated, self-administered tasks for receptive vocabulary (McIntire & Miller, 2007). Concurrent validity, a version of criterion validity used when measures are collected at the same time point, is one desirable psychometric property that a researcher may consider when selecting a task for use in a study, and is determined by a high correlation between a task in question and a previously validated measure.

[1]A search of articles from the *Journal of Speech, Language, and Hearing Research* from January 2013 to June 2023 indicates there are 245 articles that refer to the Peabody Picture Vocabulary Test.

The current study had two primary aims: (a) to determine concurrent validity of a human-administered vocabulary test (PPVT) and automated, self-administered receptive measures of receptive vocabulary; and (b) to explore the extent to which a smaller subset of items may have higher concurrent validity as a measure of receptive vocabulary than the original full-length, self-administered tasks. It is important to note that the aim was not to establish clinical relevance of these measures. Furthermore, the aim was not to determine the clinical validity of using a self-administered version compared to an in-person version. Instead, materials from a range of tasks that ostensibly measure receptive vocabulary were used to establish whether researchers seeking to use remote data collection could use an experimental task of receptive vocabulary that probes a similar construct as what is measured in the PPVT for use as a measure of individual differences.

Method

Participants

Two groups of participants were recruited: a *with PPVT* group (see below) and a *without PPVT* (see below) group. All participants completed a consent form and the study was approved by the institutional review board at New York University. Inclusion criteria included ages 18–40 years, native speaker of U.S. English, and not living outside the United States for more than 6 months before the age of 6 years.

Fifty-five participants ("with PPVT" group) with U.S. IP addresses were recruited through Prolific (2014), resulting in an opt-in sample obtained through this provider, and were paid $20 for their participation. One participant was excluded because English was not their first language, and another was excluded because they completed the PPVT portion but then did not complete the remaining online assessments. The remaining 53 participants (gender self-report fill-in-the-blank: 24 female, 26 male, one gender fluid, one nonbinary, one did not report; aged 18–31 years, M = 26 years) were all raised in the United States and learned English from birth. Ten of the 53 participants indicated a history of speech or language disorders, and were included in the analyses because the goal of the study was to determine the correlations among the tasks, not to exclude people based on an a priori idea of what their vocabulary should be. Four of the 53 participants did not pass the headphone screen (see next section). For these four individuals, their scores on the tasks that required listening to prerecorded stimuli (LexTALE, Nonword Repetition, Recalling Sentences) were entered as NA, but their performance on the other tasks that

involved reading on the screen or interacting with an experimenter over videoconferencing (for the PPVT), where the research assistant was able to determine that the participant did hear the items, were included in the analysis.

For the sole purpose of examining correlations among fully automated, self-administered assessments and increasing the sample size for those analyses, an additional 43 participants ("without PPVT" group) were recruited from the undergraduate psychology research pool at New York University and received course credit for their participation. These participants did not complete the PPVT. Five were excluded because English was not their first language and nine were excluded due to having lived outside of the United States for more than 6 months before the age of 6 years. The remaining 29 participants (gender self-report fill-in-the-blank: 18 female, 11 male; aged 18–26 years, M = 20 years) were all raised in the United States and learned English from birth. Three of these participants indicated a history of speech or language disorders. One of the 29 participants did not pass the headphone screen (see next section), but their data were included for the tasks that did not require listening over headphones.

Procedure

The 53 usable *with PPVT* participants were first provided with a link to complete the PPVT through a video call on a computer. During this session, the PPVT was administered by a trained researcher. At the end of this interactive session, participants were given the link to complete the rest of the remote, self-administered experimental tasks, which included the remaining vocabulary tasks and the control tasks. The 29 usable *without PPVT* participants who did not complete the PPVT were given a link to the remote, self-administered experimental tasks through the research pool at the university.

Prior to the online experimental tasks, participants completed a headphone screen to ensure that they were listening over headphones following Woods et al. (2017). This screen consists of six trials of a three-alternative forced-choice task in which participants must indicate which of three tones is quietest. All tones are 200 Hz with a duration of 1,000 ms and presented dichotically. One of the three tones is lower in amplitude and another one is presented as 180° out of phase across the two channels. If participants are listening over headphones, they will select the stimulus with the absolute lowest amplitude as the quietest stimulus. However, if they are listening in free field (with speakers, not headphones), then the out-of-phase stimulus is selected due to phase cancellation. If participants fail the screen, they are given up to two additional attempts to pass before moving on to the rest of the experiment.

Participants also completed a recording check in which they produced some sample speech which was played back for them to ensure that the computer could pick up their voice for the tasks that required a spoken response (Nonword Repetition, Recalling Sentences). For the experimental tasks that required listening (LexTALE, Nonword Repetition, Recalling Sentences), all sound files were amplitude normalized to the same level to ensure that the volume set at the beginning of the study would be appropriate for all tasks.

The remote, self-administered experimental tasks (both vocabulary and control tasks) were presented in random order. Following the experimental tasks, participants completed a participant questionnaire about their language history. Aside from the PPVT, all tasks (both vocabulary and control tasks) and questions were administered through Gorilla Experiment Builder without a researcher (http://www.gorilla.sc; Anwyl-Irvine et al., 2020, 2021).

The Prolific participant pool was used because it has been established as a high-quality participant pool (Douglas et al., 2023; Palan & Schitter, 2018) and, in our experience, provides a lower barrier to entry for web-based research compared to alternative platforms such as MTurk. Gorilla Experiment Builder was used because it has been established as providing high-quality experimental design and deployment (e.g., high-precision timing) in web-based environments (Anwyl-Irvine et al., 2020, 2021) and because it provides a lower barrier to entry for web-based research because it does not require programming knowledge and it provides extensive support materials (e.g., tutorials, video walkthroughs, support desk; https://gorilla.sc).

Measures of Receptive Vocabulary

Five different tasks that examine a participant's receptive vocabulary were selected. The PPVT (Dunn & Dunn, 2007) was administered by a trained research assistant (see below). Three of the other four tasks (Vocabulary Size Test [VST], Word Familiarity [WordFam], and LexTALE) have open-access versions that can be administered without a researcher or clinician and be conducted on a computer, including in a remote testing context (i.e., without an experimenter present). The Shipley task included items sourced from Table 2 of Shipley (1940). As mentioned above, the aim of this study was not to establish clinical relevance of these measures, but rather to establish whether multiple behavioral tasks probe a similar construct as what is measured in the PPVT. This could allow researchers to use these tasks as a measure of individual differences.

A trained research assistant presented the pictures from Form A of the PPVT (Dunn & Dunn, 2007) via a video call. The images from the PPVT were presented to

participants via screen sharing. All other aspects of the administration of the PPVT were the same as would be given in person; that is, participants had as much time as they wanted to respond and the item was allowed to be repeated. The research assistant read a word and asked the participant to indicate which picture went with the word. During traditional administration, participants point to the picture they believe matches the word. However, all pictures on the PPVT have a number under them that the participant can use rather than pointing. As the participants in the current study were all adults, this did not pose any difficulties. For the purposes of data analysis, each participant's raw score was used (maximum score is 228).

Items from a self-paced, synonym-matching task were sourced from Shipley (1940). This publication indicates that participants were provided with a target word and four additional words on paper. Participants select which of the four options has the same meaning as the target word. The task consists of 40 items, all of which are publicly available in Table 2 of the original publication. In the current study, the target word appeared on the top center of a computer screen and the four response options appeared under in a single row. Rather than marking a response by hand, participants clicked on the response and were then advanced to the next trial. Data were scored for whether participants identified the correct synonym to match the target word.

The VST (Nation & Beglar, 2007) is a self-paced definition-matching task. It was originally administered on paper, but has recently been developed and validated for web-based administration, including a longform assessment (Drown et al., 2023b) and, capitalizing on high split-half reliability observed for the longform assessment, two brief versions that each sample approximately half of the items in the longform assessment (Drown et al., 2023a). The task is open access and publicly available (see Appendix A for a link). In this task, participants see a word in a semantically uninformative sentence and must select the correct definition from among four options. Here, we used the Brief-A version of the web-based VST developed by Drown et al. (2023a), which consists of 42 items. Data were scored for whether participants identified the correct definition of the target word.

In the WordFam task (Lewellen et al., 1993; Pisoni, 2007), participants are presented with a word and asked to rate their familiarity with it on a 7-point scale. The endpoints are labeled "You have never seen or heard the word before" (corresponding to a rating of 1) and "You recognize the word and are confident that you know the meaning of the word" (corresponding to a rating of 7). The original task was self-paced and administered on paper. A self-administered version of the WordFam task has recently been developed and validated including both a longform assessment consisting of 150 items (Drown et al., 2023b) and two brief assessments each consisting of 72 items (Drown et al., 2023a). It is open access and publicly available (see Appendix A for a link). The version used in the current study, the WordFam Brief-A, consisted of 72 items. For the purposes of data analysis, the average rating across all items was used.

The original LexTALE was a lexical decision task designed to assess vocabulary among second language learners of English (Lemhöfer & Broersma, 2012). In the original task, participants were presented with written versions of words and nonwords on a computer screen and indicated if the item was a real word or a nonsense word. The Modified LexTALE task open access (see Appendix A for a link) and is presented auditorily on a subset of the original items, eliminating several of the highest frequency words (Babel, 2020). Participants hear an item and are asked to respond as quickly and accurately as possible. This version of the task consists of 60 items. Data were scored for whether participants correctly identified each item as either a word or nonword. As this version of the LexTALE is administered auditorily, participants who failed the headphone screener were not included in analyses related to this task.

Control Tasks

Three additional experimental tasks were included that were expected to have only low-to-moderate correlations with the above five measures of receptive vocabulary. Although two of the tasks (Nonword Repetition and Recalling Sentences) examine some component of language processing, they are not designed to explicitly examine receptive vocabulary. The inclusion of these control tasks provides a way to ensure that any observed correlations among vocabulary assessments are not spurious; namely, that people who are highly motivated generally perform better on all tasks, whereas someone who is not motivated or not following directions will perform poorly on all tasks. These control tasks were administered online without a researcher.

A standard flanker task (Eriksen & Eriksen, 1974) was used as a control task and was expected to not correlate highly with vocabulary. In the flanker task, participants saw a central arrow and two additional flanking arrows on either side. Participants were asked to indicate the direction of the central arrow as quickly and as accurately as possible. On half of the trials, all arrows faced the same direction (e.g., > > > > >, "congruent trials") and on half the central arrow faced a different direction from the flanking arrows (e.g., > > < > >, "incongruent trials"). The flanker task assesses inhibition skills. To get each participant's flanker score, only responses to correct

trials were used. For these trials, reaction time (RT) greater than 2,000 ms were excluded (less than 0.25%). From the remaining trials, the average of the log RT for congruent trials minus the average of the log RT for incongruent trials was used as a participant's inhibition score. Participants are expected to have a negative flanker score because people are expected to respond more quickly on congruent trials (shorter RT) than on incongruent trials (longer RT). The larger the difference is between the congruent and incongruent trials, the poorer a participant's inhibition skills.

The Nonword Repetition subtest of the Clinical Test of Phonological Processing (CTOPP; Wagner et al., 1999) was used as a control task as it measures phonological working memory and is not expected to be highly correlated with vocabulary. Although some studies have found moderate correlations between nonword repetition and vocabulary (e.g., Gathercole, 1995; Gathercole & Baddeley, 1989; Gathercole et al., 1992; Metsala, 1999), others have pointed out that both nonword repetition and vocabulary development rely on a range of cognitive and linguistic functions and that the strength of the relationship (correlation) between nonword repetition and vocabulary declines across development (Gathercole, 2006; Melby-Lervåg et al., 2012).[2] Because the current study was with adult participants, our expectation was that nonword repetition would not be highly correlated with the measures of vocabulary.

In the Nonword Repetition task, prerecorded nonsense words are presented only once and participants repeat them aloud. The test begins with one syllable items and continues to longer items. All items were presented to participants. Recordings were transcribed offline by two trained research assistants. If the participant met the ceiling rule before the last item, then that ceiling item was treated as the final item for that participant, in line with how the test is scored when administered in person. For the purposes of data analysis, the scaled score from the CTOPP was used.

The Recalling Sentences subtest of the Clinical Evaluation of Language Fundamentals–Fourth Edition (Semel et al., 2003) was used as a control task. In this task, we used recordings of the sentences to control the presentation across all participants. Participants hear sentences only once and must repeat them aloud. All items were presented but if a participant met the ceiling rule before the last item, their final score reflected only the items up to the ceiling rule. The raw score was used in data analysis.

[2]It is worth noting that the correlation between nonword repetition and vocabulary size may be significant and may technically count as moderate in strength, but the actual correlation itself it not especially strong, mostly less than $r = .50$.

Statistical Analysis

We first calculated Pearson correlation coefficient between the scores on the PPVT and the other receptive vocabulary tasks and control tasks. The correlations between the PPVT and the full-length forms of the other receptive vocabulary tasks therefore serve as a baseline concurrent validity (Mislevy & Rupp, 2010) value that we hoped to increase by selecting a subset of items to score through the resampling analysis.

We then attempted to find subsets of items that would lead to an increase in concurrent validity as measured through the correlation with the PPVT. To do this, for each of the receptive vocabulary tasks and for each possible size of a shortened task (1 to the number of items in the full task), we resampled the full list of items (without replacement) 100,000 times and calculated the correlation between that subsample of items and the PPVT. For each possible size of a shortened task, we calculated the maximum correlation across the 100,000 resamples and the PPVT to find the set of items with the largest correlation. In this case, the goal is not necessarily to shorten the number of items that a participant would complete, but rather to assess whether a subset of the full task would achieve a higher correlation with the PPVT than the baseline correlation we obtained. This is similar to previous work done to shorten surveys and item lists in order to obtain a shortened form that has better measurement properties (such as a correlation) than the full-length form originally had (Harel et al., 2018).

We used a resampling procedure rather than examining all possible subsets of the full task because, for example, the WordFam task consists of 72 items and there are 1.64×10^{20} possible subsamples of length 30 that could be examined, a number far greater than the computational power of a standard computer. It is important to note that the goal here was not to assess statistical significance of any specific correlation, nor to set an a priori correlation level above, which a correlation would be considered *good enough*, but rather to explore what the strength of the relationship is between various measures of receptive vocabulary.

Results

Descriptive Statistics

Information about performance on the eight tasks is provided in Table 1. This includes the minimum and maximum scores, mean, and median. The column for PPVT includes only the 53 participants who completed that

Table 1. Descriptive statistics on the eight tasks.

A	Receptive vocabulary tasks					Control tasks		
	PPVT (*n* = 53)	**Shipley (*n* = 82)**	**VST (*n* = 82)**	**WordFam (*n* = 82)**	**LexTALE (*n* = 77)**	**Flanker (*n* = 82)**	**NWR (*n* = 54)**	**Recalling Sentences (*n* = 46)**
Min	186	0.15	0.33	1.70	0.58	–0.19	7.00	64.00
Max	221	0.92	0.90	5.69	0.95	0.05	17.00	96.00
M	208	0.76	0.71	4.24	0.80	–0.08	11.80	85.40
SD	8.06	0.11	0.09	0.69	0.07	0.05	2.53	7.98
Mdn	210	0.77	0.71	4.30	0.80	–0.08	11.00	87.00
B	**Receptive vocabulary tasks**					**Control tasks**		
	PPVT (*n* = 53)	**Shipley (*n* = 53)**	**VST (*n* = 53)**	**WordFam (*n* = 53)**	**LexTALE (*n* = 49)**	**Flanker (*n* = 53)**	**NWR (*n* = 36)**	**Recalling Sentences (*n* = 27)**
Min	186	0.48	0.52	2.94	0.65	–0.19	8.00	64.00
Max	221	0.92	0.90	5.69	0.95	0.04	17.00	96.00
M	208	0.79	0.73	4.28	0.82	–0.08	12.05	86.96
SD	8.06	0.08	0.08	0.63	0.06	0.05	2.49	7.87
Mdn	210	0.80	0.71	4.34	0.82	–0.08	12.00	89.00

Note. The values in each column are as follows: PPVT is raw scores; Shipley, VST, and LexTALE are proportion correct; WordFam is the average familiarity rating across all items; Flanker is a difference score of congruent minus incongruent trials on the logged RT responses; NWR is the scaled score; Recalling Sentences is the raw score. Table 1A includes participants in both the *with PPVT* and *without PPVT* versions. Table 1B includes only the participants in the *with PPVT* condition, as these are the participants who form the core of the resampling analysis. Sample sizes are included in the headings. Note that some tasks have a smaller sample size either because participants failed the headphone screener or because participant recordings were cut off due to experimental error. PPVT = Peabody Picture Vocabulary Test; VST = Vocabulary Size Test; WordFam = Word Familiarity task; NWR = Nonword Repetition.

version of the experiment (*with PPVT*), whereas all other columns include the additional 35 participants who only completed the online tasks (*with PPVT* and *without PPVT*).

Concurrent Validity Analyses

We report the Pearson correlation values between all of the eight tasks in Table 2. These correlations represent the full data set, and thus aside from the column for PPVT, the correlations also include the set of participants who completed only the online tasks. It should be noted that there were missing data for some of the tasks, thus the actual sample size for several of the correlations is reduced.

First, these correlations indicate weak correlations between the three control tasks (Flanker, Recalling Sentences, Nonword Repetition [NWR]) and any of the receptive vocabulary measures (all $r < \pm.40$ and most <

Table 2. Pearson correlations, denoted *r*, among all tasks, both vocabulary tasks (Peabody Picture Vocabulary Test [PPVT], Shipley, Vocabulary Size Test [VST], Word Familiarity [WordFam], LexTALE) and control tasks (Flanker, Nonword Repetition [NWR], Recalling Sentences).

Task	PPVT	Shipley	VST	WordFam	LexTALE	Flanker	NWR
Shipley	**.71** (*n* = 53)						
VST	**.68** (*n* = 53)	**.64** (*n* = 82)					
WordFam	**.51** (*n* = 53)	**.61** (*n* = 82)	.40 (*n* = 82)				
LexTALE	.46 (*n* = 49)	**.53** (*n* = 77)	.47 (*n* = 53)	.24 (*n* = 53)			
Flanker	–.12 (*n* = 53)	–.11 (*n* = 82)	–.17 (*n* = 82)	–.02 (*n* = 82)	–.04 (*n* = 53)		
NWR	.28 (*n* = 36)	.17 (*n* = 54)	.02 (*n* = 54)	.24 (*n* = 54)	.20 (*n* = 54)	–.05 (*n* = 54)	
Recalling Sentence	.20 (*n* = 27)	.34 (*n* = 46)	.05 (*n* = 46)	.38 (*n* = 46)	.08 (*n* = 46)	.20 (*n* = 46)	.49 (*n* = 42)

Note. Correlations greater than ±.50 are represented in bold.

±.30). Second, the majority of the vocabulary task correlations are above .50, with the exception of three of the LexTALE correlations and the WordFam–VST correlation. Using these simple correlations, the two most highly correlated with the PPVT are the Shipley task (.71) and the VST (.68).

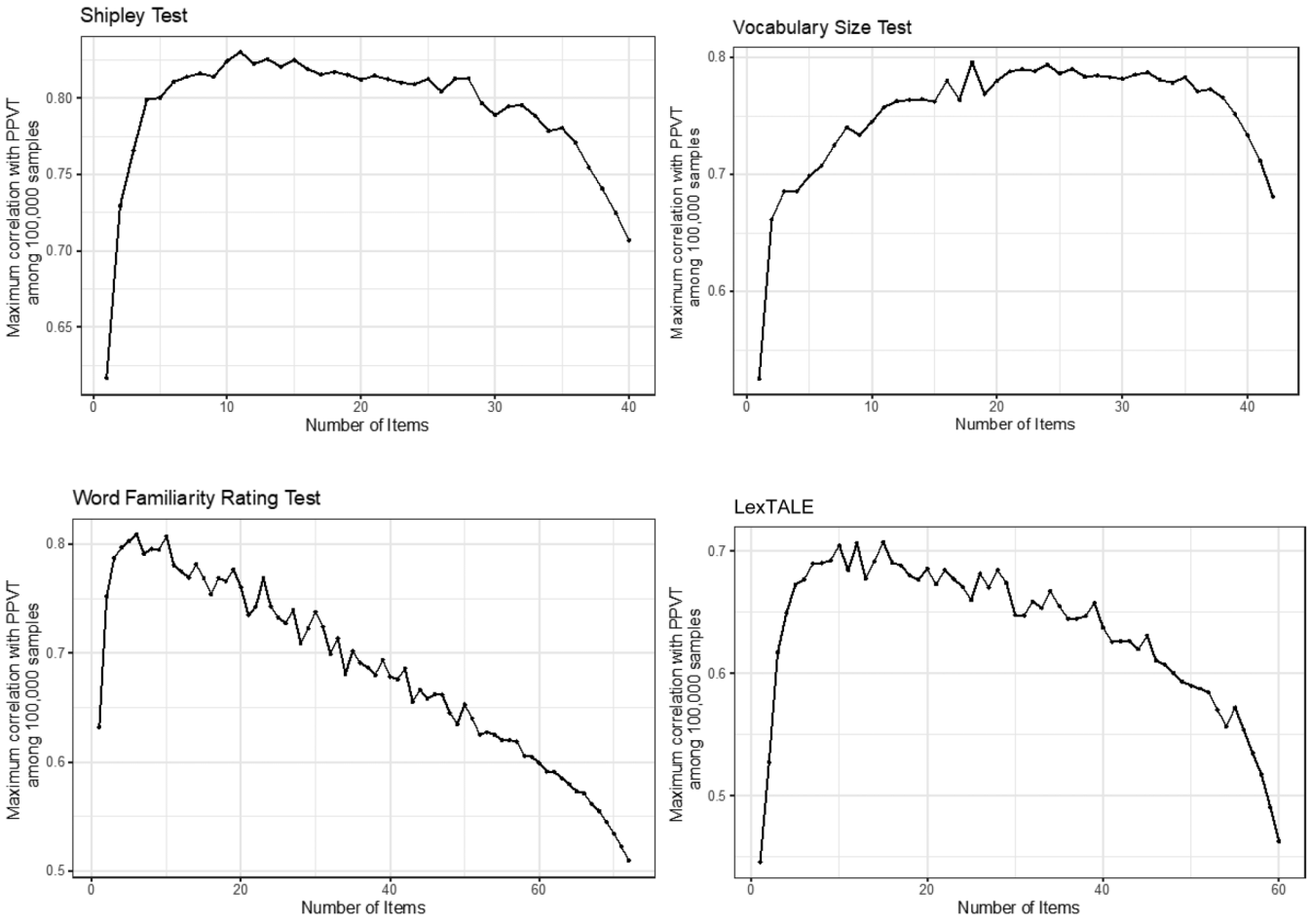
Resampling Analysis

We present our full findings in Figure 1. A subset of 11/40 items from the Shipley achieved a correlation with the PPVT of .83, an increase of .12 from the full-length task. A subset of 18/42 items from the VST achieved a correlation with the PPVT of .80, an increase of .12 from the full-length task. A subset of 6/72 items from the WordFam task achieved a correlation with the PPVT of .81, an increase of .30 from the full-length task. Lastly, a subset of 15/60 items from the LexTALE achieved a correlation with the PPVT of .71, an increase of .25 from the full-length task. A full list of items for each shortened task is provided in Appendix B. All of these correlations with these reduced sets of items achieved what is typically considered strong correlations (above .70; Akoglu, 2018; Cohen, 1988; Schober et al., 2018).

Discussion

The goal of this study was to establish how closely self-administered measures of receptive vocabulary could capture similar performance to the oft-used PPVT. We were motivated in part due to the need to move most studies to a remote format as the result of the COVID-19 pandemic, but also to understand whether multiple tasks that purportedly measure receptive vocabulary are, indeed, measuring the same construct. Our results suggest only moderate correlations among tasks that ostensibly measure receptive vocabulary when all items are included. Perhaps this is not surprising given that these tasks vary in which lexical items are included, and in how responses are

Figure 1. Maximal correlations between shortened versions of each task and the Peabody Picture Vocabulary Test (PPVT) for all possible subsets of items based on 100,000 resamples.



provided (e.g., lexical decision, picture matching, definitions, etc.). The subsetting analysis, however, which examined subsets of items offers more promising results, with correlations substantially higher. These findings suggest that using a self-administered, fully automated measure is reasonable, especially when trying to reduce time for the participant and for the researcher.

The strength of the correlations with the full list of items largely aligned with our expectations. In general, the correlations among the measures of receptive vocabulary were higher than among other tasks ($\geq .50$ on most comparisons), though with some variation. The vocabulary tasks with the lowest correlations with the other measures were WordFam and LexTALE. For WordFam, this may not be entirely surprising, since it is a measure of a participant's self-assessment of vocabulary familiarity, rather an accuracy measure. Thus, participants likely vary in how well they can assess their own abilities. Indeed, studies of self-assessment of language ability also reveal that self-assessment does not align with objective measures (Tomoschuk et al., 2019; Trofimovich et al., 2016). For LexTALE, the lower correlations could result from the assessment; namely that the LexTALE is a lexical decision task (and thus is timed), whereas the other measures of vocabulary (PPVT, Shipley, VST) are self-paced and allow participants to think and reflect.

In terms of our control tasks (Recalling Sentences, NWR, Flanker), we confirmed lower correlations between these tasks and the vocabulary tasks. Not surprisingly, the lowest correlations were found between the flanker task and the vocabulary measures (all $\pm.12$), likely because the flanker task does not rely on any language functioning.

Our second set of analyses indicated that is possible to obtain much higher correlations between the self-administered measures of receptive vocabulary and the PPVT when examining performance on a subset of items. The set of selected items is neither similar in length (ranging from 6 to 18 selected items), nor similar in which items (though to be clear, the sampling procedure did not examine every possible subset of the full list of items). Interestingly, the selected items range from those that are highly familiar (e.g., *olive*, *remember*) and those that are much lower in lexical frequency and familiarity (e.g., *denizen*, *molybdenum*).

Although the subsetting procedure resulted in fewer items that more highly correlate with the PPVT scores, we recommend administering the complete set of items and then only examining performance on the selected items. Our rationale is that performance may differ depending on whether other items are included in the task. In addition, the fully automated, self-administered tasks are relatively short in their administration, thus administering the whole set of items does not add a lot of time to an experiment.

Future Directions and Limitations

We note a few limitations of our analyses. First, this study was conducted on a relatively small number of participants and thus any results should be replicated in a larger sample. Relatively, only participants whose first language was U.S. English were included in the study and only adults were participated. Second, we only considered 100,000 resamples of the full-length task. While this may seem like a high number, it is only a small fraction of the possible number of subsets we could have considered. Therefore, our results should be taken as evidence that it is possible to substantially increase the correlations with the PPVT, noting that there may be a subset of items not explored in our analyses that achieves an even higher correlation than we found. Third, we used Pearson correlation in this study, which is only sensitive to linear relationships among the variables. Lastly, we did not explore other psychometric properties of any of the scales, including sensitivity-to-change or test–retest reliability of the PPVT or other tasks. Future work may explore how these tasks best distinguish between groups of respondents, or within a respondent across time.

Conclusions

We find these results promising. Although the different tasks may not use the same vocabulary items, the fact that there are numerous automated, self-administered measures of vocabulary that correlate highly with known, clinical measures of vocabulary suggests that researchers can save time on the part of both the research team and the participant by using these shorter, fully automated tasks of receptive vocabulary. Given that our subsetting analysis was able to generate strong correlations with the PPVT, researchers and clinicians can choose from among these other measures. The Shipley task, which requires participants to find the synonym, is the shortest to administer because there is less reading involved. If the researcher is concerned that the participant may not understand the concept of synonym, the VST offers an alternative with definition-matching. Finally, if the researcher is concerned about reading ability, the auditory LexTALE offers an additional alternative.

Data Availability Statement

The data sets used during this study are available from the corresponding author upon request.

Acknowledgments

This work was partially supported by the following grants: NSF BCS and SBP 2020805 (awarded to S.V.L., M.B.B., and D.H.), a seed grant from the Institute of Human Development and Social Change at New York University (NYU; awarded to S.V.L. and D.H.), a University Research Challenge Fund at NYU (awarded to S.V.L. and D.H.), and National Institutes of Health (NIH) R21DC016141 (awarded to R.M.T.). Its contents are solely the responsibility of the authors and do not represent the official views of the NIH, National Science Foundation, NYU, University of Oregon, University of Connecticut, or The University of Chicago.

References

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine, 18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Anwyl-Irvine, A. L., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods, 53*(4), 1407–1425. https://doi.org/10.3758/s13428-020-01501-5

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*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Babel, M. (2020). *Modified auditory LexTALE for native English listeners*. OSF.

Brownell, R. (2010). *Receptive and Expressive One-Word Picture Vocabulary Tests–Fourth Edition*. NCS Pearson.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE, 18*(3). https://doi.org/10.1371/journal.pone.0279720

Drown, L., Giovannone, N., Pisoni, D. B., & Theodore, R. M. (2023a). Validation of two measures for assessing English vocabulary knowledge on web-based testing platforms: Brief assessments. *Linguistics Vanguard, 9*(1), 99–11. https://doi.org/10.1515/lingvan-2022-0116

Drown, L., Giovannone, N., Pisoni, D. B., & Theodore, R. M. (2023b). Validation of two measures for assessing English vocabulary knowledge on web-based testing platforms: Long-form assessments. *Linguistics Vanguard, 9*(1), 113–124. https://doi.org/10.1515/lingvan-2022-0115

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test–Fourth Edition (PPVT-4)*. NCS Pearson.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*(1), 143–149. https://doi.org/10.3758/BF03203267

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur–Bates Communicative Development Inventories–Second Edition*. Brookes.

Freeman, V., & De Decker, P. (2021). Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps. *The Journal of the Acoustical Society of America, 149*(2), 1211–1223. https://doi.org/10.1121/10.0003529

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition, 23*(1), 83–94. https://doi.org/10.3758/BF03210559

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics, 27*(4), 513–543. https://doi.org/10.1017/S0142716406060383

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language, 28*(2), 200–213. https://doi.org/10.1016/0749-596X(89)90044-2

Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology, 28*(5), 887–898. https://doi.org/10.1037/0012-1649.28.5.887

Harel, D., & Baron, M., on behalf of the CSRG Investigators. (2018). Methods for shortening patient-reported outcome measures. *Statistical Methods in Medical Research, 28*(10–11), 2992–3011. https://doi.org/10.1177/0962280218795187

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

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), 316–330. https://doi.org/10.1037/0096-3445.122.3.316

McIntire, S. A., & Miller, L. A. (2007). *Foundations of psychological testing: A practical approach* (2nd ed.). SAGE.

Melby-Lervåg, M., Lervåg, A., Lyster, S.-A. H., Klem, M., Hagtvet, B., & Hulme, C. (2012). Nonword-repetition ability does not appear to be a causal influence on children's vocabulary development. *Psychological Science, 23*(10), 1092–1098. http://www.jstor.org/stable/23355500

Metsala, J. L. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. *Journal of Educational Psychology, 91*(1), 3–19. https://doi.org/10.1037/0022-0663.91.1.3

Mislevy, J. L., & Rupp, A. A. (2010). Concurrent validity. In N. J. Salkind (Ed.), *Encyclopedia of research design*. SAGE. https://doi.org/10.4135/9781412961288

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17,* 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

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

Prolific. (2014). *Prolific*. https://www.prolific.co

Sanchez, V. A., Arnold, M. L., Moore, D. R., Clavier, O., & Abrams, H. B. (2022). Speech-in-noise testing: Innovative applications for pediatric patients, underrepresented populations, fitness for duty, clinical trials, and remote services. *The Journal of the Acoustical Society of America, 152*(4), 2336–2356. https://doi.org/10.1121/10.0014418

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4)*. The Psychological Corporation/A Harcourt Assessment Company.

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology, 9*(2), 371–377. https://doi.org/10.1080/00223980.1940.9917704

Tomoschuk, B., Ferreira, V. S., & Gollan, T. H. (2019). When a seven is not a seven: Self-ratings of bilingual language proficiency differ between and within language populations. *Bilingualism: Language and Cognition, 22*(3), 516–536. https://doi.org/10.1017/S1366728918000421

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition, 19*(1), 122–140. https://doi.org/10.1017/S1366728914000832

Wagner, R., Torgesen, J. K., & Rashotte, C. (1999). *Comprehensive Test of Phonological Processing (CTOPP)*. Pro-Ed.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Zhang, C., Jepson, K., Lohfink, G., & Arvaniti, A. (2021). Comparing acoustic analyses of speech data collected remotely. *The Journal of the Acoustical Society of America, 149*(6), 3910–3916. https://doi.org/10.1121/10.0005132

Appendix A

Public links to vocabulary tasks:

VST:
https://app.gorilla.sc/openmaterials/245615
WordFam:
https://app.gorilla.sc/openmaterials/245615
LexTALE:
https://osf.io/a7ftx/

Appendix B

Items Selected After Subset Resampling

Shortened Shipley:

orifice, fortify, fascinate, denizen, serrated, hilarity, querulous, hideous, massive, divest, remember

Shortened VST:

olive, cadenza, augur, aver, malign, pussyfoot, vial, dinosaur, didactic, canonical, compound, strap, monologue, crab, bidet, copra, soliloquy, dig

Shortened WordFam:

molybdenum, immobility, mastodon, impair, cacophony, educate

Shortened LexTALE:

scornful, cairn, celestial, remuda, kilp, yonker, magrity, slain, carbohydrate, crumper, vicissitude, majestic, listless, alberation, cupidity