2016

# Determining the Validity of a Web-Based, Self-Rating Checklist Assessment of Vocabulary Knowledge

Sheida Abdi
*University of South Carolina*

# Determining the Validity of a Web-Based, Self-Rating Checklist Assessment of Vocabulary Knowledge

by

Sheida Abdi

Bachelor of Science
University of South Carolina, 2014

---

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Speech Pathology in

Speech Pathology

The Norman J. Arnold School of Public Health

University of South Carolina

2016

Accepted by:

Suzanne Adlof, Director of Thesis

Jessica Richardson, Reader

Lesly Wade-Woolley, Reader

Paul Allen Miller, Vice Provost and Interim Dean of Graduate Studies

DEDICATION

To my family for their endless love and support, David for his jokes and board games, my roommates for lending advice and snacks, Trader Joe's for providing more snacks, my roommate's pets for the stress relief, and my thesis-partner-in-crime for understanding this process.

ACKNOWLEDGEMENTS

Thank you to my committee members for selflessly dedicating time to read my thesis, and my thesis director for three years of education, encouragement, and guidance.

ABSTRACT

The purpose of this study is to examine the validity of a web-based, self-rating checklist of vocabulary knowledge. One hundred fifty-nine participants took two norm-referenced assessments in addition to one of three conditions of a developed self-rating checklist. Each condition employed a different combination of follow-up questions (synonym generation to verify participants' self-ratings) and feedback for student responses (whether or not synonyms are correct). Condition 1 did not provide any follow-up questions or feedback, Condition 2 included follow-up questions and feedback, and Condition 3 presented follow-up questions but no feedback. Results show that participants moderately overestimated vocabulary knowledge. Moderate-to-high statistically significant correlations (0.51 – 0.71) were observed between each condition of the assessment and norm-referenced assessments. Additionally, multiple regression analyses indicated that 31-67% of variance in norm-referenced assessments could be explained by scores on the self-rating checklist assessment, demonstrating concurrent validity with norm-referenced vocabulary tests. Results indicated few differences in the prediction of norm-referenced assessments between conditions differing in follow-up questions/feedback. However, participant responses to post-assessment surveys show that the presence or lack of feedback and follow-up questions had a slight effect on their perceptions of construct validity. These results demonstrate both construct and concurrent validity and suggest that a self-rating checklist can be a valid assessment of vocabulary.

*Keywords*: vocabulary, assessment, validity

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

Vocabulary knowledge is one of the strongest predictors of reading comprehension and general academic achievement in adolescents and young adults (Adlof & Perfetti, 2013). The more children read, the greater their vocabulary size (Cunningham & Stanovich, 1998). Even when the number and quality of word exposures is controlled (whether through listening or reading), evidence suggests that children vary in their word learning abilities (Kan & Windsor, 2010; Warmington, Hitch, & Gathercole, 2013). Additionally, there are children with developmental language disorders who don't have the requisite vocabulary knowledge to succeed regardless of exposure. This variation in vocabulary knowledge and word learning abilities suggests that some children would benefit from greater vocabulary instruction in schools.

Word selection is the first step in vocabulary instruction. Several approaches exist for selecting words to be taught in classrooms or during group instruction. For example, Beck and McKeown (1985) divided words into three tiers, with Tier 2 consisting of words that are high frequency for educated adults, appear in a variety of contexts, and have high utility. They suggest Tier 2 words are the most useful for instruction. Biemiller (2010) also categorizes words into three tiers: easy, high priority, and difficult. High priority words are known by 40-79% of children at the end of second grade; Biemiller suggests these words are optimal targets for vocabulary instruction for children who trail behind their peers. Hiebert's word families approach focuses on the 4,000 most common

word families (2005). Because words that share roots are inherently semantically related, Hiebert recommends teaching them in groups to efficiently increase students' word knowledge and facilitate students' abilities to infer word meanings. In addition to the considerations presented by these approaches, the selection of words for instruction should also take into account words that an individual already knows. In order to factor in an individual's vocabulary knowledge, an assessment is required.

Traditional measures of vocabulary knowledge have several limitations with regard to informing vocabulary instruction. First, norm-referenced assessments compare students to their peers in terms of total vocabulary size rather than guiding teachers on what words should be taught. These assessments also feature a fixed item set in which there is little flexibility in the words tested. Moreover, in order to preserve the validity of a norm-referenced test, instructors should not teach words that appear as items on such tests. Second, individually administered paper-and-pencil vocabulary assessments may require considerable time to administer, score, and interpret.

Third, most assessments test word knowledge in a binary fashion, with one particular "level" of knowledge tested across words, *e.g.*, surface-level knowledge of word meaning, or deep knowledge of specific word meanings. Such tests may over- or underestimate the student's true knowledge of a word. An individual's knowledge of a vocabulary word is not binary. Rather knowledge of a word develops incrementally, increasing with more exposures to a word (Nagy & Scott, 2000). Researchers have often characterized knowledge of a word as developing on a continuum, from unrecognized to completely known (Beck, McKeown, & Omanson, 1987; Christ, 2011; Dale, 1965; Miller, 1999). For example, Dale (1965) described four levels of knowledge: 1) the

individual has never seen the word before, 2) the individual has seen the word but does not know the meaning, 3) the individual recognizes some information about what contexts the word can be found in, and 4) the individual knows the word well.

In this study, we evaluated the validity of a web-based, self-assessment checklist for assessing vocabulary knowledge. A self-assessment checklist can be a useful format for a vocabulary assessment because it addresses the previously noted limitations. First, the assessment can be used by teachers to prescribe words to be studied because the assessment can directly show instructors what words children know and what words they do not know. The assessment can easily be modified based on the needs of teachers and students. Students can complete the assessment quickly, including time required to make decisions about words. Finally, the assessment can also take into account partial word knowledge if students are asked to rate their knowledge on a scale.

Two studies provide preliminary evidence of the validity of self-assessments of vocabulary. Durso and Shore (1991) investigated partial word knowledge in several experiments involving sentence decision tasks. First, students' levels of knowledge of a set of target words were measured using a self-assessment. Students made four consecutive passes through the word list, with the first pass identifying words they knew well enough to list a synonym, the second pass identifying words they knew well enough to use in a good sentence, the third pass identifying words that seemed familiar, and the last pass identifying items that looked like nonwords to the student. Next, participants made decisions about sentences containing or involving the target word. Results from each experiment showed that decision accuracy was related to participants' reported levels of word knowledge: Participants were more accurate when they knew more about

word meaning, demonstrating that self-assessed levels of word knowledge were valid indicators of true knowledge. Interestingly, participants scored above chance for choosing general contexts that were appropriate for words they claimed were *not* real words; however, they were able to answer more specific questions when they had higher levels of word knowledge. Overall, the results provided evidence of partial word knowledge and construct validity of self-assessments by demonstrating that participants' degrees of word knowledge influenced their ability to make accurate decisions about word usage.

Further evidence of the validity of checklist assessments is provided by Ackerman and Ellingsen (2014), who investigated vocabulary "overclaiming" among college students. Students completed tests of verbal ability, a checklist assessment of vocabulary knowledge, and an objective measure of vocabulary knowledge (a definition generation task using the same words in the checklist assessment). Results showed that many college students claimed to be able to define more vocabulary words on the checklist measure than they were actually able to define. However, despite this overclaiming, a strong correlation existed between self-claimed and objectively determined vocabulary knowledge. Further analysis indicated that the higher ability students were more likely to overclaim their vocabulary knowledge. This finding may be due to the increased difficulty level of the objectively determined vocabulary measure, a definition generation task. Students may have had a hard time retrieving or formulating full definitions for words they could understand, and therefore left answers blank or provided partially correct definitions. One question is whether students would be less likely to overclaim in an alternative assessment format.

The purpose of this study is to examine the validity and utility of a self-rating checklist for assessing vocabulary knowledge. This study takes place within the context of the development of a web-based vocabulary instructional program, called DictionarySquared. The DictionarySquared program is designed to provide individualized teaching of word meanings using dictionary definitions, real word contexts, and activities to promote active processing of semantics. Because of the web-based platform, students are able to access it from virtually anywhere. The checklist assessment was developed in order to prescribe words to teach within the DictionarySquared platform. Our goal was to develop a brief assessment that can be used to select words for instruction, can easily be modified, and that can also measure partial word knowledge.

We examined two types of validity: construct and concurrent. Construct validity refers to how well an assessment measures what it purports to measure. This validity can be demonstrated if students' reported knowledge of word meanings reflects their objectively measured knowledge. Concurrent validity is a measure of how an assessment compares to a previously established assessment that measures the same construct. Evidence of concurrent validity can be observed based on how well scores on our self-rating assessment of vocabulary predict scores on norm-referenced vocabulary assessments. Initial evidence of the construct validity of self-assessments of partial word knowledge was demonstrated by Durso and Shore (1991) when students' ability to accurately judge sentences containing target words was dependent on their self-assessed level of knowledge of that word. However, that study did not examine concurrent validity. Ackerman and Ellingsen (2014) also provided evidence of construct validity of

self-rating checklists, as student's self-claimed level of word knowledge was highly correlated with their objectively determined level of word knowledge. Additionally, they provided preliminary evidence of concurrent validity by demonstrating strong correlations among results on assessments of verbal ability, the self-assessment, and the objectively determined vocabulary assessment. However, they did not observe correlations between their checklist or their objective measure and assessments that specifically relate to vocabulary size, rather than general verbal ability. Furthermore, their self-rating checklist, based on Kirpatrick (1905) and Whipple (1908), asked participants to rate high levels of word knowledge, i.e., whether or not they were able to define a word. To our knowledge, no study has examined the concurrent validity of a vocabulary self-assessment that considers partial word knowledge.

In this study, we examined the construct and concurrent validity of a self-rating checklist assessment that considered partial word knowledge. Students completed one of three conditions of the developed assessment and two norm-referenced vocabulary tests. We evaluated student responses on the self-rating checklist assessment and correlations between the checklist assessment and norm-referenced assessments. Additionally, we examined the effects of three feedback methods on student overclaiming. Whereas Ackerman and Ellingsen (2014) measured overclaiming based on high levels of knowledge (being able to generate a definition of a word) we expanded the investigation of overclaiming to investigate all possible levels of knowledge by including pseudoword foils in our checklist. Our first research question addressed construct validity by examining the extent to which students accurately reported their knowledge of specific words. This was examined in two ways: (a) by including pseudo-word foils within all

conditions of the checklist to detect guessing, and (b) by including follow-up questions in two conditions of the checklist. Strong evidence of construct validity would be provided if the rate of guessing was relatively low and if students generally showed accurate responses to follow-up questions. From a different perspective, the extent to which students reported knowledge of pseudowords or were unable to accurately respond to follow-up questions would provide estimates of overclaiming.

Our second research question was: how well do results on the self-rating checklist assessment predict results on existing norm-referenced measures of vocabulary? Strong positive correlations between results on the self-rating checklist and norm-referenced assessments would provide evidence of concurrent and construct validity. Additionally, linear regression models assessed whether including information about guessing behavior improved the prediction of scores on norm-referenced assessments by explaining unique variance beyond that accounted for by self-reported knowledge of real words.

Our third research question investigated whether feedback methods influenced students' response patterns. Specifically, we wanted to assess whether the relationship between students' estimates of their own vocabulary knowledge and norm-referenced assessments varied according to whether they received follow-up questions and/or feedback. To examine this question, we compared the results of linear regression models between the three conditions.

Finally, our fourth research question examined the extent to which students perceived the checklist as having construct validity, based on qualitative data from a post-test survey. Descriptive statistics from the survey were analyzed for all participants and also compared by group.

# CHAPTER 2

# METHOD

## 2.1 PARTICIPANTS

A total of 192 participants, who were primary English language speakers between the ages of 18 and 25 years, participated in the iterative development and testing process. The first 17 participants completed pilot testing as the checklist was initially developed and revised; thus, their data is not presented here. In addition, a temporary bug in the assessment system yielded unusable data for 14 participants. Furthermore, two more participants were excluded when it was determined that they were not primary English speakers. The reported analyses involve 159 individuals who completed one of the three conditions of the checklist assessment and both norm-referenced assessments. The participants were undergraduate students from the University of South Carolina or members of the surrounding community. Participants were recruited via advertisements on campus and in the community. They received course credit or $10 for their participation.

## 2.2 CHECKLIST ASSESSMENT WORDS

The DictionarySquared ($D^2$) platform contains a core list of 1000 vocabulary words intended to span the full difficulty range of words a high school student at any level of ability may need to learn. The words are divided into 10 bands, 100 words each, of increasing difficulty. Word difficulty is estimated by frequency and age of acquisition norms (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Zeno, Ivens, Millard, &

Duvvuri, 1995). The checklist presents nine randomly selected real words from each band to each student. Thus, the words within the self-assessment checklist varied between participants.

## 2.3 CHECKLIST ASSESSMENT PSEUDOWORD FOILS

Pseudoword foils were developed using the Wuggy program (Kuleers & Brysbaert, 2010) to detect random guessing within the checklist. All participants were presented the same three pseudoword foils for each band (total of 30 pseudowords for all 10 bands). The pseudowords were created to match the mean length and orthotactic probability of the real words within each band.

## 2.4 CHECKLIST ASSESSMENT INSTRUCTIONS

Participants were presented with nine target words and three pseudowords, randomly ordered, for each band of the checklist. They were instructed to rate their level of knowledge following a scale that was adapted from Durso and Shore (1991). The scale included the following options: (a) high knowledge: the students clearly understands the word and can explain its meaning to someone else; (b) partial knowledge: the student understands the word's meaning in a sentence but cannot provide a definition out of context; (c) recognized: the student identifies the word is real, but does not know anything about the meaning; (d) unknown: the student has never seen the word or believes the word may be made-up. Participants were notified that the checklist included made up words to detect random guessing. Figures 2.1-2.3 show sample pages from Bands 1, 5, and 10 of the checklist.

## 2.5 CONDITIONS

Participants were quasi-randomly assigned to one of three conditions of the checklist vocabulary assessment.

Fifty-five participants completed Condition 1. In this condition, students simply completed the checklist as described thus far (see Figures 2.1-2.3).

Fifty-one participants completed Condition 2. In this condition, they received a follow-up question with feedback for some of the words they rated as highly known. After participants submitted their checklist answers for a given band, a pop-up presented follow-up questions for a random 25-40% of the words that they had rated as highly known within that band. In the pop-up, participants were asked to provide a single word synonym for the target word. The program compared synonyms to the list of possible synonyms provided in Merriam-Webster's Online Dictionary. When a participant's response did not match an entry in that dictionary, the participant received the following feedback, "Are you sure about the synonyms highlighted in red? You may change your answer if you like. If you are satisfied with the answer provided, press 'Continue.'" The feedback was worded in this way in order to avoid providing inaccurate feedback. That is, it was possible that participants may have provided a word that was related to the target word but not included in the Merriam Webster list. We aimed to not discourage participants by providing inaccurate feedback.

Fifty-three participants completed Condition 3. In this condition, participants received the same synonym generation follow-up questions described in Condition 2, but they did not receive any feedback regarding the accuracy of their responses (*i.e.*, "Are you sure…").

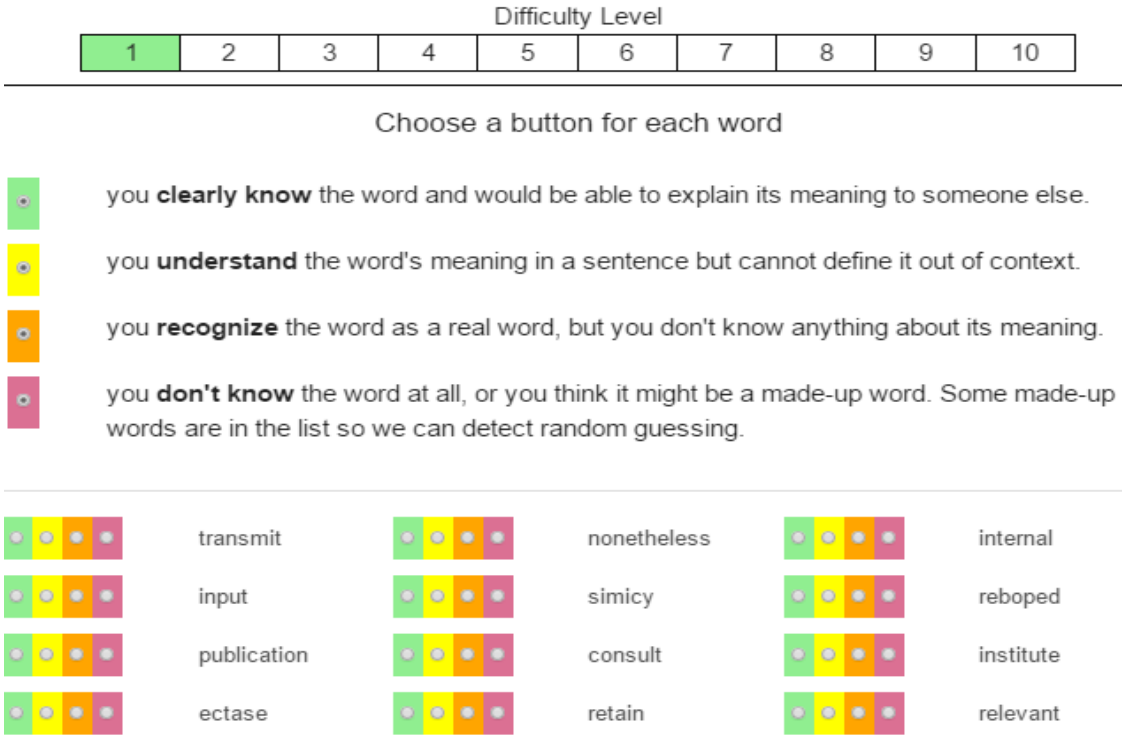



Figure 2.1 Band 1 of the checklist assessment

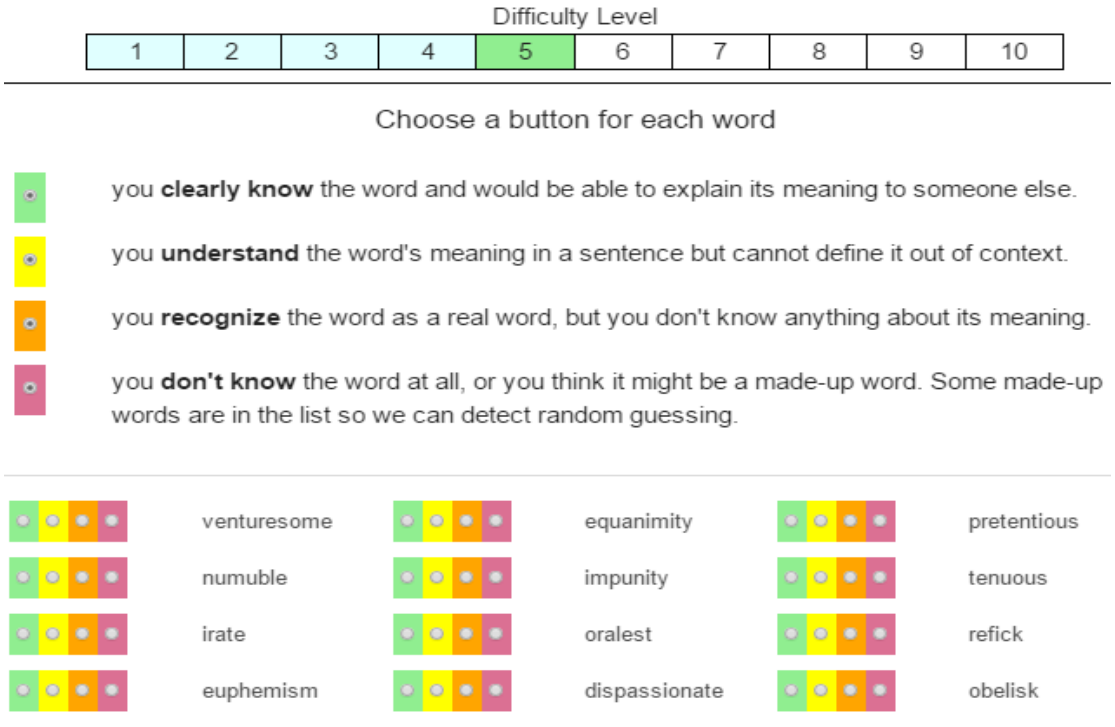



Figure 2.2 Band 5 of the checklist assessment


Figure 2.3 Band 10 of the checklist assessment

## 2.6 CHECKLIST SCORING

Four scores were derived from each of the three conditions of the self-assessment checklist. The Total Known (TK) score is the total number of real words a student rated as anything other than unknown. The maximum TK score was 90, as there were 9 real words in each of the 10 bands. The Partial Word Knowledge (PWK) score is the sum of word knowledge scores for all real words on the test. Scores for each word range from 0 for unknown words, to 3 for highly known words. The maximum possible PWK score was 270. The Total Guessed (TG) score is equal to the total number of pseudoword foils a student marked as anything other than "unknown." In other words, if a student rated a pseudoword any level of known (recognized, partially known, or highly known), it was

concluded that the student guessed about whether the word was real or not. With 30 pseudowords, the maximum score was 30. The Guessed Level (GL) score was akin to the PWK score but involved pseudowords; it represented the sum of scores for each of the 30 pseudowords (max = 90). The TG and GL scores were included to provide an estimate of the degree to which students may have adopted a strategy of overclaiming.

In addition, students' responses to synonym generation items were hand scored to assess the validity of self-ratings in Conditions 2 and 3 (note that there were no synonym generation items in Condition 1). Student responses in Conditions 2 and 3 were hand scored to give credit to correct answers that may have not been matched to synonyms listed in Merriam Webster's website. Student responses that were not found in Merriam Webster but included in Thesaurus.com were considered correct. Fifteen percent (15%) of all correct responses were not initially considered correct by the automated scoring procedure that relied only on the Merriam Webster list. Proportions of correct and incorrect synonyms were evaluated to observe whether people demonstrated knowledge of words they claimed to know.

## 2.7 NORM-REFERENCED ASSESSMENTS

In addition to the checklist self-assessment, participants completed two norm-referenced vocabulary assessments, the Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4; Dunn & Dunn, 2007) and the Gates-MacGinitie Reading Test, Fourth Edition, Level AR (Adult Reading) vocabulary subtest (GMRT-4; MacGinitie, MacGinitie, Maria & Dreyer, 2002), which were administered in a counterbalanced order along with condition A, B, or C of the checklist assessment. The PPVT-4 is a norm-referenced measure of receptive vocabulary that is commonly used to assess surface level knowledge

of up to 228 sample words, which range in frequency from very common to very rare. In the test, a clinician says a word, and the student is asked to select, from a field of four, the picture that best represents the word (*e.g.*, the word "eating" may feature four pictures: a child eating, a woman writing, a man cooking, and a child putting away dishes). The PPVT-4 manual provides norms for individuals ages 2;6-90+ and reports good to excellent reliability statistics, with split-half reliability ranging from .89-.97 across age groups, alternate form reliability ranging from .87-.93, and test-retest reliability ranging from .92-.96.

The GMRT-4 vocabulary subtest assesses a deeper level of vocabulary knowledge for 45 words. The test presents brief contexts with an underlined target word. Individuals taking the test are asked to select the best replacement word or phrase for the underlined word out of a field of five that best preserves the original context. The contexts provide little information about the target word other than part of speech. The incorrect answer choices are selected based on a variety of criteria including shared semantic features with the correct answer or words that appear to be a correct answer based on incorrect reading of the original context. The Adult Reading norms are based on individuals in their first year of community college and demonstrate good reliability statistics (internal consistency =.88)

### 2.8 EXIT SURVEY

After completing all three assessments, participants completed an exit survey with the following open-ended questions:

1. What did you think of the assessment, overall?

2. Do you think the assessment adequately measured your vocabulary knowledge? Why or why not?
3. If you could select one thing, what did you like best about the assessment?
4. If you could make any improvements to the web assessment, what would you do?
5. What comments do you have about the other assessments?

# CHAPTER 3

# RESULTS

Prior to parametric analyses, the distribution of data for each measure was inspected. Although our target population included adults with typical vocabulary knowledge, three individuals displayed norm-referenced vocabulary scores that appeared to be outside of the range of normal (> 1 SD below the mean). These individuals were significant outliers relative to their group, with scores greater than 1.5 interquartile ranges below the 25$^{th}$ percentile for their group. Therefore, they were excluded from further analysis.

Descriptive statistics reflecting participant performance on the norm-referenced assessments (PPVT and GMRT) and the checklist assessment are provided in Table 3.1. Descriptive statistics for the accuracy of responses to follow up questions in Conditions 2 and 3 are provided in Table 3.2. The first research question involved the construct validity of students' self-reported ratings of their vocabulary knowledge. GL means ranged from 10.39-11.46 out of a possible 90 in Conditions 1, 2, and 3, suggesting a low rate of guessing (11%-12%). Mean accuracy of follow-up question in Conditions 2 and 3 indicating moderate accuracy at 60% and 54%, respectively. Overall, the data on guessing and follow-up question accuracy demonstrate that participants had generally low-to-moderate rates of guessing. When participants did guess, they typically rated the lowest possible known rating, suggesting that that participants attempted to accurately

rate their knowledge. Thus, these results provide further supporting evidence of the construct validity of checklist assessments.

Because we were interested in examining potential differences related to the different conditions of the assessment, it was important to test whether the groups showed similar levels of vocabulary knowledge. One-way analysis of variance confirmed that the participants assigned to each condition of the checklist assessment did not differ in their vocabulary skills, as measured by the PPVT and the GMRT. However, there were significant group differences in the total number of words reported known on the checklist as well as the overall PWK score. Follow-up *t* tests indicated that participants in Condition 1 rated significantly more words known at significantly higher levels of knowledge than participants in Conditions 2 and 3 ($p < .05$), who did not differ from each other ($p > .23$). Note that the pattern of group means for PWK and TK of the checklist matches the pattern of means observed for norm-referenced assessments even though the norm-referenced assessment scores were not significantly different. Results indicated that groups did not differ in the total number of guesses or the level of guesses ($p > 0.83$). Thus, differences between checklist conditions did not appear to influence participants' guessing behavior. Finally, results indicated that follow up question accuracy was not statistically significantly different between Conditions 2 and 3 ($p = .08$).

Table 3.1 Descriptive Statistics for Participant Scores in All Conditions

| | Condition 1 Mean (SD) | Condition 2 Mean (SD) | Condition 3 Mean (SD) | F (2, 155) | *P* | $\eta_p^2$ |
|---|---|---|---|---|---|---|
| *Norm-Referenced Assessments* | | | | | | |
| PPVT-4 SS | 107.69 (13.22) | 105.04 (10.25) | 104.94 (9.76) | 1.019 | 0.363 | 0.013 |
| GMRT-4 ESS | 631.94 (36.43) | 623.24 (37.37) | 620.00 (30.40) | 1.668 | 0.192 | 0.021 |
| *Checklist Scores* | | | | | | |
| Total Known | 73.20 (14.17) | 67.46 (15.49) | 65.79 (14.43) | 3.724 | 0.026 | 0.046 |
| PWK Score | 170.26 (47.28) | 148.14 (41.61) | 138.14 (35.40) | 8.221 | < 0.001 | 0.097 |
| Total Guessed | 9.20 (7.40) | 9.36 (8.66) | 8.96 (7.32) | 0.034 | 0.967 | < 0.001 |
| Guessed Level | 11.46 (11.23) | 11.42 (11.53) | 10.39 (8.68) | 0.181 | 0.835 | 0.002 |

Table 3.2 Descriptive Statistics for Follow-Up Question Accuracy in Conditions 2 and 3

| Condition 2 Mean Accuracy (SD) | Condition 3 Mean Accuracy (SD) | *t* | *p* | *D* |
|---|---|---|---|---|
| 60.00 (18.36) | 53.96 (13.18) | 1.77 | 0.079 | 0.35 |

## 3.1 CORRELATIONAL ANALYSIS

The second research question asked how well checklist measures predict performance on norm-referenced measures. The third research question considered differences between conditions of the checklist assessment. Correlation and regression analyses addressed these questions. Scatter plots of associations between norm-referenced assessments and self-rating checklist scores indicated linear relationships in most cases (see figures 3.1-3.5, 3.7, 3.9, and 3.11). Pearson correlations for these variables are displayed in Table 3.3. Moderate to strong positive correlations were observed between the self-assessment checklist and the norm referenced assessments across most conditions, suggesting that individuals with higher scores on norm-referenced vocabulary assessments also rated their word knowledge higher on the checklist assessment. However, the association between GL scores and norm-referenced scores in Conditions 2 and 3 appeared to be potentially nonlinear (see figures 3.6, 3.8, 3.10, and 3.12), with the highest levels of guessing exhibited by people who scored near the mean in Conditions 2 and 3. This relationship was not observed with participants'

scores in Condition 1. Thus, in the regression models, we examined both linear and quadratic relationships between guessing and norm-referenced vocabulary scores.

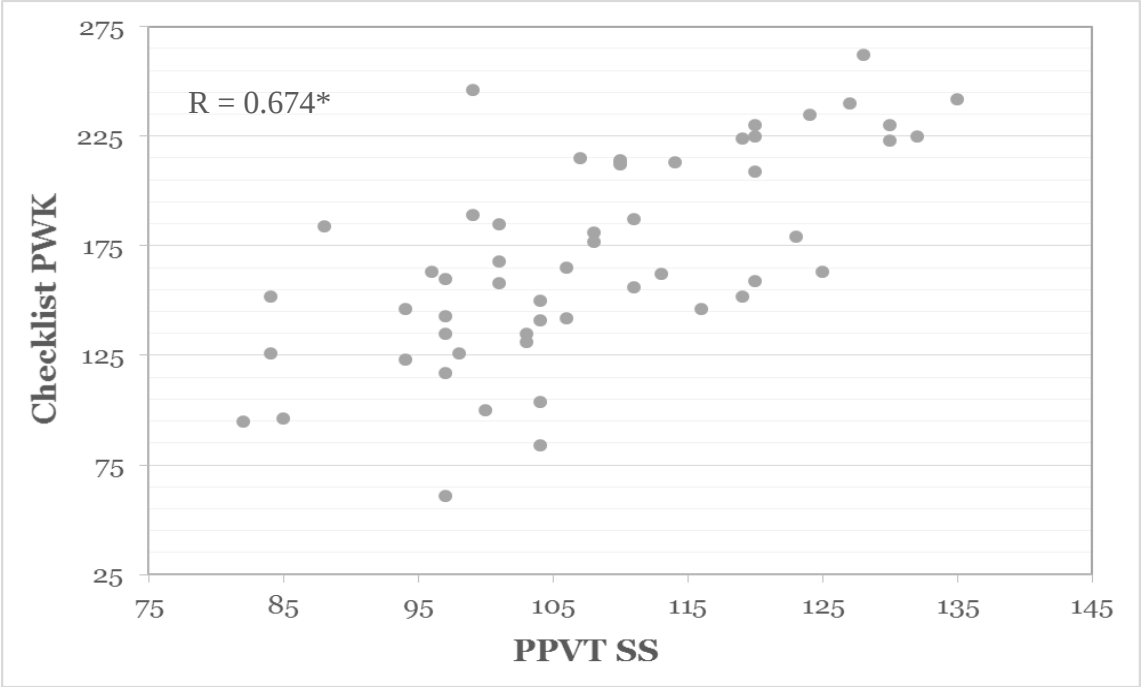
Table 3.3 Norm-Referenced Vocabulary Scores and Checklist Scores Correlations

| Condition | Assessment | TK | PWK | TG | GL |
|---|---|---|---|---|---|
| 1 | PPVT-4 SS | 0.498*** | 0.674*** | -0.191 | -0.152 |
| | GMRT-4 ESS | 0.527*** | 0.636*** | -0.263 | -0.262 |
| 2 | PPVT-4 SS | 0.434** | 0.516*** | -0.095 | -0.104 |
| | GMRT-4 ESS | 0.494*** | 0.601*** | -0.059 | -0.061 |
| 3 | PPVT-4 SS | 0.326* | 0.507*** | 0.132 | 0.132 |
| | GMRT-4 ESS | 0.516*** | 0.709*** | 0.157 | 0.172 |

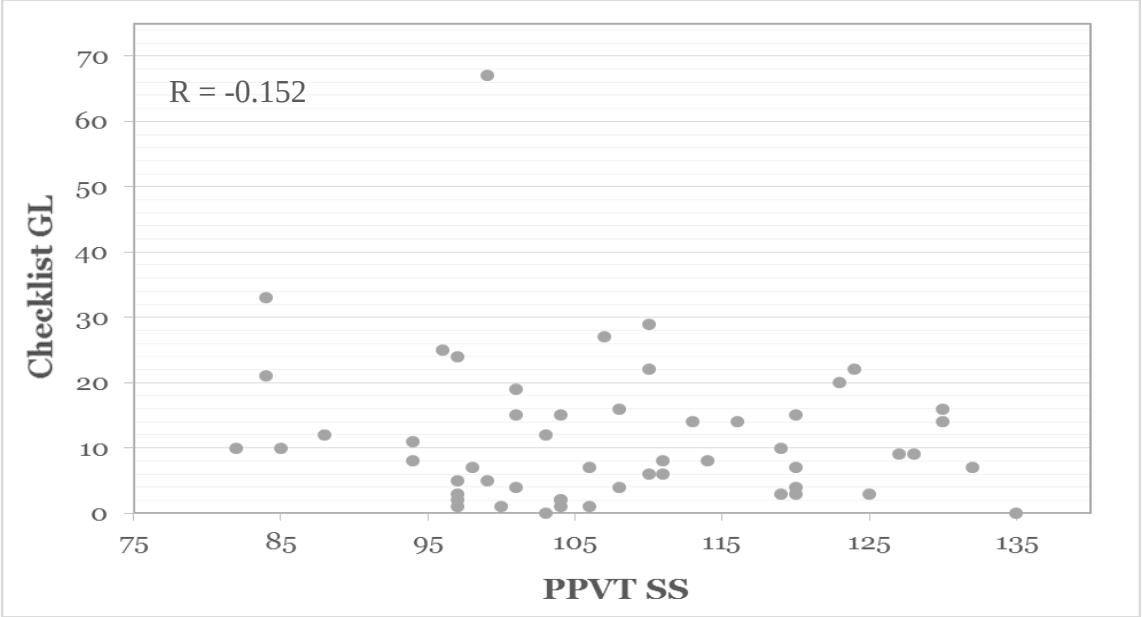
* $p < 0.05$ ** $p < 0.01$ ***$p < 0.001$

3.2 LINEAR REGRESSION

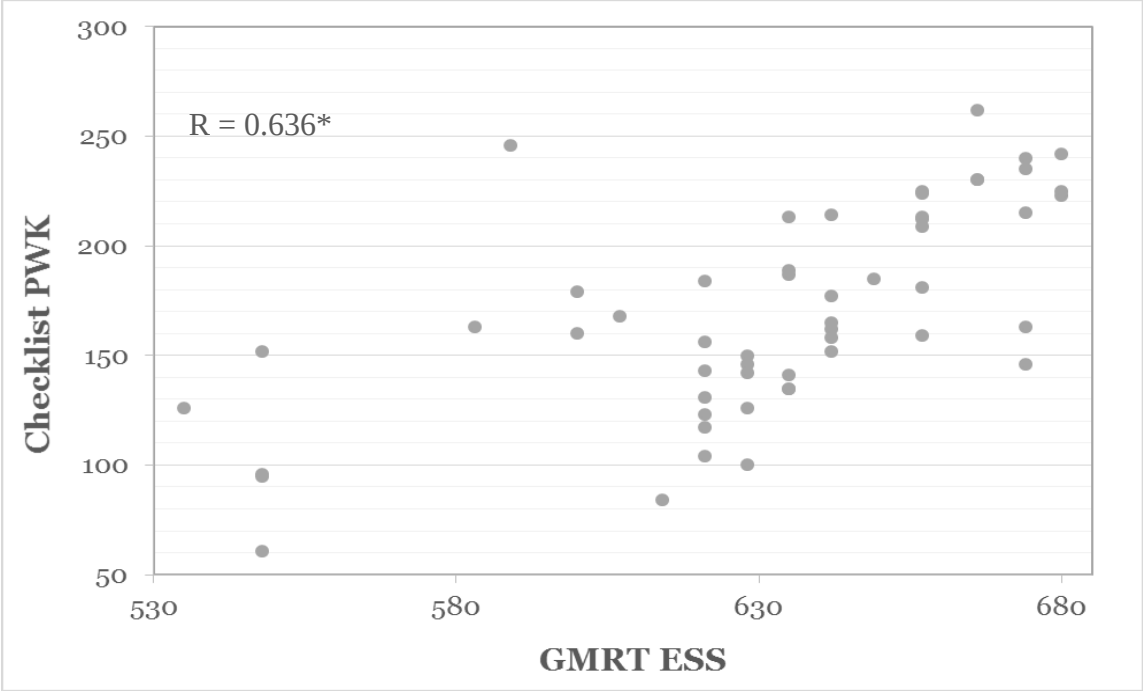
Multiple regression analyses were used to examine the degree to which scores on each condition of the checklist assessment predicted individual differences on the PPVT and GMRT (Tables 3.4 and 3.5). In the correlational analyses described previously, the TG and the GL scores were highly correlated, with $r > 0.9$ for each condition. Similarly, the TK and PWK scores were highly correlated, $r > 0.85$ for each condition. These high correlations were expected, as the TK/TG scores were derived from the PWK/GL scores. To avoid excessively collinear variables, we only included PWK and GL scores in the regression models. For each outcome variable, we entered predictors in three sequential steps: PWK, GL, and the square of GL (to test the quadratic function).



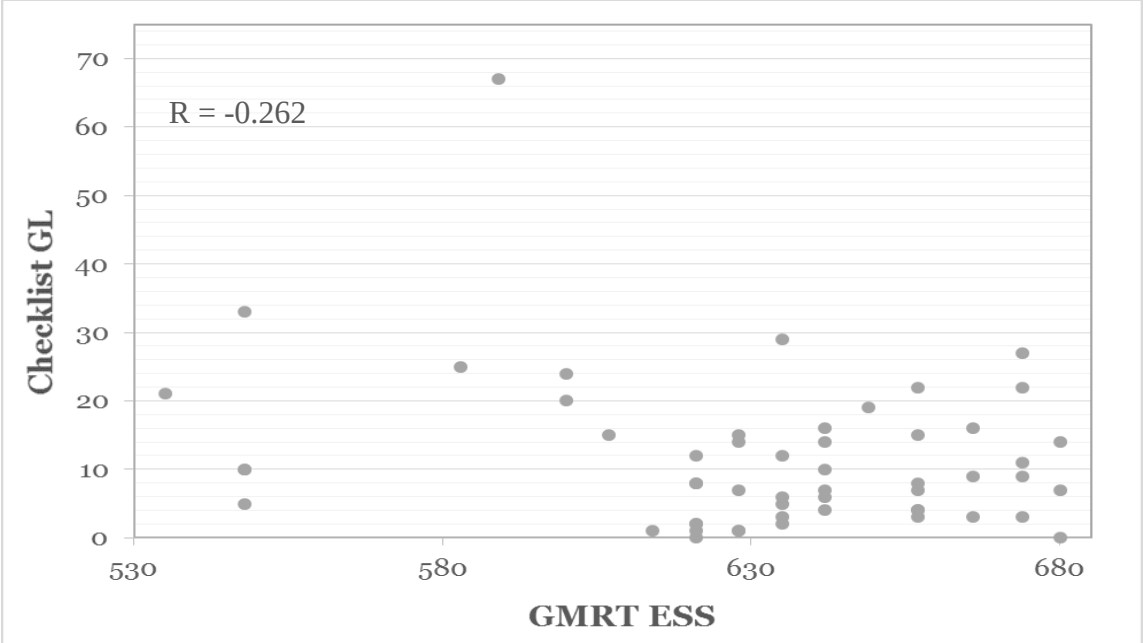

Figures 3.1 Condition 1 relationship between PWK scores and PPVT results.



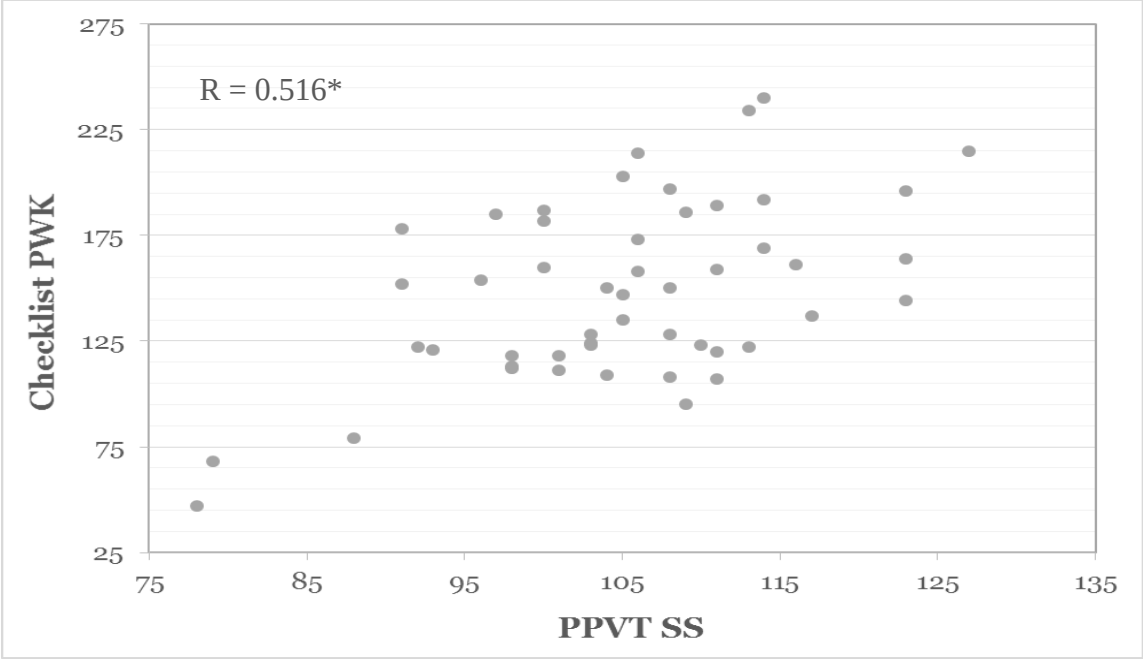

Figures 3.2 Condition 1 relationship between GL scores and PPVT results.



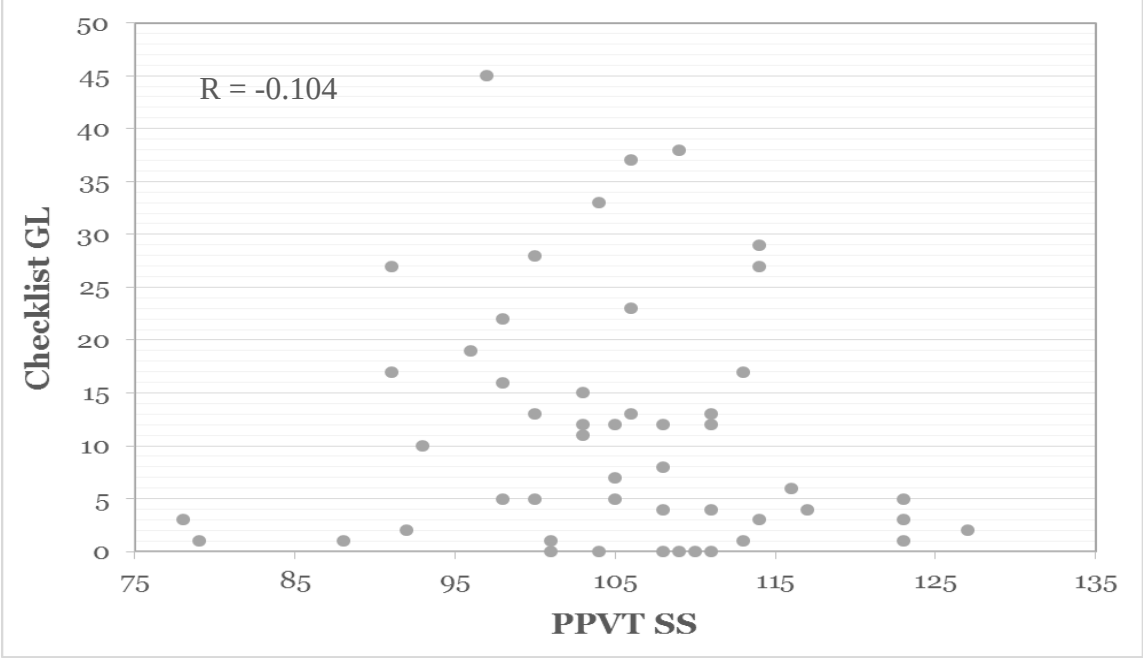

Figures 3.3 Condition 1 relationship between PWK scores and GMRT results.



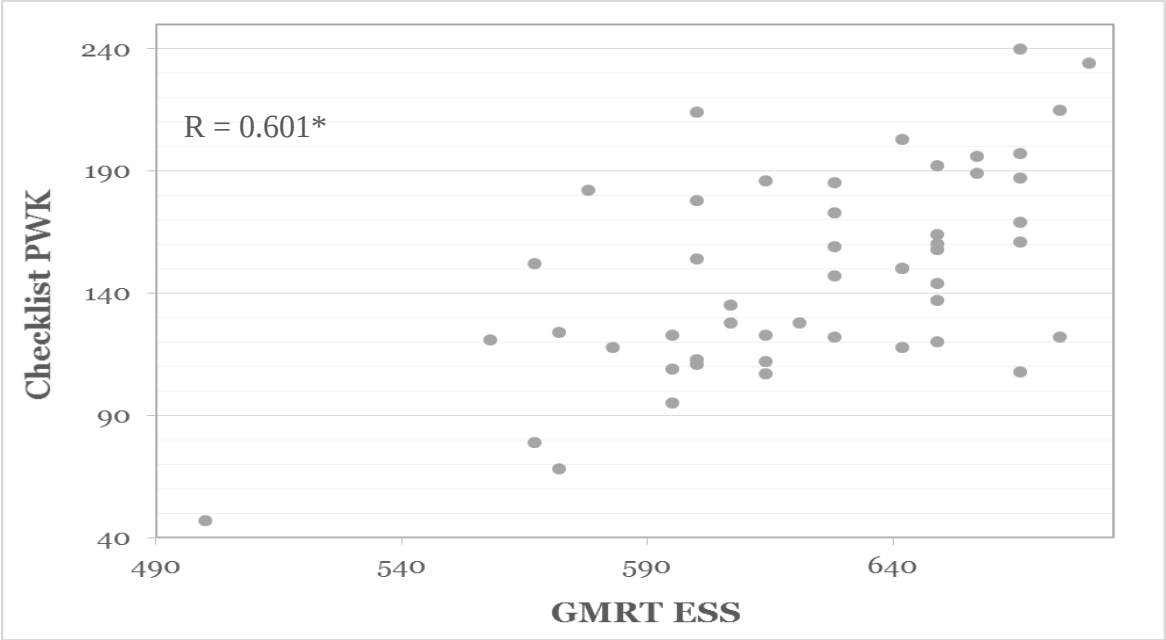

Figures 3.4 Condition 1 relationship between GL scores and GMRT results.



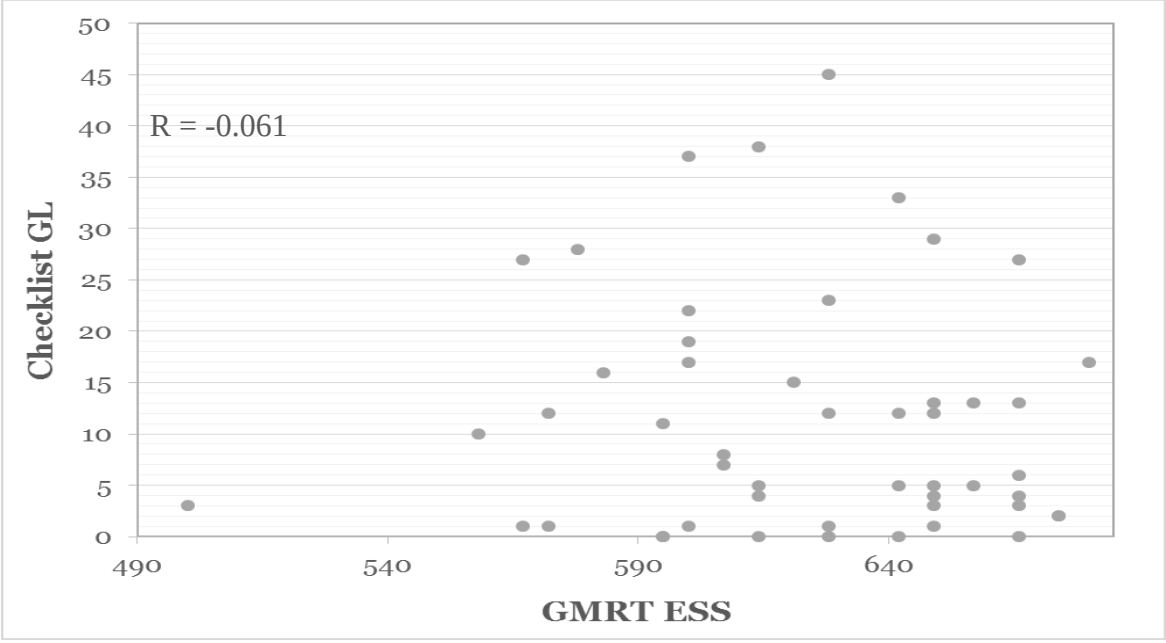

Figures 3.5 Condition 2 relationship between PWK scores and PPVT results.



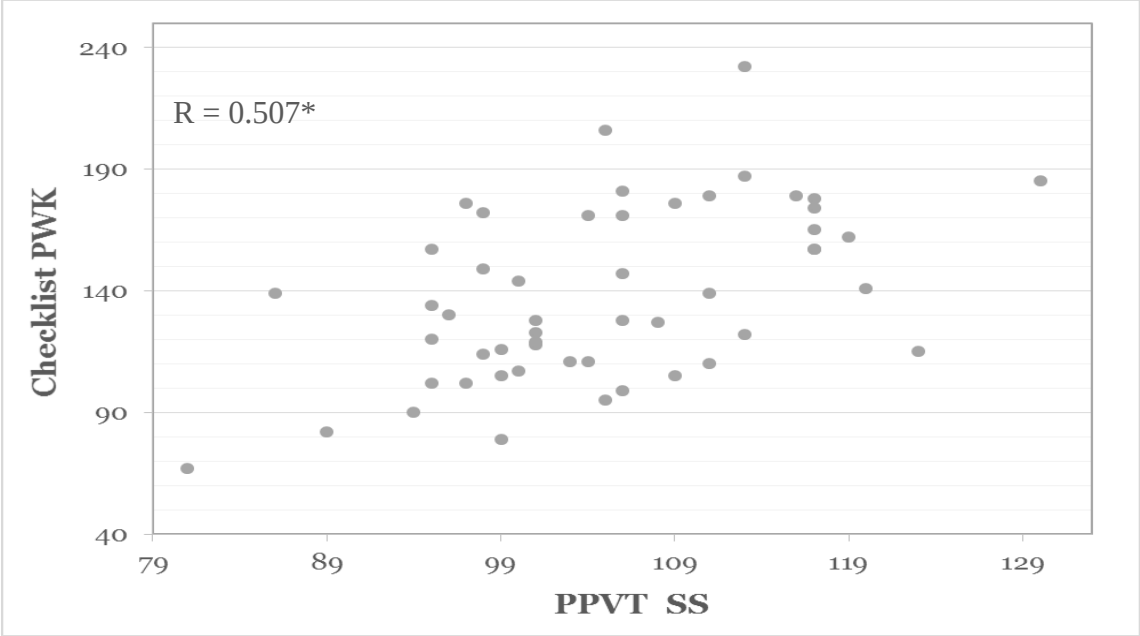

Figures 3.6 Condition 2 relationship between GL scores and PPVT results.



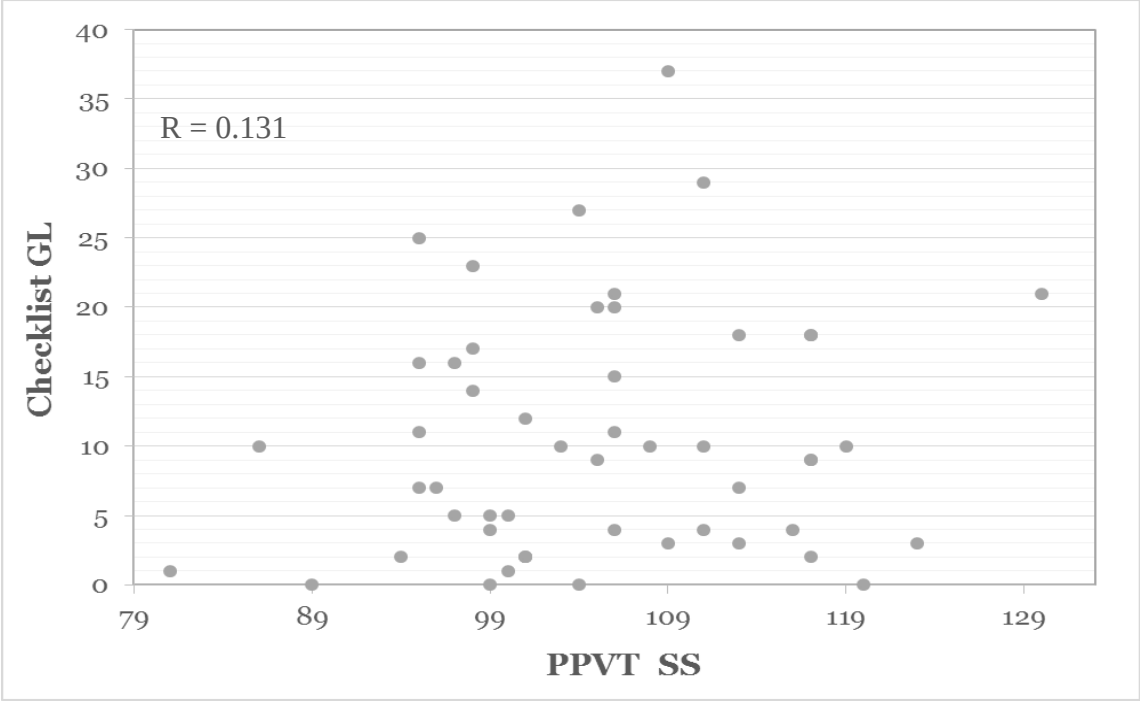

Figures 3.7 Condition 2 relationship between PWK scores and GMRT results.



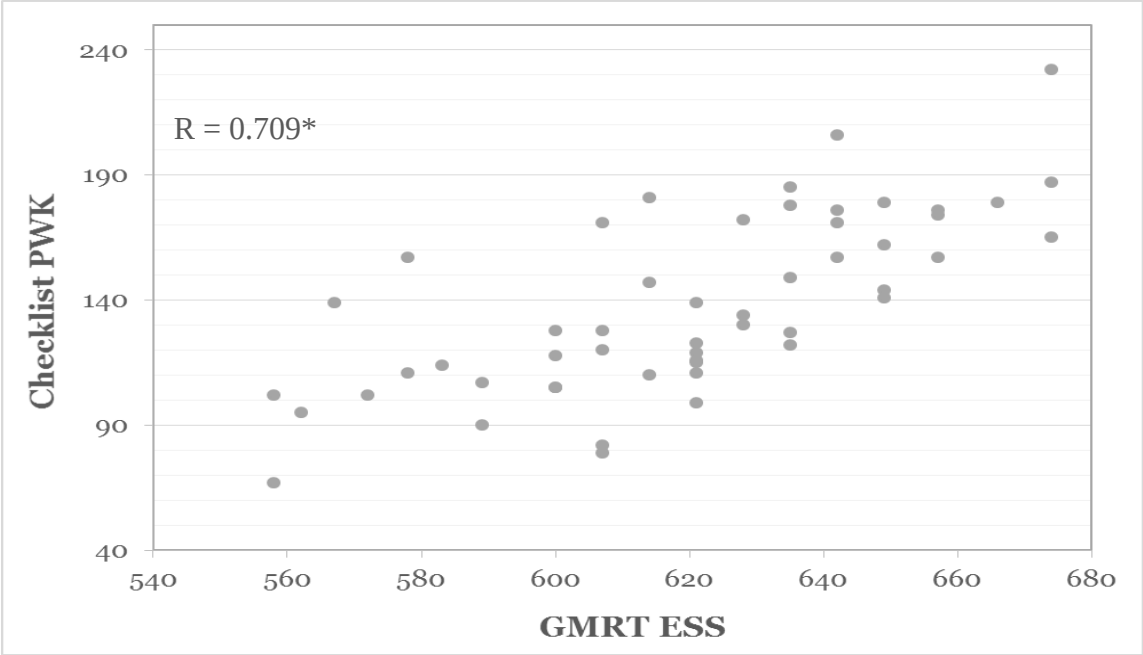

Figures 3.8 Condition 2 relationship between GL scores and GMRT results.



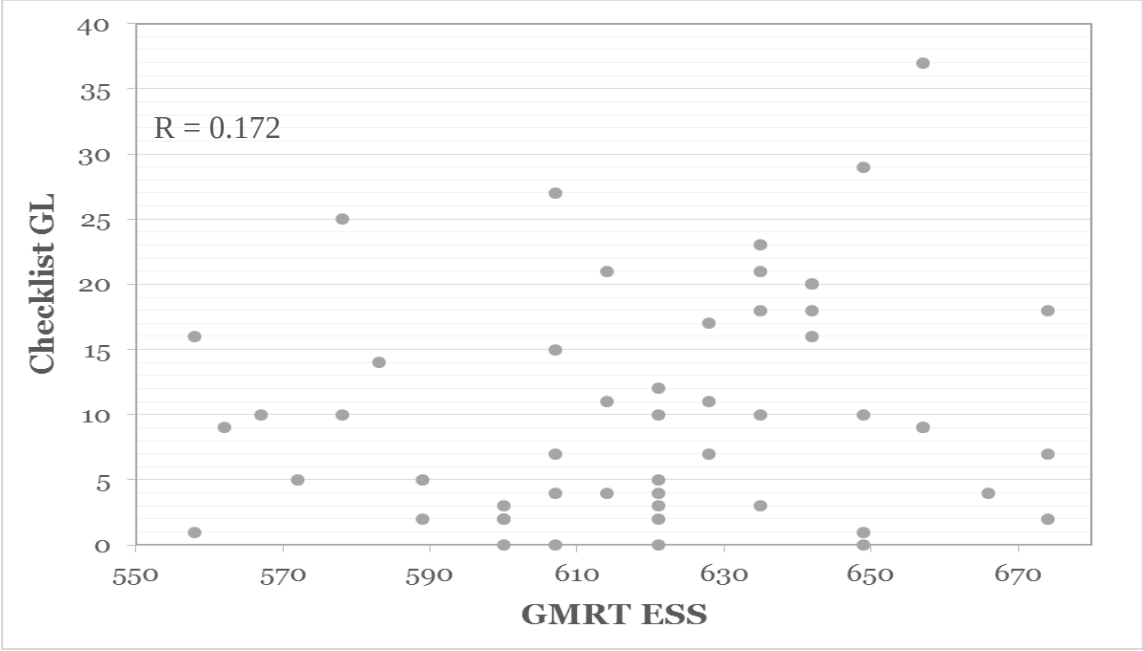

Figures 3.9 Condition 3 relationship between PWK scores and PPVT results.




Figures 3.10 Condition 3 relationship between GL scores and PPVT results.




Figures 3.11 Condition 3 relationship between PWK scores and GMRT results.




Figures 3.12 Condition 3 relationship between GL scores and GMRT results.

Then we reversed the order of the last two steps. This allowed us to determine the extent to which guessing behavior uniquely predicted norm-referenced vocabulary scores above and beyond that predicted by self-reported scores, as well as whether the linear or quadratic function (or both) of GL scores were most important in explaining individual differences in norm-referenced vocabulary test scores.

In Condition 1, the best fitting model for predicting PPVT scores included PWK and GL. Whereas PWK scores alone explained 45% variance, GL scores explained an additional 15% unique variance for a total of 61% variance in PPVT scores explained. Similarly, the best fitting model for predicting GMRT scores also included PWK and GL. PWK scores alone explained 41% variance, GL scores explained an additional 24% unique variance for a total of 65% variance in GMRT scores explained.

In Condition 2, the best fitting model for predicting PPVT scores likewise included PWK and GL scores. PWK scores alone explained 27% variance. The addition of GL scores contributed an additional 16% of the variance in PPVT scores, for a total of 43% variance explained. A similar pattern was observed for predicting GMRT scores in Condition 2. PWK alone explained 36% variance, and GL scores explained an additional 16% unique variance, for a total of 52% variance explained.

In Condition 3, a total of 26% of variance in PPVT scores was accounted for in a final regression model including only PWK scores. Neither GL nor the square of GL contributed significant unique variance, although the *p*-value for GL was 0.054. However, the best fitting model for predicting GMRT scores in Condition 3 included PWK, GL, and the square of GL scores. PWK explained 50% unique variance when entered first. GL explained 12% unique variance when entered last, and the square of GL

explained 5% unique variance when entered last. In total, the full model accounted for 67% of the variance in GMRT scores.

Table 3.4 Results of multiple regression analyses examining variance in PPVT scores

| Condition | Step | Variable | ß (final model) | Total $R^2$ | $R^2$ Δ | *F*Δ | *p* |
|---|---|---|---|---|---|---|---|
| 1 | 1 | PWK | 0.805 | 0.454 | 0.454 | 43.214 | < 0.001 |
| | 2 | GL | -0.397 | 0.605 | 0.151 | 19.426 | < 0.001 |
| | 3 | $GL^2$ | -0.014 | 0.605 | 0.000 | 0.005 | 0.943 |
| | 2 | $GL^2$ | -0.014 | 0.574 | 0.120 | 14.425 | < 0.001 |
| | 3 | GL | -0.397 | 0.605 | 0.030 | 3.827 | 0.056 |
| 2 | 1 | PWK | 0.754 | 0.267 | 0.267 | 17.452 | < 0.001 |
| | 2 | GL | -0.780 | 0.430 | 0.164 | 13.489 | 0.001 |
| | 3 | $GL^2$ | 0.330 | 0.442 | 0.012 | 0.976 | 0.328 |
| | 2 | $GL^2$ | 0.330 | 0.380 | 0.114 | 8.622 | 0.005 |
| | 3 | GL | -0.780 | 0.442 | 0.062 | 5.086 | 0.029 |
| 3 | 1 | PWK | 0.706 | 0.257 | 0.257 | 17.298 | < 0.001 |
| | 2 | GL | -0.475 | 0.312 | 0.055 | 3.890 | 0.054 |
| | 3 | $GL^2$ | 0.179 | 0.315 | 0.004 | 0.261 | 0.611 |
| | 2 | $GL^2$ | 0.179 | 0.293 | 0.036 | 2.483 | 0.121 |
| | 3 | GL | -0.475 | 0.315 | 0.023 | 1.580 | 0.215 |

Table 3.5 Results of multiple regression analyses examining variance in GMRT scores

| Condition | Step | Variable | ß (final model) | Total $R^2$ | $R^2$ Δ | *F*Δ | *p* |
|---|---|---|---|---|---|---|---|
| 1 | 1 | PWK | 0.802 | 0.405 | 0.405 | 35.373 | 0.000 |
| | 2 | GL | -0.521 | 0.646 | 0.241 | 34.737 | 0.000 |
| | 3 | $GL^2$ | 0.003 | 0.646 | 0.000 | 0.000 | 0.985 |
| | 2 | $GL^2$ | 0.003 | 0.594 | 0.189 | 23.706 | 0.000 |
| | 3 | GL | -0.521 | 0.646 | 0.052 | 7.384 | 0.009 |
| 2 | 1 | PWK | 0.834 | 0.362 | 0.362 | 27.186 | < 0.001 |
| | 2 | GL | -0.709 | 0.523 | 0.161 | 15.860 | < 0.001 |
| | 3 | $GL^2$ | 0.026 | 0.530 | 0.007 | 0.716 | 0.402 |
| | 2 | $GL^2$ | 0.260 | 0.479 | 0.117 | 10.601 | 0.002 |
| | 3 | GL | -0.709 | 0.530 | 0.051 | 4.981 | 0.031 |
| 3 | 1 | PWK | 1.032 | 0.503 | 0.503 | 50.538 | < 0.001 |
| | 2 | GL | -1.082 | 0.621 | 0.118 | 15.279 | < 0.001 |
| | 3 | $GL^2$ | 0.652 | 0.670 | 0.049 | 7.169 | 0.010 |
| | 2 | $GL^2$ | 0.652 | 0.553 | 0.051 | 5.563 | 0.022 |
| | 3 | GL | -1.082 | 0.670 | 0.117 | 16.993 | < 0.001 |

## 3.3 EXIT SURVEY

The fourth research question addressed students' perceptions of construct validity, by examining exit survey data. Survey data was available for 105 participants (33, 39,

and 33 participants in Conditions 1, 2, and 3, respectively). Participant responses (short answers to open-ended questions) were read and grouped into categories based on similar responses to each question. Surveys from 32 participants were excluded because their responses pertained to the norm-referenced assessments, whereas the instructions stated to only evaluate the web-based self-assessment. Several participants made more than one comment for questions 1, 2, and 4. Each comment was included in the data analysis. When fewer than 5 individuals gave similar responses, these responses were collapsed under the category "Other." Non-responses were considered a separate category from Other. Question 1 asked for participants' general thoughts about the assessment and was intentionally designed to be a warm-up question. The topics of participants' responses to Question 1 varied widely. Therefore, descriptive analyses were restricted to responses to Questions 2-4.

### 3.4 EXIT SURVEY: QUESTION 2

*Do you think the assessment adequately measured your vocabulary knowledge? Why or why not?* There were 105 total responses to question 2, which are broken down by condition in Table 3.6. The results suggest that the majority of participants overall thought the assessment adequately measured their vocabulary knowledge. The proportion varied somewhat between groups receiving different conditions of the checklist, with 66.7% of participants in Condition 1, 71.6% of participants in Condition 2, and 80.7% of participants in Condition 3 responding favorably.

### 3.5 EXIT SURVEY: QUESTION 3

*If you could select one thing, what did you like best about the assessment?* Participants made a total of 112 comments about what they enjoyed the most about the

Table 3.6 Participant Responses on Survey Question 2

| Response | Percentage of all Responses (n =105) | Percentage of Responses from Condition 1 Participants (n= 33) | Percentage of Responses from Condition 2 Participants (n= 39) | Percentage of Responses from Condition 3 Participants (n= 26) |
|---|---|---|---|---|
| Yes | 54.3% | 51.5% | 56.4% | 69.2% |
| Partial | 12.4% | 15.2% | 15.2% | 11.5% |
| No | 28.6% | 27.3% | 27.3% | 11.5% |
| IDK/NR | 4.8% | 6.1% | 2.6% | 7.7% |

*Note: Column sums 100.0% ± 0.1% due to rounding*

assessment in Table 3.7. The most common response made by participants was the self-rating system (i.e., being able to rate their levels of knowledge), comprising approximately 21% of total comments. Such responses more often came from participants in Condition 2 (41.7%) than from participants in Conditions 1 and 3 (25.0% and 33.3%, respectively). Other favorable comments mentioned the length of the assessment (13.4% of total comments), the overall easiness of completing the assessment (12.5% of total comments), the appeal of providing synonyms for highly known words in Conditions 2 and 3 (13% of total comments), and the assessment design (11.6% of total comments). Of the participants who referenced overall easiness, the highest percentage participated in Condition 1 (57.1%), where no follow-up questions or feedback were

provided. Accordingly, 60.0% of participants who reported enjoying the challenge of completing the assessment were in Condition 3. One participant from Condition 3 commented on how the synonym generation task made her more conscientious about rating her levels of knowledge.

### 3.6 EXIT SURVEY: QUESTION 4

*If you could make any improvements to the web assessment, what would you do?*

107 total comments were made regarding potential improvements that could be made to the program. As shown in Table 3.8, nearly half of all responses were grouped in the "Other" category, as they involved comments made by fewer than five people. The majority of responses pertained to the follow-up questions (or lack thereof, in Condition 1). Approximately 16% of participants overall suggested no changes to the assessment, but approximately 37% of participants requested changes to the follow-up task. Suggested changes to the follow up task included allowing more than one word for the follow-up task in Conditions 2 and 3 (20.6% of total responses) and designing a new follow-up task (15.9% of total responses). Of the participants who suggested a new task, 20% were from Condition 1 and suggested a follow-up task be added. Condition 2 comments were less likely to suggest using a different follow-up task (7.7%) but more likely to suggest using more than one word in the follow-up task (30.8%) possibly due to the feedback they received for their answers. The category of "other" included comments about item variety, altering the assessment design, adding visuals or audio, and no response.

Table 3.7 Participant Responses on Survey Question 3

| Response | Percentage of all Responses (n =112) | Percentage of Responses from Condition 1 Participants (n= 35) | Percentage of Responses from Condition 2 Participants (n=43) | Percentage of Responses from Condition 3 Participants (n=34) |
|---|---|---|---|---|
| Level of Knowledge Range | 21.4 | 25.0 | 41.7 | 33.3 |
| Length | 13.4 | 40.0 | 26.7 | 33.3 |
| Overall Easiness | 12.5 | 57.1 | 21.4 | 21.4 |
| Providing Synonyms | 12.5 | 0.0 | 50.0 | 50.0 |
| Aesthetics/Design | 11.6 | 30.8 | 38.5 | 30.8 |
| Item Variety | 8.0 | 44.4 | 44.4 | 11.1 |
| The Challenge of Completing the Assessment | 4.5 | 20.0 | 20.0 | 60.0 |
| Nothing/No Response | 4.5 | 20.0 | 80.0 | 0.0 |
| Other | 11.6 | 38.5 | 38.5 | 23.1 |

*Note: Column sums 100.0% ± 0.1% due to rounding*

Table 3.8 Participant Responses on Survey Question 4

| | Percentage of all Responses (n =107) | Percentage of Responses from Condition 1 Participants (n= 35) | Percentage of Responses from Condition 2 Participants (n=39) | Percentage of Responses from Condition 3 Participants (n=33) |
|---|---|---|---|---|
| Allow more than one word responses in follow-up task | 20.6 | 0.0 | 30.8 | 30.3 |
| Use a different follow-up task | 15.9 | 20.0 *participants requested addition of a follow-up task | 7.7 | 21.2 |
| No change | 15.9 | 48.6 | 10.3 | 39.4 |
| Other | 47.7 | 51.4 | 51.3 | 9.1 |

*Note: Column sums 100.0% ± 0.1% due to rounding*

# CHAPTER 4

# DISCUSSION

The primary purpose of this study was to examine the construct and concurrent validity of a self-rating checklist assessment of vocabulary knowledge. Our first research question observed the extent to which students accurately reported their knowledge on the self-rating checklist assessment by evaluating GL scores and student responses on follow-up questions. Ackerman and Ellingsen (2014) concluded that college students are likely to overclaim their vocabulary knowledge. In their study, 98% percent of participants overclaimed their vocabulary knowledge at least once, and an average of 44% of words were incorrectly defined in the definition generation task. While Ackerman and Ellingsen defined overclaiming as words participants claimed to know but could not define, we extended the definition of overclaiming to also include pseudowords that participants claimed to know. We noticed similar rates of overclaiming among all three conditions in our study. In Conditions 1, 2, and 3, 97%, 88%, and 93% of participants guessed at least once, and the number of guesses in all three conditions was about 30% (9 out of 30 pseudowords in each group). Additionally, follow-up question accuracy means in Conditions 2 and 3 were 60% and 54%, respectively. These rates indicate moderate levels of overclaiming. However, when we consider participants' ratings of levels of partial word knowledge, we see that participants' GL scores range from 10.39-11.46 out of a total possible GL score of 90. These findings suggest that while participants do guess occasionally, most do not drastically overestimate their knowledge. The participants

appeared to honestly rate their levels of knowledge by rating items as low as possible when they guessed. These findings support the notion of construct validity and answer the first research question by suggesting that participants attempted to accurately rate their levels of knowledge on the self-rating checklist assessment. These results also demonstrate the value of incorporating partial word knowledge in vocabulary assessments, which other researchers have not assessed. Binary measures of vocabulary knowledge may lead to higher rates of overclaiming (*e.g.*, when assessment formats allow for guessing) or underestimates of word knowledge (*e.g.*, when participants know some information about a word, but are unable to define it out of context). Ratings of partial word knowledge allow for the examination of the degree to which participants overestimate, providing a more valid representation of word knowledge.

Ackerman and Ellingsen (2014) observed that students with higher ability were more likely to overclaim their vocabulary knowledge. When we examined guessing with pseudoword foils, we observed similar rates across ability levels. While guessing and skill (determined by results on norm-referenced assessments) were not significantly correlated in our study, skill and follow-up question accuracy were, converging with the findings of Ackerman and Ellingsen. This may be due to the increased difficulty of the follow-up question task, a synonym generation task, similar to the definition generation task in the assessment used by Ackerman and Ellingsen. In order to accurately provide a synonym or definition for words, participants must have high levels of knowledge of the words.

The second research question evaluated how well results on the self-rating checklist assessment predicted results on existing norm-referenced measures. In all three

conditions of the assessment, participants' results on the self-rating checklist were positively correlated with their results on norm-referenced measures ($r$ = 0.507-0.709) and to a greater degree than TK scores were ($r$ = .33-.53). These findings extend those of Ackerman and Ellingsen (2014) who demonstrated preliminary evidence of concurrent validity with a self-rating checklist that did not account for partial word knowledge. Our results indicate that accounting for partial word knowledge explains more variance in norm-referenced assessment scores. Furthermore, our results underscore the value of including pseudoword foils. Whereas Ackerman and Ellingsen hypothesized that the inclusion of foils in checklist assessments could be distracting and could fail to accurately assess vocabulary knowledge, we found that including foils to measure and account for guessing behavior in linear regression models led to a more accurate prediction of norm-referenced vocabulary scores. Taken together, these results provide strong evidence of the concurrent validity of checklist self-assessment vocabulary measures.

Our third question asked whether the presence of feedback or follow-up questions influenced student response patterns. Our results for this question are inconclusive. On one hand, the overall results of regression models appeared similar between the conditions. On the other hand, participant survey responses seem to suggest differences among conditions. The three groups did not differ significantly in norm-referenced vocabulary scores, or in their rates of guessing, and participants in Conditions 2 and 3 showed no significant difference in follow-up question accuracy. However, the sample size may not have been large enough to detect small effects, for example, in norm-referenced vocabulary scores or in follow-up question accuracy. Despite the lack of a significant difference between Conditions 2 and 3, participants in Condition 1 rated their

vocabulary knowledge as higher than participants in Conditions 2 and 3. It is difficult to fully evaluate whether this is related to a difference in response patterns because the pattern of mean PWK ratings across the three conditions matches the pattern of group means for the norm-referenced assessments. That is, it may be that the PWK score was a more sensitive indicator of true differences in vocabulary knowledge that were not statistically significant for the norm-referenced measures. Turning to regression models, the general pattern of findings was similar across conditions, although the amount of variance explained varied. In almost all cases, including an estimate of guessing improved model fit and explained unique variance beyond students' self-reported vocabulary knowledge, suggesting similar results across conditions. Alternatively, student responses on survey questions appear to show differences do exist among groups. In Condition 1 where there was no follow-up question, 20% of participants discussed how they felt as though they were more likely to overclaim their knowledge. Participants in Conditions 2 and 3 (31% and 30%, respectively) reported they were more likely to underclaim their vocabulary knowledge because of the difficulty of their follow-up task. Given these varied results, a definitive conclusion cannot be made regarding whether or not differences exist between the conditions.

The fourth research question examined participants' perceptions of construct validity. When asked if they believed if the assessment adequately measured their vocabulary knowledge, the majority of participants (66.7%) responded favorably by saying they believed the assessment was an adequate or partially adequate measure. People across all three conditions thought the self-rating checklist was a good assessment, but more people in Condition 3 responded favorably. Overall, the majority of

participants believed that this assessment format could accurately assess their vocabulary knowledge, demonstrating good perceptions of construct validity among participants.

Altogether, the findings of this study suggest that a self-rating checklist assessment can be a valid assessment of vocabulary knowledge, displaying both concurrent and construct validity. These findings can be considered robust, as concurrent and construct validity were essentially replicated across three separate samples, with three different conditions of a checklist assessment.

Overall, there was little evidence that including follow-up questions and/or feedback improved the prediction of norm-referenced scores. The checklist scores in Condition 1 accounted for a larger amount of variance in norm-referenced vocabulary scores than checklist scores in Condition 2 or Condition 3, except for the case of GMRT scores, where Condition 3 accounted for slightly more variance than Condition 1 (67% vs. 65%). However, an important limitation should be acknowledged. The study employed a between-subjects design to evaluate possible effects of condition. Participants in Condition 1 reported knowing significantly more words than individuals in Conditions 2 or 3. This pattern with PWK and TK scores in Condition 1 shows a similar pattern to norm-referenced assessment scores in Condition 1. Although not significantly different, mean PPVT-4 and GMRT-4 scores were numerically higher for Condition 1 than Conditions 2 and 3, suggesting that groups may not have had equivalent vocabulary knowledge at the start of the study. Future studies employing a within-subjects design may be better able to examine differences between conditions.

Future directions to provide further support for a web-based, self-rating checklist of vocabulary knowledge may include more detailed analyses of participant responses

within each band of the assessment to observe performance across levels of word difficulty. Additionally, the reliability of checklist scores remains to be determined. While our results and the results of Ackerman and Ellingsen (2014) contribute to the validity of such assessment, determining its reliability is important in order to suggest clinical utility of a self-rating checklist of vocabulary knowledge.

A web-based self-rating checklist assessment can be advantageous for several reasons. Students can take it quickly, it is easily adjustable, it can take into account partial word knowledge, and teachers can use the assessment to determine what words students need to be taught according to their theory of learning. When accounting for partial word knowledge, unlike traditional vocabulary assessments, this format can provide more accurate results regarding vocabulary knowledge. The results of this study show that a self-rating checklist that takes into account partial word knowledge can be a practical and valid assessment format with possible educational and clinical utility.

REFERENCES

Ackerman, P.L., & Ellingsen, V.J. (2014). Vocabulary overclaiming — A complete approach: Ability, personality, self-concept correlates, and gender differences. *Intelligence*, *46*, 216-227.

Adlof, S.M., & Perfetti, C.A. (2013). Individual differences in word learning and reading ability. In A. Stone, B. Ehren, E. Silliman, & G. Wallach (Eds.), *Handbook of Language and Literacy Development and Disorders, 2nd Edition* (pp. 246-264). New York: Guilford.

Beck, I. L., & McKeown, M. G. (1985) Teaching vocabulary: Making the instruction fit the goal. *Educational Perspectives*, *23*(1), 11-15.

Beck, I., McKeown, M., & Omanson, R. (1987). The effects and use of diverse vocabulary instructional techniques. In M. McKeown & M.E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 147-163). Hillsdale, NJ: Erlbaum.

Biemiller, A. (2010). *Words worth teaching: Closing the vocabulary gap*. Columbus, OH: SRA/McGraw-Hill.

Christ, T. (2011). Moving past "right" or "wrong": Toward a continuum of young children's semantic knowledge. *Journal of Literacy Research*, *43*(2), 130-158.

Cunningham, A.E., & Stanovich, K.E. (1998). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*, 934–945.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English*, *42*, 895–901.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test (4th ed.)*. Minneapolis, MN: NCS Pearson.

Durso, F. T., & Shore, W. J. (1991). Partial knowledge of word meanings. *Journal of Experimental Psychology: General*, *120*, 190–202.

Hiebert, H. H. (2005) In pursuit of an effective, efficient, vocabulary curriculum for elementary students. In E. H. Hiebert & M. L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 243-264). Mahwah, NJ: Erlbaum.

Kan, P. F., & Windsor, J. (2010). Word learning in children with primary language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *53*, 739-756.

Kirkpatrick, E. A. (1905). A vocabulary test. *Popular Science Monthly*, *LXX*, 157–164.

Kuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627-633.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods*, *44*, 978-990.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2002). *Gates-MacGinitie Reading Tests, Fourth Edition, Forms S and T*. Itasca, IL: Riverside Publishing.

Miller, G. A. (1999). On knowing a word. *Annual Review of Psychology*, *50*, 1-19.

Nagy, W. E. & Scott, J. A. (2000). Vocabulary processes. In M. L. Kamil, P. Mosethal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research (*Vol. 3, pp. 269-284). Mahwah, NJ: Erlbaum.

Warmington, M., Hitch, G. J., & Gathercole, S. E. (2013). Improving word learning in children using an errorless technique. *Journal of Experimental Child Psychology*, *114*, 456-465.

Whipple, G. M. (1908). Vocabulary and word-building tests. *Psychological Review*, *15*(2), 94–105.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science.