Published in final edited form as:
Lang Assess Q. 2012 ; 9(2): 152–171. doi:10.1080/15434303.2011.613504.

Construct Validity and Measurement Invariance of the Peabody Picture Vocabulary Test-III Form A in the Performance of Struggling Adult Readers: Rasch Modeling

Hye Pae,
University of Cincinnati

Daphne Greenberg, and
Georgia State University

Robin D Morris
Georgia State University

Abstract

Purpose—The aim of this study was to apply the Rasch model to an analysis of the psychometric properties of the PPVT-III Form A items with struggling adult readers.

Methods—The PPVT-IIIA was administered to 229 African-American adults whose isolated word reading skills were between third and fifth grades. Conformity of the adults' performance on the PPVT-III items was evaluated using the Winsteps software.

Results—Analysis of all PPVT-IIIA items combined did not fully support its use as a useful measure of receptive vocabulary for struggling adult readers who were African Americans. To achieve an adequate model fit, items 73 through item 156 were analyzed. The items analyzed showed adequate internal consistency reliability, unidimensionality, and freedom from differential item functioning for ability, gender, and age, with a minor modification.

Discussion—With an appropriate treatment of misfit items, the results supported the measurement properties, internal consistency reliability, unidimensionality of the PPVT-IIIA items, and measurement invariance of the test across subgroups of ability, age, and gender.

The Peabody Picture Vocabulary Test (PPVT) has been widely used in the Unites States for decades as a receptive vocabulary or verbal ability measure by clinicians, educators, and researchers for both children and adults (Bell, Lassiter, Matthews, & Hutchinson, 2001; Campbell, Bell, & Keith, 2001; Carvajal, Nowark, & Fraas, 2000; Gerde & Powell, 2009; McLaren & Richards, 1986; Stockman, 2000; Walker, Givens, Cranford, Holbert, & Walker, 2006; Washington & Craig, 1999; Vaughn, Beaver, Wexler, DeLisi, & Roberts, 2011). PPVT scores have been used in multiple arenas as proxy scores of verbal intelligence in the clinical setting (Carvajal, Nowark, & Fraas, 2000), as scores used to determine eligibility for intervention programs (Majsterek & Lord, 1991), as scores used to identify children at risk for language delay (Dollaghan & Campbell, 2009), and as indicators of receptive vocabulary knowledge in research (Gerde & Powell, 2009). Recently, its use has been expanded to the bilingual area (Dixon, 2011; Millett, Atwill, Blanchard, & Gorin, 2008; Quiroz, Snow, & Zhao, 2010).

Correspondence should be sent to Hye Pae, PhD, University of Cincinnati, School of Education, 615R Teachers College, P.O. Box 210022, Cincinnati, 45221 United States., hye.pae@uc.edu.

The PPVT is an isolated single word, receptive vocabulary measure designed to be individually administered to test-takers in a wide range of ages, ranging from toddlers to elderly individuals. The examinee is asked to indicate verbally or nonverbally which of four pictures on the easel page best represents the meaning of an orally presented word by the examiner. For example, the examiner says the word *gigantic* while simultaneously showing the test-taker an easel page containing pictures of a target and three foils. An entry set is recommended based on the test-taker's age, but the examinee has to establish a basal (0 or 1 error in a set) to move on to the next set. If a basal is not established, backward administration is performed until he/she establishes a basal point. Each set includes 12 items and the test has 17 sets, comprising 204 stimuli in total. The task continues with an increasing degree of difficulty until the test-taker makes 8 errors out of 12 words in a set. The raw score is computed by the total number correct, and is converted into a standard score based on age, if necessary.

The PPVT is continuously updated to reflect more up-to-date vocabulary and changes in the general population make-up. For example, one of the changes made to the third edition of the PPVT was an inclusion of more ethnic minority participants in the normative sample. The PPVT-Revised (PPVT-R) comprised 14.6% ethnic minority representation in the normative sample, and the PPVT-III increased the ethnic minority proportion to 34%, which was higher than twice that of the PPVT-R, as the rate of the minorities increased in the general population (Williams & Wang, 1997).

Individuals with an ethnic minority background have been found to perform differently on a series of vocabulary tests (Stockman, 2000). This difference in performance can be attributed to the disproportionate representation of minority groups in the lower socioeconomic status as well as to variations in cultural, linguistic, and social experiences (Stockman, 2000). Administering a norm-referenced test like the PPVT and using its score uniformly, regardless of the culturally diverse backgrounds of test-takers, may result in inadequate score interpretation and use in research and clinical settings.

Although the normative group is composed of individuals ages 2 to 90 years in the United States population, psychometric research studies have focused on evaluations of the PPVT in children (Campbell, Bell, & Keith, 2001; Miller & Lee, 1993; Restrepo, Schwanenflugel, Blake, Neuharth-Pritchett, Cramer, & Ruston, 2006; Stockman, 2000; Webb, Cohen, & Schwanenflugel, 2008). Despite the expansion of its use to the adult population (Bell, Lassiter, Matthews, & Hutchinson, 2001; Carvajal, Nowark, & Fraas, 2000; Greenberg, Wise, Morris, Fredrick, Rodrigo, Nanda, & Pae, 2011; Vaughn, Beaver, Wexler, DeLisi, & Roberts, 2011), the psychometric evaluation and validity testing of this instrument for use with the adult population, particularly those with minority or unique backgrounds, have been absent. Since they are different from children in terms of prior knowledge, real-life experience, and semantic repertoire, adults may manifest a different profile of responses to the test. Because of these different profiles, if the norm-referenced measure is used uniformly and the same decision rule is applied to different populations, test fairness and comparable validity are possibly threatened in the interpretations of the test results and the clinical judgments made on the basis of the test scores.

Measuring Vocabulary and Rasch Modeling

Given the pervasive use of the PPVT, the importance of the PPVT-III's validity and precision cannot be understated. Such a norm-referenced measurement tool is expected to have a high degree of precision and accuracy, discriminating high performers from those who perform poorly, as well as functioning indifferentially across race, gender, and age. The evaluation of the difficulty on a particular test item is a challenge because the test-taker's

ability (person ability) and item difficulty are inherently abstract, but related, constructs. The Rasch probabilistic measurement model calibrates these two latent constructs (i.e., person ability and item difficulty) on the same metric scale to address this challenge (Linacre, 2010a). Rasch modeling provides an excellent methodology for quantifying person ability and item difficulty as well as evaluating test validity and reliability. This study is the first, to our knowledge, to examine the construct validity and measurement invariance of the PPVT-III Form A (PPVT-IIIA) through test differential functioning (DTF) and item differential functioning (DIF) in a unique minority group of African-American adults who struggle with reading. Of particular interest was the examination of whether the PPVT-IIIA became distorted relative to these particular participants' responses such that it made the resulting measurement potentially inaccurate. This is an important question because test fairness entails the principles of justice and beneficence (Kunnan, 2010). If a test shows comparable and equivalent construct validity and when fair score interpretations and decisions are made for different populations, an instrument is considered a bias-free tool, and no harm is inflicted on individuals. In order to achieve accuracy and appropriateness in score-based interpretations and decisions about African-American adults who struggle with reading, stability and accuracy in measurement must also take into account gender, age, and skill levels. In addressing these issues, Rasch modeling provides a detailed assessment of the response patterns, item fit, dimensionality, and the detection of item biases.

The use of total scores under a classical testing theory (CTT) framework to determine an individual's skills may mask his/her true ability because each item entails a unique meaning and a difficulty level. Advanced Rasch modeling offers a more precise measurement model under a one-parameter item response theory, which focuses on the probability that a person makes a particular response pattern according to his/her level of an underlying latent variable (Cohen, Kim, & Baker, 1993; Linacre, 2010a). Testing for differential item functioning ensures that test items function uniformly across various groups within a population, such as age, gender, or ability.

The Rasch model provides indicators of how well each item fits within the latent construct, utilizing the logarithmic transformation for estimates of person ability (i.e., person trait level) and item location (i.e., item difficulty; Linacre, 2010a). The person ability and item difficulty parameters are estimated simultaneously to produce estimates of an equal interval scale measured in logits (log odd units), which are independent of both the items and sample employed (Bonds & Fox, 2007; Smith, 2001). Therefore, the basic Rasch assumptions include (1) the relationship between person ability and item difficulty; that is, each person is related to an ability, while each item is characterized by a difficulty, and (2) unidimensionality, which means that person ability and item difficulty can be expressed by numbers along one dimension and that the probability can be computed from the difference between the numbers (Bond & Fox, 2007; Schumacker, 2004). Psychoeducational tools, such as the PPVT, require the evaluation of the extent to which an item is useful in assessing the underlying construct as well as the possible redundancy the item exhibits relative to other items on the same scale (Fox & Bond, 2007; Linacre, 2010a; Waugh & Addison, 1998).

The core elements of Rasch modeling are as follows:

Item Fit Statistics

Item fit is evaluated using infit and outfit statistics to assess the residual differences between expected and actual test-takers' responses. The infit statistic refers to the information-weighted statistic of the squared residuals (unexpected persons' responses) which are close to the item's location on the logit scale (Linacre, 2010a). The outfit statistic is not weighted and refers to an "outlier-sensitive fit statistic" (Linacre, 2010a).

Unidimensionality and Local Independence

A one-dimensional underlying construct is one of the fundamental requirements of the Rasch model. Unidimensionality is also evaluated using the principal component analysis (PCA) to identify common variance in the residuals. PCA of the residual is an analysis of the residual variance that shows unexplained relations between the item residuals after accounting for the primary Rasch dimension (Linacre, 2010a). Local independence is assumed to be met if a dominant unidimensional construct is extracted, because the residuals are not sufficient enough to affect measurement (Bond & Fox, 2007).

Measurement Invariance

Measurement invariance is a critical property of scientific measurement and refers to the stability of item and person parameters across repeated calibrations within the limits of measurement error (Bond & Fox, 2007). The Rasch model hypothesizes the invariance principle that the relative difficulty level of the item should be consistent across subsamples, and test items should not behave variably to the particular subgroup. If an item functions inconsistently for a certain group, the item diminishes the validity of the measure for the construct under consideration. Therefore, differentially functioning items in the specific group should be eliminated from the instrument tool to secure the construct validity (Bond & Fox, 2007). If the item of a test functions in a consistently different way for one group of test-takers than for another, the measure for one group may not be comparable with the measure for another, yielding a violation of the invariance principle.

Aims and Objectives

The purpose of this study was to apply the Rasch model to the investigation of the psychometric measurement properties of the PPVT-III by estimating both item and person parameters. Three research questions were addressed in this study:

1. How well did item difficulty represent construct validity in the PPVT-IIIA for a sample of African-American adults who struggled with reading?
2. How did the indicators of receptive skills of struggling adult readers cluster along the unidimensional construct of the Rasch measurement model?
3. Were the test and items of the PPVT-III invariant between two subgroups by ability (i.e., higher and lower skill groups) and by gender?

The first question examined the probabilistic relationship between item difficulty and person ability along a single continuum. The second question addressed whether the data form a single latent trait which explained all the variance in the data. The final question examined measurement invariance to evaluate whether items were functioning equivalently across subgroups, given that the Rasch model requires item estimation to be independent of any subgroup of individuals taking the test. The rationale for the measurement invariance evaluation across subgroups by ability and gender is based on significant differences found in a series of research studies. Specifically, statistical differences in performance by gender were found in computer-based test versions (Gallagher, Bridgeman, & Cahalan, 2002), item differential estimates in the tests of English as a foreign language proficiency (Ryan & Bachman, 1992), L2 comprehension and vocabulary learning in the video-based computer-assisted language learning program (Lin, 2011), and task performance in tape-mediated assessment of speaking (Lumley & O'Sullivan, 2005). Ability-group differences were also found in language learning (Pae, Sevcik, & Morris, 2010; Qingquan, Chatupote, & Teo, 2008). The measurement invariance was examined through DTF and DIF. We investigated DTF by doing separate analyses for each subgroup, and then comparing the two sets of item

difficulties. DIF was investigated by estimating two subsamples' difficulties for the specific items while controlling for all the other item difficulties and person measures.

METHOD

Participants

The participants were 229 struggling adult readers[i] whose isolated word-reading skills fell between the third- and fifth-grade levels on the Letter/Word Identification subtest of the Woodcock-Johnson Tests of Achievement-III (Woodcock, McGrew, & Mather, 2001). The participants' mean age was 35 years of age (SD=15.77), ranging from 16 to 72. All the participants were African-American English native speakers. Males accounted for 28% of the sample and females were 72%. Their formal educational level was 10.10 years of formal schooling (SD=1.61; range=5–14).

Measure

Form A of the PPVT-III (Dunn & Dunn, 1997) was administered. As indicated earlier, the PPVT-III is designed to measure an individual's receptive vocabulary knowledge and verbal ability for Standard American English, and is normed on American English speakers with an age range from 2 to 90 years old. In this study, standard test procedure was followed, with examinees being asked to point verbally or nonverbally to the picture that best described the stimulus word upon the examiner's verbal presentation of a single word. According to the examiner's manual of the PPVT-IIIA (Dunn & Dunn, 1997), internal consistency alphas for the age groups from 2 to 90 range from .92 to .98 (median: .95), and split-half reliability ranges from .86 to .96 (median: .94). The test-retest coefficients range from .91 to .94.

The participants' mean raw score on the PPVT-IIIA was 133.83 (SD=18.45; range=75–182), when the standard test basal and ceiling rules were applied. The PPVT-IIIA manual specifies a basal rule to be one or no item error in an item set, and a ceiling rule to be 8 or more errors in an item set. According to the PPVT manual, a test-taker who is suspected of having lower receptive vocabulary skills below the 25th percentile should begin with a lower item than the suggested entry point, since the standard basal criterion was derived from the probability that 50% of the normative age group would meet the basal (Dunn & Dunn, 1997). Previous research has indicated that struggling adult readers' vocabulary skills are commensurate with their reading skills rather than with their chronological ages (Byrne, Crowe, Hale, Meek, & Epps, 1996; Sabatini, Sawaki, Shore, Scarborough, 2010). On the basis of the participants' word reading level (grade 3 through grade 5), the participants were administered items from Set 7, which begins with item 73 and is considered the entry point for ages 8 and 9 (typical ages of children in grade 3). Only participants who established a basal point at Set 7 were included in this study. When participants reached their ceiling items, the test was discontinued. The mean standard score was 72.39 (SD=9.8; range=40–94), and mean age equivalency was 10.69 years (SD=2.76; range=3.11–22.00).

Procedure

The PPVT-IIIA was administered by trained graduate students. Prior to testing the adult struggling readers, testers were given extensive training by the project's psychometrician to ensure appropriate administration by carefully going over issues of adult literacy sensitivity and the assessment protocol.

[i]This sample was part of a larger study on struggling adult readers.

Data Analysis

The data were analyzed using the Winsteps software (Linacre, 2010b) to obtain item difficulty and person ability measures on the PPVT-IIIA. Since Rasch modeling is less concerned about statistical power than about stability estimation, and Winsteps does not impute data, the items after the ceiling were treated as missing data. According to Linacre (2010c), it is fairer and more accurate to score unadministered above ceiling items as "missing" than to decide whether the test-taker would have succeeded or failed. For research question 1, items 73 to 204 (the last possible item) were analyzed in order to evaluate overall item functioning for this adult sample. Research questions 2 and 3 were examined using items 73 through 156, since the majority of the adult sample was given those items before reaching their ceiling.

Construct validity was evaluated through the evaluation of dimensionality, hierarchical differentiation of items, and item dispersion along the latent variable. Fit statistics and PCA of the standardized Rasch residuals were used to examine the magnitude of variance in the measure which was explained by the first order factor (Linacre, 2010a). Following Linacre's (2010a) recommendation for high-stakes tests, mean square (MNSQ) fit statistics between 0.8 and 1.2 were considered acceptable and Zstd values between −2.00 and +2.00. Analyses were rerun after deleting misfit items to evaluate whether deleted items affected the accuracy of the test and whether the error rate of the model estimates was reduced without misfit items (Hart & Wright, 2002).

The influence of vocabulary skill level on person ability and item difficulty was examined through DTF and DIF to evaluate if there was a systematic bias toward the two ability subsamples. There are multiplicities involved in group assignment. Although a median split can be arbitrary, the results of many research studies have shown robust group differences using a median split. For instance, a median split differentiated the two groups on three of five outcome variables, and two of five growth variables in studies by Torgesen et al. (2001) and Vellutino et al. (1996). Moreover, a multitude of research articles have adopted a median-split method in the field (Bundgaard-Nielsen, Best, & Tyler, 2011; Isaacs & Trofimovich, 2011; Pae, Sevcik, & Morris, 2010). Hence, a median-split was used to classify the ability groups (i.e., high and low) in this study[ii]. The standard score of the median was 74. Since it was expected that the participants would perform uniformly on the test, uniform DIF was applied to examine the differentiation between the two groups.

RESULTS

Construct Validity and Predictive Validity (Research Question 1)

The overall goodness-of-fit of the items and persons indicated that these data fit the Rasch model. A unit normal distribution of the standardized residuals N(0,1) showed reasonable values [N(0, 1.02)]. The person reliability, which is equivalent to test reliability in a CTT model was r = .91 and separation index was 3.22. The item reliability was r = .96 and the separation index was 4.80.

In order to evaluate whether the item responses on the PPVT-III were aligned with the abilities of the persons, point-measure correlation coefficients were obtained. There were 13 items that resulted in negative correlations, indicating that the responses to these items were contradictory to the direction of the latent variable (these items were: 201, 199, 200, 202, 194, 195, 203, 188, 187, 184, 74, 175, and 193— listed from the largest negative correlation

[ii]Although we ran DTF and DIF for age by splitting the sample into two groups (one with 40 and younger and the other 41 and older), the results are not reported in this paper to avoid redundancy. The results showed a very similar pattern to those of ability and gender, indicating little variance across the two groups.

coefficient to the lowest). These items showed a disparity between observed correlations and the expected correlations in the Rasch model.

The observed item response was analyzed for the latent trait and item location on the Rasch scale. The item distribution map is plotted in Figure 1, along the same latent trait, to illustrate the distribution of item-difficulty estimates and person-ability estimates on the same logit scale. Figure 1 indicates that the lower on the scale, the easier the items are and the less able the student. The vertical dashed line and adjacent numbers represent the common logit scale of person ability and item difficulty on the same scale. The items were distributed along the logit scale from −6 to +5. The lower and upper ends of item difficulty for the PPVT-IIIA did not map adequately with the person ability. The easiest and the most difficult items have no persons assigned to provide good information about them. For instance, there was only one item (i.e., 148) near 1 logit, indicating an item-targeting problem. An item-targeting problem suggests less precision of measurement and larger person standard error than expected (Linacre, 2010a).

Overall, the item-person representation indicated that the items were not well matched for this sample. Although the item difficulty spanned ten units on the logit scale, Figure 1 shows that two logit points at the bottom and one logit point at the top were represented by only a few items (items 74 and 77 at the bottom of the scale and item 184 at the top). The participants were packed between −1 and +2 logits. It would be very hard to locate persons precisely at either end of the scale represented by the PPVT-IIIA items. Items 74 and 77 were too easy for the participants, while item 184 was too difficult for the sample. The person distribution was center-heavy in comparison to the item distribution.

In a similar vein, the precision (standard error) and accuracy (good-fit) of the measure were examined through a bubble chart as seen in Figure 2. The item arrangement showed a very similar pattern to that of Figure 1; the easiest item at the bottom and the most difficult at the top. The size of the bubble map indicated the standard errors of the measures along the vertical axis, the latent variable (Linacre, 2010a). Item location for the PPVT-IIIA demonstrated that the items also covered a narrow range −2 and +2 logits. Items 77 (*towing*) and 74 (*nostril*) were easiest for the participants, whereas item 184 (*reposing*) was the most difficult. The horizontal axis represents the fit (i.e., the accuracy of the measure) of the data to the latent variable. The overfit on the left shows that the responses are too predictable, whereas the underfit on the right indicates that the responses are too unpredictable from the Rasch model's perspective (Linacre, 2010a).

Based on these results, only items 73 through 156 were reanalyzed. This decision was made primarily because the majority of the participants were administered item 73 through item 156. To examine the extent to which the response to an item aligned with the underlying ability of a person, a point-measure correlation was obtained. For example, item 74 showed a negative correlation (point-measure correlation = −.08), while the expected correlation was .09. This indicated that the responses to this item contradicted the direction of the latent variable. As another example, item 77 showed a weak point-measure correlation coefficient (point-measure correlation = .07), meaning that it under-discriminated high performers from low performers. The fit statistics showed that item 74 was the most misfit (outfit mean square = 5.46). Although the infit mean squares fell within the recommended range, the outfit mean square values were relatively high, suggesting that the items were too unpredictable. Since it appeared that it distorted the model or degraded the measurement system, we eliminated item 74 from further analyses. The remaining items were then recalibrated and reevaluated. Without item 74, item 77 behaved much better, but items 90, 84, 106, and 75 became underfits, unproductive for the construction of the model. Hence, these five items were eliminated in the analyses. The removal of the misfit items enhanced

the Rasch model. Table 1 shows the misfit order and fit statistics. Figure 3 exhibits the distribution of the measure, before and after the removal of the misfits, along the x-axis, and the distribution of the fit on the y-axis. The capital letters indicate unpredictable items, while the lower case letters label predictable items. As seen in Figure 3a (before removal), item 74 (denoted as "A", red arrow), was an unpredictable outlier. The red rectangle in the figure 3a indicates unproductive items in the model. Figure 3b (after removal) shows an improved cross-plot with all items within the fit range.

We also examined the person measure cross-plot. It was apparent that a couple of persons (red arrows) were impacted by item 74 (see Figure 4a, before removal). After eliminating the misfit items, the model improved significantly (see Figure 4b, after removal). Since there was no intention to eliminate the misfit person from the pool, no further investigation on the person was performed.

Unidimensionality Structure and Item Fit (Research Question 2)

The unidimensional structure of the PPVT-IIIA was assessed through PCA using the corrected fit items only to evaluate the amount of variance explained by different components of the data. The standardized residual variance resulted in 26.8% (persons 9.6% and items 17.1%) of the variance explained by the Rasch model. Reckase (1979) has noted that the variance explained by the first factor should be greater than 20% as a minimal amount of variance for the identification of unidimensionality. Based on this criterion, the variance accounted for in this sample evidenced a practical amount for unidimensionality. Besides, given the participants' narrow range of ability to be included in the study (word reading skills between grade 3 to grade 5), the amount of variance explained could be considered a reasonable dispersion of persons, items, and person-item targeting for these participants. It also resulted in a first contrast with an eigenvalue of 4.1 (3.8% empirical and 5.1% modeled variances).

The distribution of each item's loading on the first contrast dimension was plotted against their item sets (7 to 13). Each item's stimulus word (in a set) is shown related to its factor loading in Figure 5. There is a slight trend from bottom left to top right. However, there is a contrast between the word "*upholstery*" and "*mammal*" in set 12. The word "*island*" is consistent with the pattern formed in set 7. Overall, no prominent sub-dimension was identified.

Invariance: Differential Test Functioning and Differential Item Functioning (Research Question 3)

The Rasch model hypothesizes that test items should not behave differently in any particular subgroups evaluated. If an item functions differently for certain groups, the item may be a threat to the validity of the measure for that construct. Therefore, differentially functioning items in the specific group should be eliminated from the instrument to obtain the construct validity under consideration. In order to evaluate test and item invariance, we divided the sample into two different subgroups according to gender and ability, and conducted item estimation for the test for each. A scatterplot of DTF is displayed in Figure 6, showing each item as a point in which the item estimates are invariant within error. The plot compares the performance of the subsamples by items. The high-ability participants produced the item estimates plotted on the y-axis, and the low-ability counterpart's estimates were used to calculate the item difficulty on the x-axis. The dotted line in the center is a Rasch modeled relationship line required for invariance, while the two solid lines are the 95% control confidence bands. As can be seen, there was a slight dispersion away from the commonality line. This was because the estimates were represented by a number of components relating to quantity and quality, including measurement error, in the usual Rasch estimation

procedure. It seems reasonable to say that the item estimates show invariance across the two subgroups.

DIF evaluates whether or not different subgroups with the same latent trait (e.g., gender or ability) had a different probability of responding to a test item, providing an indication of unexpected item behaviors on a test. Since the change in an item's difficulty was assumed consistent across different person groups, uniform DIF estimates were obtained. DIF estimated the difficulty of each item for each group, while placing constraints on all the other item difficulty and person ability measures. Figure 7 displays a DIF size plot, which was relative to the item difficulty. The graphical representations of DTF and DIF were useful diagnostic tools for evaluating the potential impact on the level of the entire scale (DTF) and on the item level (DIF). The gender plot showed less item fluctuation than the ability plot. Despite the slight differences in the estimates between the two subgroups, these results demonstrated evidence that items displayed no DIF across the subsamples. The group-specific item functioning demonstrated that the same underlying true ability had a similar probability to give a certain response.

In order to validate the graphic presentation, we hypothesized that there was no DIF for this instrument between the two person sub-groups. A Bonferroni *t*-test using the probability of each item was performed. There were no significant differences between the two subgroups of gender or ability ($t = -0.19$, *ns* for gender; $t = 0.18$ *ns* for ability).

DISCUSSION

A Rasch analysis was performed to assess person ability and item difficulty on PPVT-IIIA responses of struggling adult readers who were African Americans. Validity, unidimensionality, and DIF studies have been lacking in this particular population. This study is the first to carry out a rigorous psychometric evaluation of the PPVT-IIIA for African-American adults who struggle with reading. The results provide support for the useful measurement properties, reliability, unidimensionality, and measurement invariance of the PPVT-IIIA, with improvements possible with some modifications to the measure. An elimination of misfit items from the PPVT-IIIA yielded no substantial DIF.

In order to address the three research questions, two steps were undertaken to perform the analyses. First, we analyzed the items from item 73 (the first item of Set 7) to the last test item to evaluate whether the PPVT-IIIA was an appropriate measure to gauge the receptive vocabulary skills of the participants. The decision to analyze item 73 on was made because all the participants met the basal requirement at Set 7. The purpose of basal and ceiling rules is to shorten testing time, diminish the test-taker's frustration, and increase the proportion of useful items to predict an examinee's degree of proficiency. The basal rule assumes that items below a certain point would have been passed, if they had been administered, whereas the ceiling rule assumes that items beyond a certain point would have been failed, if they had been administered. Therefore, only actual administered items above item 72 were included in these analyses. This was felt to provide the best sample of each subject's range of vocabulary knowledge within the standardized administration framework for this test. An initial examination of construct-related validity and reliability in the PPVT-IIIA revealed that some items (i.e., items 74, 77, and 184) were less relevant to the instrument's construct. Second, on the basis of the results of the first step, we eliminated misfit items and analyzed a restricted range of the items (items 73 – 156; sets 7 – 13). The removal of the misfit items improved the Rasch measurement model.

The Rasch model placed item and person estimates on a common scale measured in logits, which determined the probability of a participant's correct answer on an item of the PPVT-

IIIA. The item-person map demonstrated that the range of item difficulty was more extended beyond the high and low ends of the measure, which resulted in the item logit values outside the range of person ability. This suggested a floor effect and a ceiling effect with the instrument not measuring lower and higher levels of ability properly, which could be regarded as an inadequacy of the test. A floor effect indicates that some items (the high end of the person-item map) are too difficult for the population to produce correct responses (that is, the questions on the instrument are extraordinarily difficult to measure a test-taker's true ability), while a ceiling effect (the low end of the person-item map) suggests that some items are too easy for the population to produce incorrect responses (that is, the questions on the instrument are insufficiently difficult to measure a test-taker's true ability). This implies that without some modifications, the PPVT-IIIA may not be an appropriate test, as a whole, for African-American adults who exhibit low literacy levels.

It is possible to speculate that the PPVT's broad aim[iii] of evaluating a wide range of ages and vocabulary skill levels played a part in this finding of extended item-difficulty variability beyond the person-ability distribution. As example, the test includes numerous low-frequency words to help address individual differences at the high end of the vocabulary scale (Pearson, Hiebert, & Kamil, 2007). In order to address this broad aim of spanning wide age ranges and skill levels, the PPVT uses a large number of items, but specifically defines basal and ceiling rules to save test administration time and reduce test-takers' frustration that might be caused by administering too many items that are either too easy or too difficult for him/her, but, at the same time, have a highly probable estimate if the entire test is given. In addition, the PPVT manual suggests beginning with lower items for test-takers for which there are suspicions of not performing at their chronological ages. In this respect, the standard procedure was followed in this study. The question raised by this standard administration procedure and these results is whether the range of item difficulty found in the Rasch model was restricted due to the floor and ceiling administration guidelines of the measure. In the case of struggling adult readers, should one give all the items on the PPVT without considering the floor and basal administration guidelines to obtain a more valid index of their actual vocabulary ability? Clearly one such limitation of this approach is the number of items that would need to be given, and the subject's potential frustration on those higher level items which they may not know. At the same time, another option would have been to administer the PPVT using the participant's actual age levels (Set 13; ages 17-adult) and use the standard basal and ceiling rules to administer the test. In this situation, the results may be similar to those found given that the basal rule requires backward item administration until the basal is established. Due to the Rasch model's conjoint measurement of persons and abilities, the items that were too easy and difficult indicate less precision of measurement at either end of the scale for this population. Given their unique and heterogeneous linguistic and language histories, this might not be surprising.

When it came to each item, some items were overfits or underfits. These items did not contribute to the unidimensionality in the PPVT-IIIA for the given population, indicating that the responses to these items were not aligned well with the abilities of the participants. As these data did not support the assumption that the structure of the PPVT-IIIA was reliably represented by the unidimensional trait, further investigations on these items are needed. The person and item reliability indices were also calculated to ensure consistency using reliability coefficients, which were related to the number of statistically different performance strata that the test identified in the sample. The results of the first analysis showed the presence of unmodeled high variability in the responses to the PPVT-IIIA and that the measure was unable to define the hierarchy of persons along the measured construct.

[iii] An anonymous reviewer raised this important point, of which we are appreciative.

After the diagnostics of the scales, a procedure of misfitting item deletion took place. The elimination of doubtful items resulted in the pattern of item difficulties to be more consistent with the model expectancy. In other words, the fit statistics and separation index not only fell within the acceptable ranges, but reliability and residual values also got better.

By and large, both gender and ability DIF demonstrated no remarkable differential functioning. However, a microscopic examination through DIF revealed that some items behaved in a slightly different way when the sample was broken down into two subgroups. It was unclear as to whether these small psychometric inequalities between the subgroups stemmed from a measurement bias, true difference, chance, or randomness of measurement. However, special attention needs to be placed on these items, especially when judgments and important inferences about test-takers' verbal abilities, and/or decisions regarding qualifications for special services are made based on the PPVT-IIIA scores (Carvajal, Nowark, & Fraas, 2000; Dollaghan & Campbell, 2009; Gerde & Powell, 2009; Majsterek & Lord, 1991).

Overall, the property of the Rasch model supported the comparison of person ability and item difficulty estimates, showing independence of the distribution of those abilities and difficulties in the subgroups of ability and gender, except for a very small segment of the item pool. With a modification, splitting the sample into two subgroups to calculate item difficulty estimates yielded invariant item estimates within the limits of measurement error, indicating that the PPVT-IIIA items were, in general, functioning equivalently across ability and gender categories through good item selection. As Bond and Fox (2007) note, the stability of item and person parameters across subgroups (i.e., measurement invariance) is a critical property of scientific measurement because test items should not behave differently to the particular subgroup. The results did not deviate from the invariance principle, after the misfit items were removed from analyses.

To summarize, an identification of misfit items for a particular population is important to detect item bias. A single item can carry particular cultural, linguistic, and/or social bias to a certain group. A real issue in measurement is that if test score differences are due to item biases rather than true differences, the test is considered to be unfair (Bond & Fox, 2007). Test bias reflects psychometric inequalities across groups. This study identified particularly misfit items for the African-American struggling readers. This is important to note. The elimination of the misfit items yields significant improvement of unidimensionality. Bond and Fox (2007) also suggest that if the principle is "not instantiated in practice, we should be motivated to examine the reason for that inadequate item [sic] and avoid using any such item in the measure in its current form" (p. 70). In this regard, as Messick (1995) states, "… the validity is an evolving property and validation is a continuing process" (p.741).

The scrutiny of the item's accuracy and precision was addressed by identifying the questionable items which degraded the measurement system or distorted the unidimensiionality. These results provide scientific evidence for potentially biased items which require particular consideration for a conceptual construct of the PPVT-IIIA for African-American struggling adult readers. These items call for special attention especially when important judgment about the test-taker is made based on the test score. When the PPVT scores are uniformly used across individuals with different backgrounds, sensitivity to and awareness of the measurement complexities should be emphasized. The results of this study ask clinicians and researchers to be cognizant of the implications of the test scores when they use the PPVT scores for diagnostic or research purposes with struggling adult readers or when important inferences are to be made about test-takers.

The findings of this study contain theoretical, research, and clinical implications. Theoretically, the results offer an insight into similarities and differences between adults and children with respect to latent constructs of receptive vocabulary skills, given the results of previous research studies with children (Campbell, Bell, & Keith, 2001; Miller & Lee, 1993; Restrepo et al., 2006). Except for a few items that behaved against the latent model, the majority of the items clustered together along the unidimensional construct of the Rasch model. From a research point of view, this study provides psychometric evidence of possibly biased items for this particular group. Further research can determine whether certain items are more susceptible to a measurement bias for African-American adults who have difficulty reading. In addition, this study suggests that a more narrowly targeted or adaptive test for this particular population[iv] might provide more precise information about vocabulary knowledge for struggling adult readers. Further research is necessary to address methodological and theoretical issues in the development of such a test. Clinically, the results of this study provide evidence that the psychometric properties of some items are questionable when the PPVT-IIIA is used for this particular group for the purpose of clinical assessment of receptive vocabulary and its score interpretation.

Although this study examined the utility, feasibility, and psychometric property of the PPVT-IIIA measure, the source of threats that a few items may have to the validity of the instrument tool is still unknown. Further research studies are needed to substantiate these findings and evaluate the complexity of the PPVT-IIIA instrument. Especially, the items identified as misfits are subject to further investigation.

Some limitations of this study need to be noted, which are also linked to future directions. First, the participants were restricted to African-American adults who read isolated words at the third- and fifth-grade level. This restriction might have resulted in skewness or deviations from the expected model. A further examination of goodness-of-fit of the PPVT-IIIA to the measurement model in appropriately targeted samples is recommended to confirm the findings of this study. A direct comparison to other populations, such as children, expert adult readers, and other ethnic groups, will offer valuable information about the instrument. Second, since an identification of misfitting items that disrupted the unidimensional construct of the PPVT-IIIA was the aim of this study, further Rasch analyses after eliminating misfit items was limited in this study. A full analysis of the item characteristics of the misfits would be beneficial. At the same time, a comparison of the good-fit and misfit items with respect to item properties would provide a deeper understanding of the instrument. Finally, the data were cross-sectional; therefore, the precision and accuracy of a one-time administration may be questioned because the test-takers might have exhibited a transitory response pattern, due to the test-taker's nervousness, not feeling well at the time of testing, chance or random guess, or other examinee idiosyncrasies, rather than a true indicator of a latent construct. A longitudinal analysis would corroborate the results of this study. Since the PPVT-III is one of the widely used instruments in the U.S. for children and adults, additional studies that test the utility of the PPVT-III are in need.

A possible subsequent study could include an evaluation of both Form A and Form B of the PPVT-III at the same time through equating and linking methods, such as parallel equating, common-person equating, or concurrent equating. An adaptation or modification of the PPVT as a spoken, receptive, vocabulary measure for the population of English language learners is also recommended. Nonetheless, the merit of this study entails a robust contribution to the field because there has been no study that examined this well-known

[iv]This is also a point raised by an anonymous reviewer.

instrument with respect to Rasch modeling, DTF, and DIF with a sample of African-American adults who are struggling readers.

Acknowledgments

This study was supported by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, the National Institute for Literacy, and the U.S. Department of Education – grant # R01 HD43801.

We thank Mike J. Linacre for valuable suggestions and two anonymous reviewers for constructive feedback on the earlier version of this paper.

References

Bell NL, Lassiter KS, Matthews TD, Hutchinson MB. Comparison of the Peabody Picture Vocabulary Test-third edition and Wechsler Adult Intelligence Scale-third edition with university students. Journal of Clinical Psychology. 2001; 57(3):417–422. [PubMed: 11241372]

Bond, TG.; Fox, CM. Applying the Rasch model: Fundamental measurement in the human sciences. 2. Mahwah, NJ: Lawrence Erlbaum Associates; 2007.

Bundgaard-Nielsen RL, Best CT, Tyler MD. Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese vowel categories. Applied Psycholinguistics. 2011; 32:51–67.

Byrne ME, Crowe TA, Hale ST, Meek EE, Epps D. Metalinguistic and pragmatic abilities of participants in adult literacy programs. Journal of Communication Disorders. 1996; 29:37–49. [PubMed: 8722528]

Campbell JM, Bell SK, Keith LK. Concurrent validity of the Peabody Picture Vocabulary Test-third edition as an intelligence and achievement screener for low SES African American children. Assessment. 2001; 8:85–94. [PubMed: 11310729]

Carvajal H, Nowark SJ, Fraas AC. Saving time: Using the Peabody Picture Vocabulary Test-III as a screening test of intelligence with undergraduates. College Student Journal. 2000; 34(2):281–283.

Cohen AS, Kim SH, Baker FB. Detection of differential item functioning in the graded response model. Applied Psychological Measurement. 1993; 17:335–350.

Dollaghan CA, Campbell TF. How well do poor language scores at ages 3 and 4 predict poor language scores at age 6? International Journal of Speech-Language Pathology. 2009; 11(5):358–365.

Dunn, LM.; Dunn, LM. The Peabody Picture Vocabulary Test. 3. Circle Pines, MN: American Guidance Service; 1997.

Gallagher A, Bridgeman B, Cahalan C. The effect of computer-based tests on racial-ethnic and gender groups. Journal of Educational Measurement. 2002; 39(2):133–147.

Gerde HK, Powell DR. Teacher education, book-reading practice, and children's language growth across one year of Head Start. Early Education and Development. 2009; 20(2):211–237.

Greenberg D, Wise J, Morris RD, Fredrick L, Rodrigo V, Nanda A, Pae HK. A randomized-control study of instructional approaches for struggling adult readers. Journal of Research on Educational Effectiveness. 2011; 4(2):101–117.

Hart DL, Wright BD. Development of an index of physical functional health status in rehabilitation. Arch Phys Med Rehabil. 2002; 83:655–665. [PubMed: 11994805]

Isaacs T, Trofimovich P. Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. Applied Psycholinguistics. 2011; 32(1):113–140.

Kunnan AJ. Test fairness and Toulmin's argument structure. Language Testing. 2010; 27(2):183–189.

Lin L-F. Gender differences in L2 comprehension and vocabulary learning in the video-based CALL program. Journal of Language Teaching and Research. 2011; 2(2):295–301.

Linacre, JM. A user's guide to Winsteps. 2010a. http://www.winsteps.com/winman/index.htm?guide.htm

Linacre, JM. Winsteps (Version 3.70.02). Chicago: Winsteps.com; 2010b. [Computer Software]

Linacre, JM. personal written correspondences. 2010c.

Lumley T, O'Sullivan B. The effect of test-taker, audience and topic on task performance in tape-mediated assessment of speaking. Language Testing. 2005; 22(4):415–437.

Majsterek DJ, Lord EN. An evaluation of the PPVT-R and UMI for screening preschoolers who are at risk for reading disabilities. Assessment for Effective Intervention. 1991; 16:173–179.

McLaren KP, Richards HC. Adaptive behavior scale cognitive triad: Discrimination and classification of institutionalized mentally retarded adults. American Journal of Mental Deficiency. 1986; 91:304–307. [PubMed: 3799738]

Messick S. Validity of psychological assessment. American Psychologist. 1995; 50(9):74–149.

Miller LT, Lee CJ. Construct validation of the Peabody Picture Vocabulary Test-Revised: A structural equation model of the acquisition order of words. Psychological Assessment. 1993; 5(4):438–441.

Millett J, Atwill K, Blanchard J, Gorin J. The validity of receptive and expressive vocabulary measures with Spanish-speaking kindergarteners learning English. Reading Psychology. 2008; 29:534–551.

Pae HK, Sevcik RA, Morris RD. Cross-language correlates in phonological process and naming speed: Evidence from deep and shallow orthographies. The Journal of Research in Reading. 2010; 33(4): 335–436.

Pearson PD, Hiebert EH, Kamil ML. Vocabulary assessment: What we know and what we need to learn. Reading Research Quarterly. 2007; 42(2):282–296.

Qingquan N, Chatupote M, Teo A. A deep look into learning strategy use by successful and unsuccessful students in the Chinese EFL learning context. REK Journal. 2008; 39(3):338–358.

Quiroz BG, Snow CE, Zhao J. Vocabulary skills of Spanish English bilinguals: impact of mother child language interactions and home language and literacy support. International Journal of Bilingualism. 2010; 14(4):379–399.

Reckase MD. Unifactor latent trait models applied to multi-factor tests: Results and implications. Journal of Educational Statistics. 1979; 4:207–230.

Restrepo MA, Schwanenflugel PJ, Blake J, Neuharth-Pritchett S, Cramer SE, Ruston HP. Performance on the PPVT-III and the EVT: Applicability of the measures with African American and European American preschool children. Language, Speech, and Hearing Services in Schools. 2006; 37:17–27.

Ryan K, Bachman LF. Differential item functioning on two tests of EFL proficiency. Language Testing. 1992; 9:12–29.

Sabatini JP, Sawaki Y, Shore J, Scarborough H. Relationships among reading skills of low-literate adults. Journal of Learning Disabilities. 2010; 43:122–138. [PubMed: 20179307]

Schumacker, RE. Rasch measurement: The dichotomous model. In: Smith, EV.; Smith, RM., editors. Introduction to Rasch measurement: *Theory, models, and applications*. Maple Grove: JAM Press; 2004. p. 226-253.

Smith EV. Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. Journal of Applied Measurement. 2001; 2:281–311. [PubMed: 12011511]

Stockman IJ. The new Peabody picture vocabulary test-III: An illusion of unbiased assessment? Language, Speech, and Hearing Services in Schools. 2000; 31:340–353.

Torgesen JK, Alexander A, Wagner R, Rashotte C, Voeller K, Conway T. Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. Journal of Learning Disabilities. 2001; 34:33–58. [PubMed: 15497271]

Vaughn MG, Beaver KM, Wexler J, DeLisi M, Roberts GJ. The effect of school dropout on verbal ability in adulthood: A propensity score matching approach. Journal of Youth Adolescence. 2011; 40:197206.

Vellutino F, Scanlon D, Sipay E, Small S, Pratt A, Chen R, Denckla M. Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. Journal of Educational Psychology. 1996; 88:601–638.

Walker MM, Givens GD, Cranford JL, Holbert D, Walker L. Auditory pattern recognition and brief tone discrimination of children with reading disorders. Journal of Communication Disorders. 2006; 39:442–455. [PubMed: 16487537]

Washington JA, Craig H. Performance of at-risk African-American preschoolers on the Peabody picture vocabulary test III. Language, Speech, and Hearing Services in Schools. 1999; 30:75–82.

Waugh RF, Addison PA. A Rasch measurement model analysis of the revised approaches to studying inventory. British Journal of Educational Psychology. 1998; 68:95–112. [PubMed: 9589625]

Webb M-YL, Cohen AS, Schwanenflugel PJ. Latent class analysis of differential item functioning on the Peabody picture vocabulary test-III. Educational and Psychological Measurement. 2007; 68(2): 335–351.

Williams, KT.; Wang, T. Vocabulary Test-third edition. Circle Pines, MN: American Guidance Service; 1997. Technical references to the Peabody Picture.

Woodcock, RW.; McGrew, KS.; Mather, N. Woodcock-Johnson III test of achievement. Itasca, IL: Riverside Publishing; 2001.

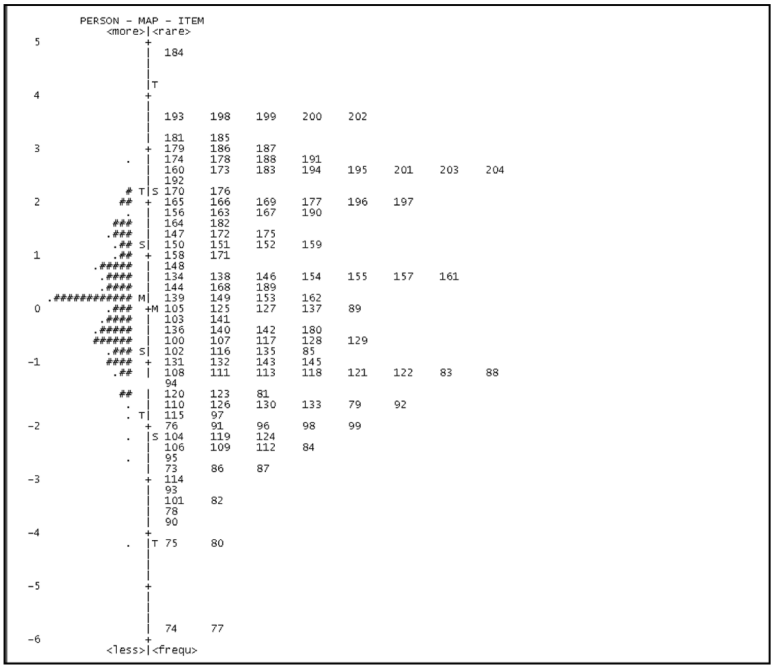

Figure 1.
Item-Person Map
Note: Each “#” represents 3 people and each “.” represents 1 person

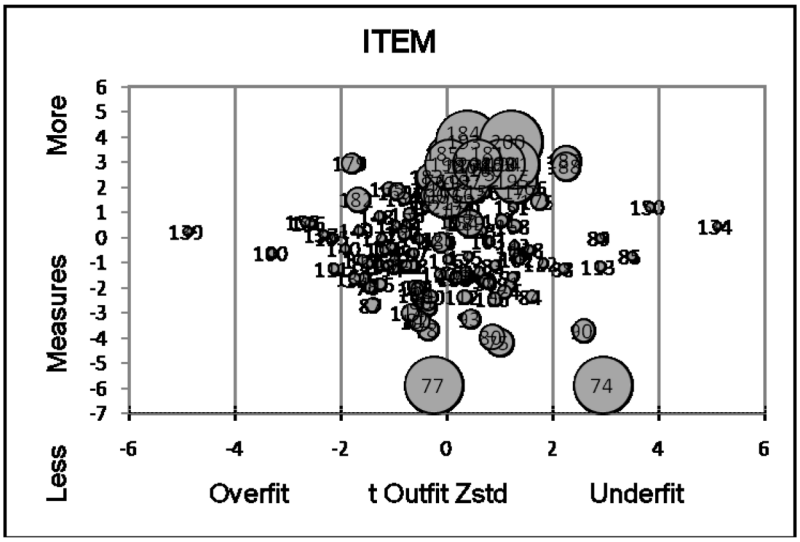

Figure 2.
Bubble Map

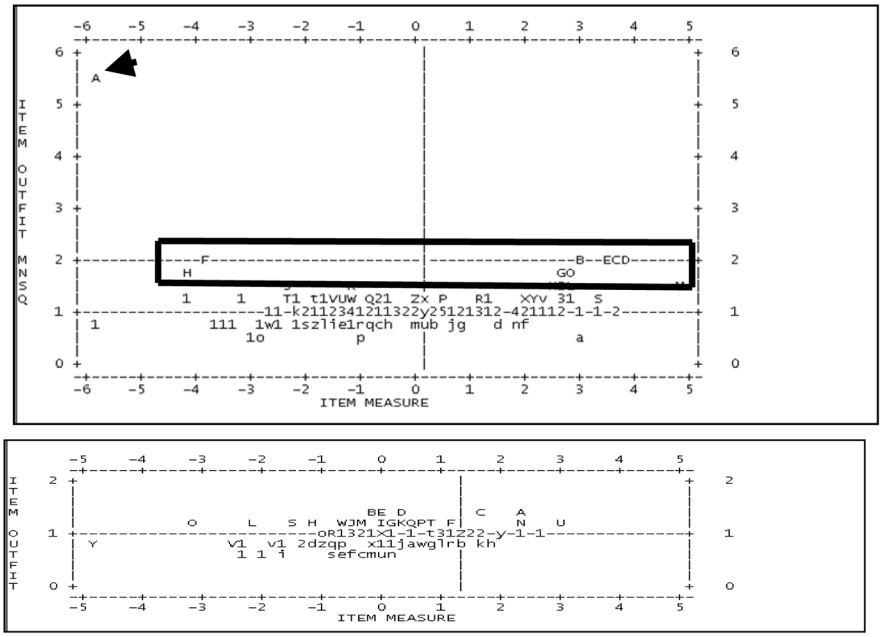

Figure 3.
Item Cross-Plot of Outfit vs. Measure

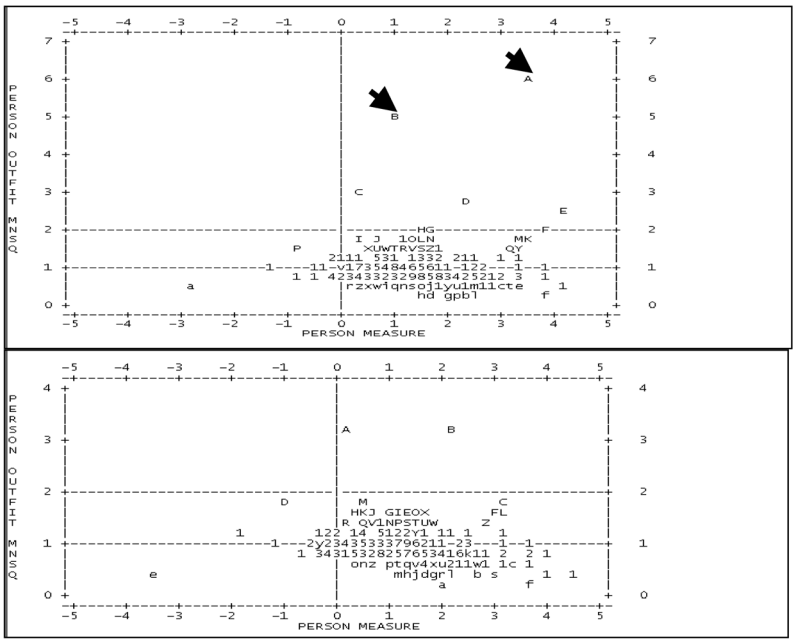

Figure 4.
Person Cross-Plot of Outfit vs. Measure

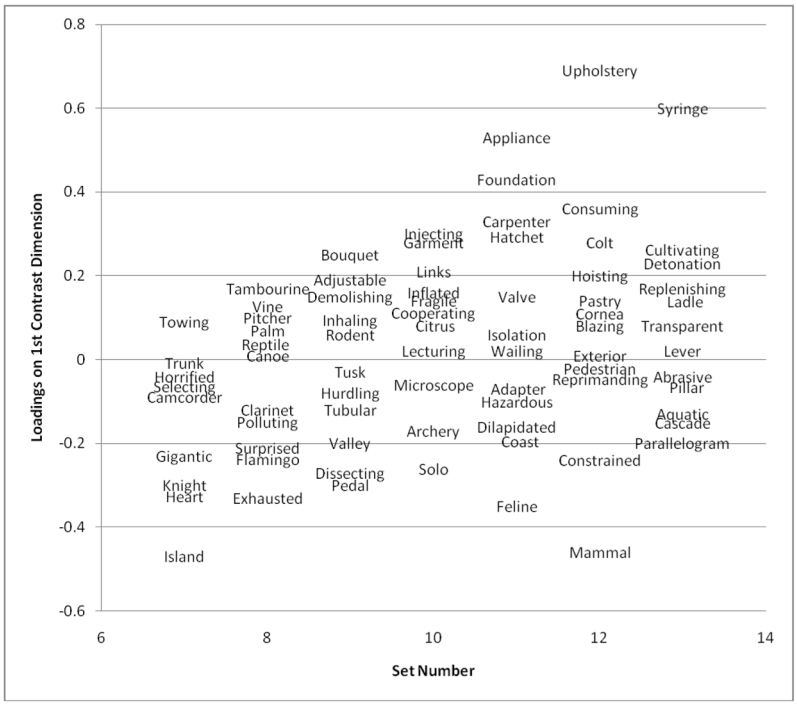

Figure 5.
Factor Loading of Each Item in the Set of the PPVT-III.

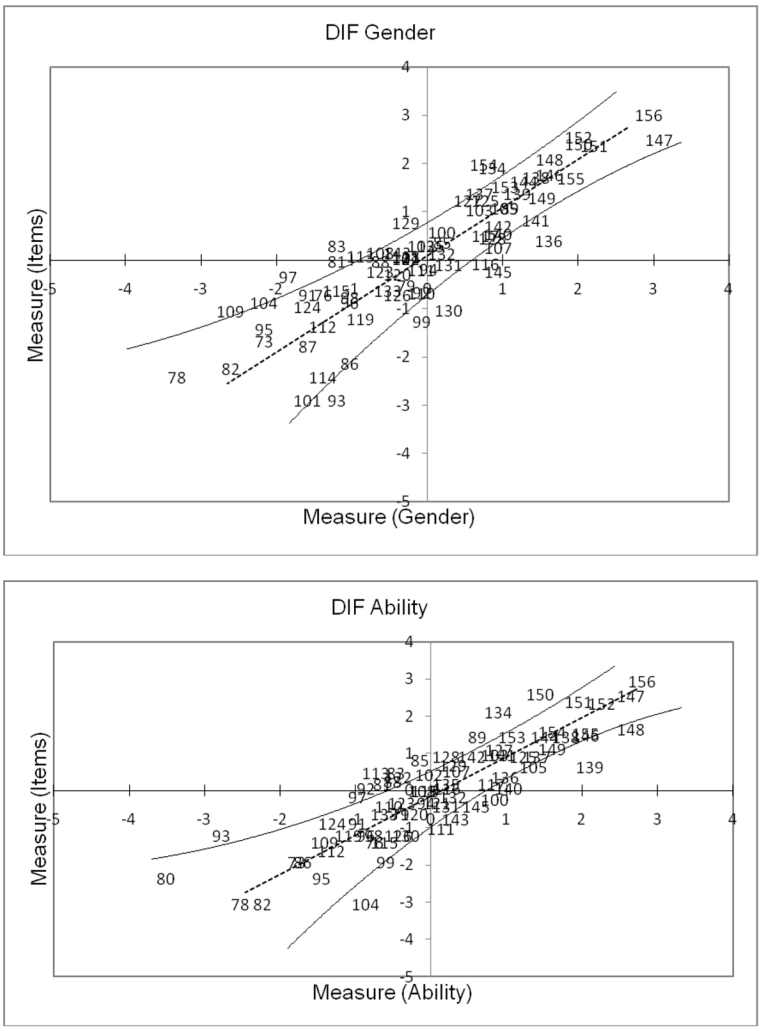

Figure 6.
Differential Test Function (DTF) by Ability and Gender

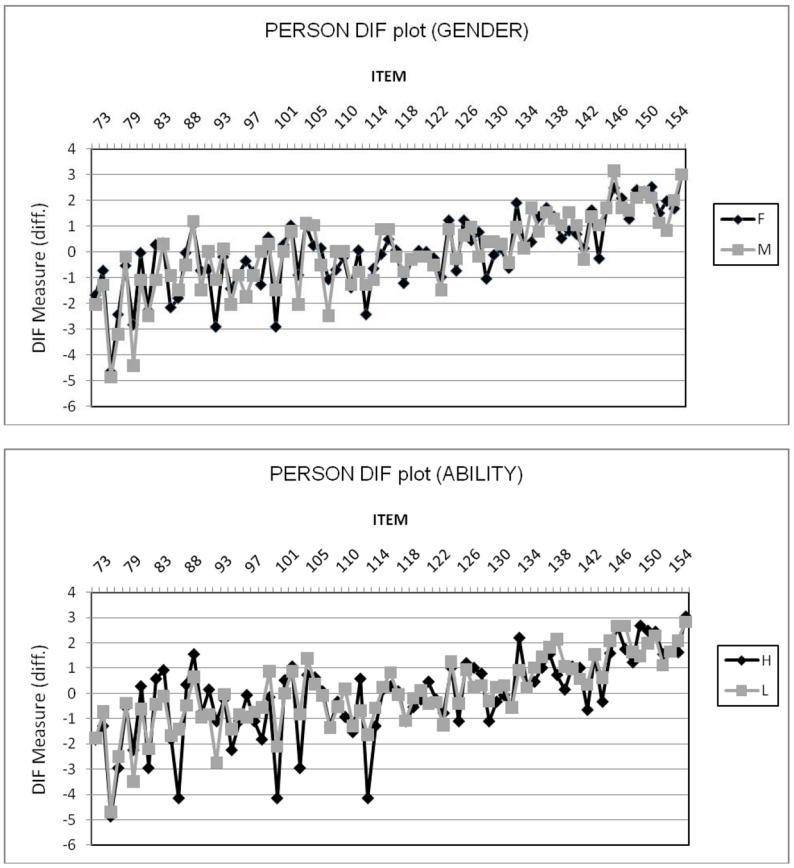

Figure 7.
Differential Item Functioning by Ability and Gender

Table 1

Misfit Order and Fit Statistics

Item	Stimuli Word	Measure	Infit		Outfit		PT-Measure Corr	
			MNSQ	ztsd	MNSQ	Ztsd	Corr	Exp. Corr
74	Nostril	−5.87	1.06	.4	5.46	3.0	−.−8	.09
90	Interviewing	−2.67	1.07	2.5	3.02	.14	−.01	.14
84	Wrench	−1.14	1.13	.8	2.00	2.8	.13	.29
106	Adjustable	−1.26	1.00	.01	1.72	2.0	.23	.25
75	Vase	−3.01	1.16	.5	1.5	.9	.03	.17