TPM®

COMPARING THE PERFORMANCE OF SYNONYM AND ANTONYM TESTS IN MEASURING VERBAL ABILITIES

WAHYU WIDHIARSO
HARYANTA
GADJAH MADA UNIVERSITY

This study investigates whether synonym and antonym tests measure similar domains of verbal abilities and have comparable psychometric performance. The data used in this study are subsets of the data collected during 2013-2014 graduate admission testing at Gadjah Mada University (UGM), using three forms of the Potensi Akademik Pascasarjana (PAPS) [Graduate Academic Aptitude Test]. Confirmatory factor analysis revealed that synonym and antonym tests assess similar domains of verbal abilities. A model integrating items from both tests to represent a single dimension better explained the data than a model separating the two tests into manifestations of different dimensions. High correlations among dimensions in unidimensional model showed interrelatedness for the domains of verbal abilities such as verbal knowledge, comprehension, and reasoning. Additional analysis using item-level analysis showed that antonym items tended to be more difficult than synonym items. This finding indicates that, although both tests assess similar content, responding to an antonym test requires more complex cognitive process than responding to a synonym test.

Key words: Synonyms; Antonyms; Verbal abilities; College admission test; Rasch analysis.

Correspondence concerning this article should be addressed to Wahyu Widhiarso, Faculty of Psychology, Gadjah Mada University, Jl. Humaniora, No.1, 550436 Yogykarta, Indonesia. Email: wahyu_psy@ugm.ac.od

College admission tests change rapidly, with test developers frequently modifying the test design, scoring procedure, or even content to better represent the construct being measured. Within the last ten years, the most important test in this field — the Scholastic Aptitude Test (SAT) — has undergone many changes in terms of both content and composition of test sections (Zwick, 2004). Another test — the Graduate Record Examinations (GRE) — has also changed its content, adopted a new design, and established new score scales for several subtests (Educational Testing Service, 2015). Each university admission test is unique due to a wide range of factors, such as content domains, specification, and type of task or items. Previous studies suggested that, for example, verbal reasoning explains a significant portion of aptitude test score variance in science but not in economics (Schult, Fischer, & Hell, 2015). For this reason, some universities could emphasize quantitative domains, because scholars in scientific activity frequently work with numbers. Conversely, other universities could emphasize verbal domains because most of their students' activities primarily relate to primarily use of everyday language. However, skills related to verbal ability are required for work in all disciplines, as learning requires listening and reading, as well as demonstrating one's knowledge both verbally and in writing (Burton, Welsh, Kostin, & van Essen, 2009).

Verbal ability is an important dimension in human intelligence. According to Cattell-Horn-Cattell's (CHC) model of general intelligence, factor "g" is composed of crystallized intelligence as reflected in verbal abilities and fluid intelligence as reflected through nonverbal abilities (Reynolds & Kamphaus, 2003). In the CHC model, verbal ability is included in crystallized intelligence (Gc), defined as individual capacity to store verbal or language-based declarative (knowing "what") and procedural (knowing "how") knowledge, acquired through the "investment" in other abilities during formal and informal educational and general life experiences (McGrew, 2005). Verbal ability is also supposed to be associated with fluid reasoning (Gf) which is defined as a facility of reasoning, particularly where adaptation to new situations is required and crystallized learning assemblies are of little use (Wasserman & Tulsky, 2005). In this case, verbal is just one of several contents (e.g., numerical, spatial, abstract, and mechanical) used to perform common cognitive tasks, but the cognitive functions which play on that performance involve perception, memory, and reasoning.

Previous research showed that verbal abilities have a strong relationship with several activities of academic achievement such as reading, writing, and mathematics (Rindermann, Michou, & Thompson, 2011; Walker, Greenwood, Hart, & Carta, 1994). Because verbal abilities are so highly valued in academic and intellectual environments (Izard et al., 2001; Petrides, Chamorro-Premuzic, Frederickson, & Furnham, 2005), most scholastic aptitude tests are composed of several subtests for assessing various domains of verbal abilities. In the university admission setting, these tests assess capacity to analyze relationships among component parts of sentences and to recognize relationships among words and concepts (Educational Testing Service, 2005). Most well-known scholastic aptitude tests for admission assess various dimensions, including verbal abilities. For example, the Cognitive Abilities Test (CogAT), which purports to assess reasoning abilities, assesses verbal abilities in addition to quantitative and figural domains. Commonly used measurement domains of verbal abilities are verbal comprehension, verbal reasoning, and verbal fluency (Janssen, De Boeck, & Vander Steene, 1996).

Several cognitive tests have been proposed to measure these domains; they include synonyms, antonyms, analogies, classification, reading comprehension, and sentence completion. These tests were generally developed to measure relatively similar verbal ability domains so that the tests might be interchangeable. However, each of the tests is unique in how they assess specific verbal ability domains. For example, certain tests might measure verbal reasoning, yet at the same time they provide thorough and comprehensive information about an individual's vocabulary. Information about the effectiveness and efficiency of every such test is important for test constructors in order to develop test batteries composed of multiple tests to assess specific domains.

Giving a correct synonym involves the generation of one or more synonym candidates and the (subsequent or simultaneous) evaluation of these generated words (or words being generated) on their degree of synonymy with the stimulus word. This componential model does not specify the temporal organization of these two components in the solution process of the total task. Generating and evaluating synonym candidates can be sequential or parallel cognitive processes. The open synonym task was decomposed into a generation and an evaluation component.

The use of synonym- and antonym-based tests to measure cognitive abilities, particularly in the verbal ability cluster, is still controversial. Many popular tests are composed of subtests including synonyms and antonyms to measure individual attributes associated with verbal abilities. The Concept Mastery Test (Form T) employs synonym and antonym subtests because the test

constructors believe that the level of vocabulary, general knowledge, and ability to make inferences of the sort required by verbal analogy items represent different levels of concept mastery (Grigorenko, Sternberg, & Ehrman, 2000). The Woodcock-Johnson III test of Cognitive Abilities (WJ III COG) also employs both synonym- and antonym-based subtests to measure verbal comprehension in the verbal ability cluster (Schrank, 2005). Other tests, such as the Armed Services Vocational Aptitude Battery (ASVAB; Wolfe & Held, 2010) and the Brazilian Adult Intelligence Battery (BAIAD; Wechsler et al., 2014), also use synonym- and antonym-based subtests. Synonym- and antonym-based tests are also often used in clinical settings for purposes such as assessing deficits in complex language functions (Copland, Chenery, & Murdoch, 2000). A number of tests employ only one of these subtests (synonyms or antonyms); one example is the old version of the SAT, which only employed antonym items. This type of items is no longer included in the current version of the SAT, the logic being that these items present words without any context and simply encourage rote memorization (Lawrence, Rigol, Essen, & Jackson, 2004).

Controversy is still strong regarding whether synonym and antonym tests reflect overall skills or simply the information that individuals have learned. This is an important distinction, as aptitude tests for student admission should predict success in future studying, and should therefore not simply measure prior learned knowledge. Synonym and antonym tests are closely associated with vocabulary tests, which are mostly used to assess word familiarity in language fluency. However, several scholars (e.g., Sincoff & Sternberg, 1987) suggest that antonyms items, as used in verbal scholastic aptitude tests, are more difficult than vocabulary items, because antonym items emphasize precision and reasoning more than familiarity. We support this notion and propose that synonym and antonym tests both assess some degree of verbal comprehension and reasoning.

Another concern, besides what synonym and antonym tests actually assess, is their psychometric performance. Despite the fact that these are both very common types of tests, few studies have been conducted to examine them. There are three possible reasons why both tests may be employed concurrently. First, if both tests assess a similar content domain of verbal ability, test constructors employ both of them because one test might be more difficult than the other, and they want to develop tests that consist of items representing different levels of difficulty (see Wilson, 2005). In this context, synonym and antonym tests are assumed to represent complementary methods. Second, if test constructors want to apply differential weighting to different dimension of verbal abilities, the dimension that is weighted more heavily should be represented by more items or facets. Different dimensions may be differently weighted when the measurement domain is exceptionally broad, when there is a specific measurement objective, or when there is a diverse population of test takers. For example, certain college admission tests might give more weight to verbal comprehension. If so, the synonym and antonym tests would be assumed as tests that measure different dimensions. Third, if synonym and antonym tests assess different domains of verbal knowledge, then both tests are assumed to be different methods. However, employing two tests that assess similar domains can be inefficient due to inherent redundancy.

Synonym and antonym tests are different from other verbal reasoning tests because they emphasize the mastery of vocabulary words. Both of these tests are usually presented as multiple choice items where test takers must choose which one of four or five alternative words has the most similar meaning (for synonym tests) or most opposite meaning (for antonym tests) to the stimulus word. Often, the stem word in question has some connection to all of the possible options; test takers must therefore thoroughly examine the root, suffixes, and prefixes to find the best option. Syno-

nyms and antonyms are not opposites; rather, the opposite of synonymity is heteronymy, while the opposite of antonymy is a lack of opposition (Hermann, Conti, Peters, Robbins, & Chaffin, 1979). Two words can have a direct or indirect antonymous relationship (Brown, Frishkoff, & Eskenazi, 2005). For example, *wet* and *dry* are direct antonyms, while *humid* and *arid* are indirect antonyms, stemming from the opposition between *wet* and *dry* (Gross & Miller, 1990). Success in solving such indirect antonym questions requires knowledge of the direct antonym of the words in question.

THE AIMS OF THIS STUDY

The present study examines whether synonym and antonym tests assess different measurement domains within verbal abilities. Prior studies (e.g., Ward, 1982) have assumed that synonym and antonym tests are identical methods for measuring verbal comprehension. However, we argue that synonym and antonym tests may measure different domains of verbal abilities because both tests have different emphases: specifically, although both tests assess vocabulary, antonym tests assess additional fluid skills, such as verbal reasoning and working memory, as this type of test requires more contextualized vocabulary knowledge.

In addition to the fact that synonym and antonym tests may assess different domains, the level of cognitive processes required to answer items on these tests may be different. Several studies have compared synonym and antonym item difficulty, but the results have been mixed (Sabouri, 1998). We posit that solving antonym items requires higher level of cognitive processes than solving synonym items.

The present study also investigates whether the item difficulty for these types of tests is different. Solving antonyms requires both vocabulary knowledge and verbal reasoning to understand the context of the items in question. Knowing the embedded context of the word can help test takers to determine the meaning of the word in question; accordingly, antonym items are considered more difficult than synonym items (Medical College Admission Test, 1970).

The present study employed a dataset from the Gadjah Mada University Potensi Akademik Pascasarjana (PAPS) [Graduate Academic Aptitude Test], one of the main instruments for graduate admission assessment. The PAPS is a standardized test that measures several abilities representing a general intelligence. Individuals' scores on the verbal, quantitative, and analytic reasoning sections of the PAPS provide cognitive readiness to be successful in graduate school. PAPS was developed using Carroll's (1993) hierarchical organization of cognitive abilities, and consists of verbal ability, mathematical ability, and spatial ability as three content domains encircling general factor of intelligence.

METHOD

Participants and Procedure

The data used in this study are a subset of the data collected via the PAPS graduate admission test to evaluate potential students for a master program at the Indonesian Gadjah Mada University (UGM), during the years 2013-2014. The test was administered to 6357 applicants for

a broad range of disciplines programs (e.g., science, humanity). Three forms of PAPS (administered in the language of Bahasa Indonesia) were employed in this study: Form A (N = 3470), Form C (N = 1390), and Form D (N = 1497). We employed all existing datasets in the data bank of PAPS, so sampling procedures for selecting participants for this study were not applied. It is possible that overlap is present in samples: indeed, students might have taken the test more than once because their previous scores did not meet the requested criteria. We did not employ dataset of PAPS Form B because this form is usually used for different purposes. Approximately half of the total sample (n = 3051; 48%) were male and the remaining students (n = 3306; 52%) were female. All participants were Indonesian, had a bachelor's degree, and their age ranged from 22 to 38 years. Most of people in Indonesia have a strong interest in enrolling in the courses provided by UGM because this university is one of the biggest and oldest universities in the country. Therefore, the distribution of demographical backgrounds data (e.g., rural-urban, ethnicity, etc.) of this study represents national characteristics.

The administration of the PAPS subtests (verbal, quantitative, and logical reasoning) was carried out in 120 minutes, divided into three stages of 40 minutes each. During the first 20 minutes the examinees were allowed to work on the items of the subtest according to the instruction of the testers, namely, they were not allowed to work on items of another subtest. This applied for all the three subtests. The distribution of test books to the examinees was conducted randomly so that the examinees in the same test room received different forms of PAPS subtests.

Instrument

Each form of the PAPS (Azwar, Suhapti, Haryanta, & Widhiarso, 2009) test consists of three sections (verbal, quantitative, and logical reasoning). The verbal subtest is composed of four types of items: synonyms (12 items), antonyms (12 items), verbal analogies (10 items), and readings (six items). The present study examined only the synonym and antonym items, each of which had five possible answer options (only one correct). For the synonym items, participants were instructed to choose the word that had the meaning most similar to that of the target word. For the antonym items, participants were instructed to choose the word that had the closest opposite meaning to that of the target word (see Figure 1 for items examples).

BUILD	EXPOSED
a) start	a) open
b) invent	b) public
c) found	c) absent
d) develop	d) hidden
e) produce	e) suppressed

FIGURE 1
Example of synonym (left side) and antonym (right side) items.

Verbal analogies items were somewhat akin to Miller's Analogies test: one word pair was presented in item stem (e.g., DOG: ANIMALS = ?) and participants were instructed to choose one

word pair in the answer option that was similar to the word pair in the item stem. The total score was the total number of items answered correctly.

Data Analyses

Data were analyzed using Rasch analysis in Winsteps 3.7, with the goal of assessing and comparing item parameter of each form of PAPS. The Rasch analysis consisted of several analytical steps. The first step was checking the dimensionality of measurement because Rasch analysis assumes unidimensionality. A principal component analysis (PCA) as conducted by Winsteps 3.7 was employed for this purpose. PCA provides several dimensions regarding measurement but only the first dimension is the construct of interest. A variance of 60% or greater accounted for by the Rasch dimension is considered to be good (Linacre, 2012). Winsteps 3.7 provides eigenvalue contrast value between the first and second dimension. Small eigenvalue value contrast (below 2.0) between the first and second dimension indicates unidimensionality of measurement, meaning that any pattern in the differences within the residuals is not large enough to suggest that more than one dimension exists.

The second step was evaluating the model-data fit by examining infit and outfit statistics for each item. Infit and outfit statistics are mean square residuals between observed and expected responses, ranging from zero to positive infinity. Infit or outfit value close to 1.0 indicates an adequate fit. A value of infit and outfit statistics ranging from .8 to 1.2 is recommended for high-stakes testing (Bond & Fox, 2007), but values between .5 and 1.5 are still productive for measurement (Linacre, 2012). This study employed the second rule of thumb to define item fit to model.

RESULTS

Dimensionality Checking

A Rasch principal component analysis of the residuals of the instrument was performed. Residuals are parts of the measurement representing a random or a structure that was not explained by the Rasch analysis. We examined the dimensionality of the 24 items in the PAPS verbal subtest. The proportion of raw variances explained by the Rasch measures is 27.9%, 27.8%, and 35.4% for PAPS Form A, C, and D, respectively. Those values are lower than expected because the minimum variation that should be explained by the measures is 60% to prove unidimensionality of measurement. This finding suggests that additional factors explaining Rasch score variance in each individual may be considered. A closer examination of a standardized residual loading contrast suggested that there are one to three items that have a value above .4 in each PAPS form. According to Linacre (2012) values of ± .4 or higher are considered substantive. These items possibly explain great variance of other factors. However, the unexplained variance in first contrast in each PAPS form (in eigenvalue units) for those forms is 1.4 (5.4%), 1.4 (5.1%), and 1.6 (5.0%), respectively, suggesting that the unidimensionality holds because those values are below the cutting point (2.0 eigenvalue units).[1] An examination of standardized residual correlations suggests only one item pair (PAPS Form D). Rasch dimensionality analysis supports the hypothesis of unidimensionality and local independence of the synonym-antonym of PAPS items. Unidimensionality indicates that synonym and antonym items measure single ability,

and local independence indicates that individuals' possibility of answering correctly synonym items does not affect (or is not affected by) their answers to antonym items, and vice versa.

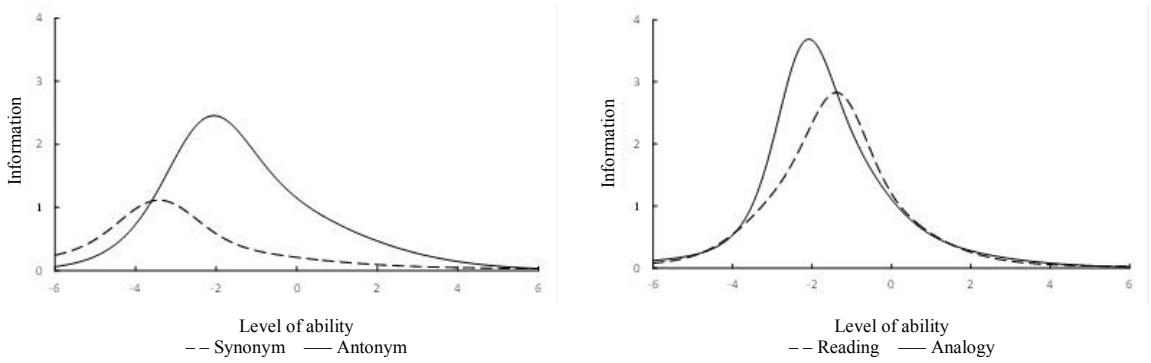
Correlations between Types of Items

Although the two dimensional models[2] better explained the data than the unidimensional model, the correlations among the four types of items that make up the verbal subtest of all three forms of PAPS are high. For example, in PAPS Form A there was a high correlation between the vocabulary items (a combination of synonym and antonym items) and the analogy items (r = .964), as well as between the vocabulary and the reading items (r = .880); this finding was consistent across all the three forms of PAPS. In PAPS Form C there was a very high correlation between the synonym and the antonym items (r = .998). This high correlation suggests that both synonym and antonym items assess a single content domain and share a high degree of method variance; that is, they are not unique methods for assessing verbal knowledge and comprehension. Moreover, high correlations among the four types of items in the verbal section of PAPS indicate high convergence, that is, sharing a high degree of similarity or even representing overlapping measurement domains in verbal abilities.

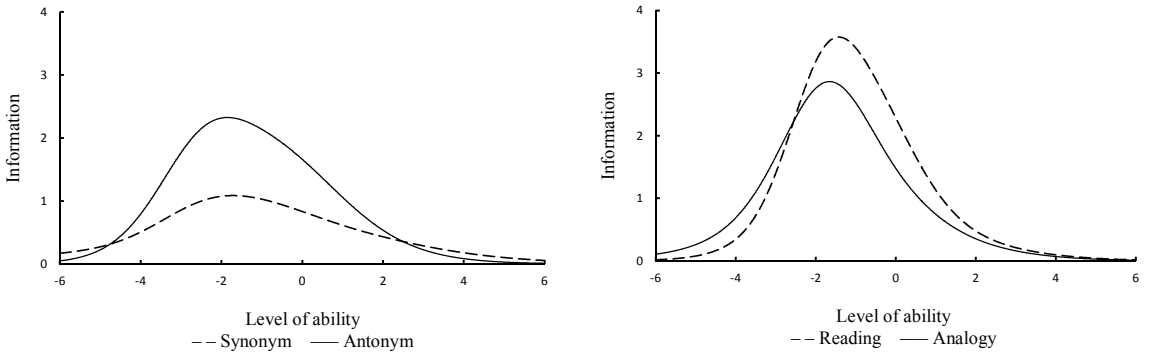
Comparing Item Parameters

A comparison of PAPS synonym and antonym items difficulty suggested that antonym items were more difficult than synonym items. For PAPS Form A, the mean item difficulty for synonym (M = .00) and antonym (M = .24) items were both located in the middle of the ability continuum. These results were consistent for Form C (synonym, M = –.53; antonym, M = .17) and for Form D (synonym, M = –1.02; antonym, M = –.72). However, although antonym items were found to be more difficult than synonym items, both types of items discriminated well between individuals with moderate levels of verbal abilities, with reported values within the range of –1 to 1. The Wald statistics for testing differences in mean item difficulty yielded significant results for Form A, $\chi^2(1)$ = 46.45, p < .001; Form C, $\chi^2(1)$ = 230.55, p < .001; Form D, $\chi^2(1)$ = 32.97, p < .001. These results suggest that individuals with low to high levels of ability (theta above –1) can solve synonym items, but only those with moderate to high levels of ability (theta above .25) can solve antonym items.

Figure 2 shows a comparison of the test information function for the synonym and antonym items across two forms of PAPS. The test information function is the sum of the information contributed by the items of a particular test. In Item Response Theory (IRT), each item of a test contributes information about individual ability. Items that discriminate well among test takers will contribute more item information. For example, for Form A, the maximum information obtained by the synonym items is 1.11, while the maximum information obtained by the antonym items is 2.45. The synonym items yield more information if they are administered to individuals with ability ranging from –3.5 to –3.2, while the antonym items yield more information if they are administered to individuals with ability ranging from –2.2 to –1.7. This finding is consistent with the other forms of PAPS, which both show antonym items as posing a higher level of difficulty than synonym items.



Form A



Form C

FIGURE 2
Item information function of synonym and antonym items.

DISCUSSION

This study found that synonym and antonym tests measure similar domains of cognitive abilities, although the level of difficulty differed for the two types of test items. Antonym items were more difficult and more discriminative than synonym items. High correlations between synonym and antonym tests and other types of tests can indicate that their measurement domains are relatively similar. This similarity may be due to the words used in synonym and antonym tests. The synonym and antonym items of PAPS are different, but share a common approach, employing uncommon, infrequent, abstract, or specialized words to assess an individual's capacity to use reasoning in diverse contexts. However, although using uncommon words in synonym or antonym tests is a useful technique to assess prior achievement, it is a poor technique to assess verbal comprehension or reasoning. Accordingly, as this technique exclusively assesses vocabulary knowledge or growth, it should be avoided in aptitude tests that aim to measure reasoning rather than domain-specific knowledge (Lohman, 2004). For this reason, the words used in the synonym and antonym items of PAPS are different from those usually used in other tests. Instead

of using difficult or obscure words, PAPS employs words that test takers would be likely to encounter in their day-to-day lives, such as *break*, *straightforward*, or *stop*. As a result, solving these items does not merely require verbal knowledge, but instead requires a combination of verbal comprehension and reasoning, meaning that the PAPS synonym and antonym items measure verbal comprehension more than verbal knowledge. The findings of the present study indicate that there is convergent validity for PAPS synonym and antonym items, and clarifies the relationship between verbal fluency and verbal comprehension abilities.

The convergence of synonym and antonym content domain coverage can have both advantages and disadvantages. One of the advantages of the fact that these tests assess similar domains is that both item types are therefore homogeneous, and can potentially achieve high internal consistency. Another advantage is the breadth of ability level covered by both item types, as synonyms tend to discriminate better among examinees with lower levels of verbal ability than antonyms do. The main disadvantage is redundancy, thus indicating inefficiency. In general, different types of items in a battery test should contain as little redundancy as possible; that is, all other things being equal, item types should correlate with each other as little as possible. (Powell, Bailey, & Clark, 1980). However, redundancy is unavoidable if the theory used to construct different tests is the same (Bacharach & Furr, 2007).

Comparing item difficulty indicated that the antonym items were more difficult than the synonym items. This was the expected finding, as antonym items — which require finding the opposite meaning — require more reasoning skills than synonym items. Correctly answering antonym items requires more verbal comprehension than answering synonym items, because synonymity refers to a pair of words with the same meaning that are interchangeable in a text, while antonymy refers to two words that may form either a conceptual identical pair (e.g., man-woman) or have the opposite meaning (e.g., good-bad) (Morimoto, 2015). Because antonym items have a broader range of possible answers, test takers must have good vocabulary, comprehension, and reasoning skills to solve them. Previous studies have found that solving antonym items requires a broader domain of cognitive skills than solving synonym items (Phillips, 2013), and that synonyms measure breadth of vocabulary while antonyms measure analogical thinking (Arjona-Tseng, 1993).

Another reason why antonyms tend to be harder to solve than synonyms is that the level of context clarity between synonym and antonym items is different. Antonym items provide more limited verbal contexts than synonym items. To solve antonym items correctly, individuals must have the ability to interpret the context of the word in question, which is often hidden in tests. Test takers with good verbal knowledge and analytical ability can expose this context to determine which word has the most opposite meaning to the word in question. Synonym items provide a clearer verbal context and more specificity than antonym items (Freedle & Kostin, 1990).

The findings of the present study should be interpreted with caution because they were only tested using the synonym and antonym items of PAPS, which might differ from the synonym and antonym items of other assessments. The finding that synonyms and antonyms both measure verbal knowledge, comprehension, and reasoning has several limitations. Although previous research supports this interpretation, this might not be applicable to the specific synonym and antonym items that appear in PAPS. Further study is required to determine whether the synonym and antonym items used in PAPS measure verbal comprehension and verbal reasoning. Even though several popular tests, such as the WJ III COG, indicate synonym-antonym subtests as an appropriate tool for measuring verbal comprehension, the administration of these tests is

different from that of PAPS. For example, in the WJ III COG, synonyms and antonyms are both the stimulus and the response given orally (Read & Schrank, 2003).

NOTES

1. A Rasch principal component analysis performed by Winsteps consists of five contrast tests between residuals and measurement variance. The first contrast shows how large the variances of secondary dimensions are. To consider a measure as unidimensional, the first contrast should explain at least the amount of variance of no more than two items.
2. In checking the dimensionality, two dimensions are, in some degree, appropriate to explain the model because the variation of the measure should be at least 60% of variance of the Rasch measurement.

REFERENCES

Arjona-Tseng, E. (1993). A psychometric approach to the selection of translation and interpreting students in Taiwan. *Perspectives*, *1*, 91-104. doi:10.1080/0907676x.1993.9961203

Azwar, S., Suhapti, R., Haryanta, & Widhiarso, W. (2009). Test Potensi Akademik Pascasarjana (A1) [The Graduate Academic Aptitude Test-A1].Yogyakarta, Indonesia: Universitas Gadjah Mada.

Bacharach, V. R., & Furr, R. M. (2007). *Psychometrics: An introduction*. Thousand Oaks, CA: SAGE Publications.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005, October). *Automatic question generation for vocabulary assessment.* Paper presented at the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, Canada.

Burton, N. W., Welsh, C., Kostin, I., & van Essen, T. (2009). Toward a definition of verbal reasoning in higher education. *ETS Research Report Series*, *ETS RR-09-33*.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York, NY: Cambridge University.

Copland, D. A., Chenery, H. J., & Murdoch, B. E. (2000). Persistent deficits in complex language function following dominant nonthalamic subcortical lesions. *Journal of Medical Speech-Language Pathology*, *8*, 1-14.

Educational Testing Service. (2005). *Guide to the use of scores*. Princeton, NJ: Author.

Educational Testing Service. (2015). *GRE® information and registration bulletin*. Princeton, NJ: Author.

Freedle, R., & Kostin, I. (1990). Item difficulty of four verbal item types and an index of differential item functioning for black and white examinees. *Journal of Educational Measurement*, *27*, 329-343. doi:10.2307/1434853

Grigorenko, E. L., Sternberg, R. J., & Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability: The canal-F theory and test. *The Modern Language Journal*, *84*, 390-405. doi:10.2307/330568

Gross, D., & Miller, K. J. (1990). Adjectives in WordNet. *International Journal of Lexicography*, *3*, 265-277.

Hermann, D. J., Conti, G., Peters, D., Robbins, P. H., & Chaffin, R. J. (1979). Comprehension of antonymy and the generality of categorization models. *Journal of Experimental Psychology: Human learning and memory*, *5*, 585-597. doi:10.1037/0278-7393.5.6.585

Izard, C., Fine, S., Schultz, D., Mostow, A., Ackerman, B., & Youngstrom, E. (2001). Emotion knowledge as a predictor of social behavior and academic competence in children at risk. *Psychological Science*, *12*, 18-23. doi:10.1111/1467-9280.00304

Janssen, R., De Boeck, P., & Vander Steene, G. (1996). Verbal fluency and verbal comprehension abilities in synonym tasks. *Intelligence*, *22*, 291-310. doi:10.1016/S0160-2896(96)90024-0

Lawrence, I. M., Rigol, G., Essen, T. V., & Jackson, C. (2004). A historical perspective on the content of the SAT. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 57-74). New York, NY: Routledge.

Linacre, J. M. (2012). *A user's guide to WINSTEPS and ministep Rasch model computer programs: Program manual 3.75*. Chicago, IL: Winstep.com.

Lohman, D. F. (2004). Aptitude for college: The importance of reasoning tests for minority admissions. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions*. New York, NY: Routledge.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities. Past, present and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 136-181). New York, NY: Guilford Press.

Medical College Admission Test. (1970). *Preparation for medical college admission test*. New York, NY: Cowles Book Company.

Morimoto, Y. (2015). *Word meaning relationship extraction device.* Tokyo, Japan: Hithaci Ltd.

Petrides, K. V., Chamorro-Premuzic, T., Frederickson, N., & Furnham, A. (2005). Explaining individual differences in scholastic behaviour and achievement. *British Journal of Educational Psychology*, *75*, 239-255. doi:10.1348/000709904x24735

Phillips, C. I. (2013). *Children's understanding of antonymy* (Doctoral thesis, University of Calgary, Alberta, Canada). Retrieved from http://hdl.handle.net/11023/963

Powell, G. E., Bailey, S., & Clark, E. (1980). A very short version of the Minnesota Aphasia Test. *British Journal of Social and Clinical Psychology*, *19*, 189-194. doi:10.1111/j.2044-8260.1980.tb00947.x

Read, B. G., & Schrank, F. A. (2003). Qualitative analysis of Woodcock-Johnson III Test performance. In F. A. Schrank & D. P. Flanagan (Eds.), *WJ III clinical use and interpretation* (pp. 47-91). Boston, MA: Academic Press.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales*. Lutz, FL: Psychological Assessment Resources, Inc.

Rindermann, H., Michou, C. D., & Thompson, J. (2011). Children's writing ability: Effects of parent's education, mental speed and intelligence. *Learning and Individual Differences*, *21*, 562-568. doi:10.1016/j.lindif.2011.07.010

Sabouri, L. L. (1998). *The interaction of suffixation with synonymy and antonymy* (Master's thesis, University of Alberta, Edmonton, Canada). Retrieved from http://www.collectionscanada.gc.ca/obj/s4/f2/dsk2/tape15/PQDD_0008/MQ34411.pdf

Schrank, F. A. (2005). Woodcock-Johnson III tests of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 371-401). New York, NY: Guilford Press.

Schult, J., Fischer, F. T., & Hell, B. (2015). Tests of scholastic aptitude cover reasoning facets sufficiently. *European Journal of Psychological Assessment*. Advance online publication. doi:10.1027/1015-5759/a000247

Sincoff, J. B., & Sternberg, R. J. (1987). Two faces of verbal ability. *Intelligence*, *11*, 263-276. doi:10.1016/0160-2896(87)90010-9

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of school outcomes based on early language production and socioeconomic factors. *Child Development*, *65*, 606-621. doi:10.2307/1131404

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement*, *6*, 1-11. doi:10.1177/014662168200600101

Wasserman, J. D., & Tulsky, D. S. (2005). A history of intelligence assessment. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 3-22). New York, NY: Guilford Press.

Wechsler, S. M., Vendramini, C. M. M., Schelini, P. W., Lourençoni, M. A., Ferreira de Souza, A. A., & Mundim, M. C. B. (2014). Factorial structure of the Brazilian adult intelligence battery. *Psychology & Neuroscience*, *7*, 559-566. doi:10.3922/j.psns.2014.4.15

Wilson, M. (2005). *Constructing measures. An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wolfe, J. H., & Held, J. D. (2010). Standard errors of multivariate range-corrected validities. *Military Psychology*, *22*, 356-366. doi:10.1080/08995605.2010.491845

Zwick, R. (2004). Part I: Standardized tests and American education. What is the past and future of college admissions testing in the United States? In R. Zwick (Ed.), *SAT rethinking the future of standardized testing in university admissions* (pp. 1-4). New York, NY: Routledge.