The Development of a Lexical Knowledge Test: Assessing Lexical Knowledge of Indonesian High School Students using the Cattel-Horn-Carroll (CHC) Theory of Intelligence

Ariestianto Waskita[a] & Nurul Arbiyah[b*]

[a]*Faculty of Psychology, Universitas Indonesia, Depok, Indonesia;* [b]*Psychology Research Method Department, Faculty of Psychology, Universitas Indonesia, Depok, Indonesia*

*Corresponding Author:
Nurul Arbiyah
Psychology Research Method Department
Faculty of Psychology, Universitas Indonesia
Jl. Lkr. Kampus Raya, Depok, Jawa Barat
Indonesia, 16424
Tel.: +62 217270004
Email address: nurul.arbiyah31@ui.ac.id

The Development of a Lexical Knowledge Test: Assessing Lexical Knowledge of Indonesian High School Students using the Cattel-Horn-Carroll (CHC) Theory of Intelligence

As the most comprehensive contemporary theory, the Cattell-Horn-Carroll Theory of Intelligence (CHC) is a suitable framework for a new standard reference for senior high-school majors in Indonesia. This study attempted to construct a lexical knowledge measurement, called *Tes Pengetahuan Leksikal* (TPL), as one of the Narrow Abilities in the CHC theory, so that it could be administered in a group and serve as a reference for senior high school students in selecting a major. This lexical knowledge measurement was administered to 104 students (M_{age}=16.01; SD_{age}= 4.67). Initially, this measurement contained a set of 40 items, divided into synonym and antonym sections. Eventually, it was reduced to 20 items through a psychometric item analysis. The reliability testing resulted in the score of α = .65. Items in this measurement varied in difficulty and were able to discriminate individual levels of lexical knowledge. Distractors within each item were well-performed and distributed evenly among false choices. This measurement correlated positively and significantly with TIKI-M's subtest 3 with r = .286, which indicated that the TPL was a valid test for measuring lexical knowledge. Scores from this measurement could be interpreted using a scaled score (M = 10, SD = 3) method.

Keywords: CHC; lexical knowledge; senior high school students; major; test construction

Introduction

Within the educational setting, the need for evaluating and predicting students' abilities and skills is of paramount importance. In order to objectively evaluate student ability, standard measures are needed to can allow schools and teachers to make better decisions. Examples of such measures are national exams or intelligence tests, or in one case, a standard reference was recommendation by a junior high-school counselor. However, this type of measure lacks a strong and measurable base, and its objectivity is prone to question. These achievement test measurements did not measure purely students' abilities in a field or subject; they were affected by many other factors, including socio-economic influences, parental roles, and teacher and class contexts (Sirin, 2005; Fan & Chen, 2001; Multon, Brown, & Lent, 1991).

In addition to competence and achievement, other objective measures aimed to identify students' interests and talents through the use of intelligence tests. Since the beginning of the intelligence measurement period, one objective was to discover a person's aptitude in a field of expertise (Lohman, 2006). Intelligence tests were also considered to be able to measure one's maximum performance (Ackerman & Heggestad, 1997). Friedenberg (1995) stated that an ability test or maximum performance test could be used to measure one's potential. In addition to predicting talent, intelligence could also be used to predict vocational interests (Passler, Beinicke, & Hell, 2015). Intelligence nowadays has been interpreted as a collection of cognitive abilities that support individual performance (Carroll, 1993). Passler, Beinicke, and Hell (2015) stated that intelligence is the "motor" of interest. Intelligence determined the depth of one's potential to understand or master a field. However, several issues arise with regard to intelligence tests in Indonesia, First, most intelligence tests in Indonesia are out of date (i.e., WISC, WAIS). They

were adapted from old versions of the tests that were published abroad and there has been little attempt to revise or update them. Second, the content of the Indonesian intelligence tests was leaked on the internet. The test items, administration procedures, and test interpretations can be freely accessed by any Indonesian person. Consequently, conclusion drawn from the test results will raise questions about the test's accuracy in reflecting individual performance. Thus, the construction of new intelligence tests is needed as a means of tackling the problem.

Cattell-Horn-Carroll (CHC) Theory

The CHC Theory is one of the most comprehensive theories of intelligence for measuring various cognitive abilities (Schneider & McGrew, 2012). The taxonomy of cognitive capacity in CHC theory is the most frequently used in the contemporary framework of conceptualization and measurement of cognitive abilities (McGrew & Wendling, 2010). This theory combines the Gf-Gc Theory of Cattel-Horn and Carroll's Three-Stratum Model (McGrew, 2005, 2009). CHC has become the intelligence theory with the most abundant supporting evidence for accuracy (Kaufman, 2009; McGrew, 2005). With a detailed classification of cognitive abilities, CHC deserves to be the basis for measurement of individual capabilities and potential. Up until now, there has been no measurement based on the CHC theory that identifies specific cognitive abilities (Narrow Abilities) that correspond with the competencies expected in Indonesian education.

Lexical Knowledge

Lexical knowledge is a Narrow Ability of Domain Specific Knowledge (Gkn) of CHC theory. It is one of the competencies required by high school students. Lexical knowledge is the knowledge of the definitions of words and their underlying concepts (Schneider & Mcgrew, 2012). Whereas language development is more about understanding words in context, lexical knowledge is more about understanding the word definitions in isolation (sans context). Lexical knowledge is also an obvious precursor skill for reading decoding and reading comprehension (Schneider & McGrew, 2012).

Competence in the area of lexical knowledge is especially important for Social Scientific and Language majors (Ackerman, 2003; Qian, 2002; Nassaji, 2006). This is because, as the Broad Ability of lexical knowledge, the level of general comprehension (GC) is closely related to grades in subjects in the fields of humanity, citizenship and business (Ackerman, 2003). Lexical knowledge is also connected to language and second-language proficiency (Zareva, Schwanenflugel, & Nikolova, 2005) (Laufer & Nation, 1995; Nassaji, 2006). These areas are the main focus of subjects in the Social Sciences and Language majors.

The purpose of this study is to construct a lexical knowledge measurement using the CHC Theory of Intelligence. The goal is to create a measurement that has good psychometric qualities; that is, it consists of items that can differentiate among students with high and low lexical knowledge. It also has gradual and varying degrees of difficulty in accordance with its construction purpose, has distractors that work properly, and has the right norms to use as recommendations for senior high-school students in Indonesia. The constructed instrument is named *Tes Pengetahuan Leksikal* (TPL) and was designed as part of the CHC-based Intelligence Battery Test, for senior high school students. The composite scores from the TPL and other ability tests under the CHC Theory domain will serve as a description of student potential in

areas that are important to all high school students. Consistent with the purpose of construction, this description of student potential will be used by senior high schools as one reference in determining their major.

Methods

Participants

There were 104 participants in total (M_{age} = 16,01; SD_{age}= 4,67), who came from three senior high schools in Jakarta. 67.31% participants of this study are female and 32.69% are male. 36 participants came from the Natural Science major (34.62%) and 68 participants (65.38%) came from the Social Science major. Lexical Knowledge itself is more related to Social Science (ex: History, Economics, or Geography, etc.) than to Natural Science (ex: Biology, Chemistry, or Physics).

Measures

There were two instruments used in this study. First was the TPL test. The TPL was designed to measure lexical knowledge. Lexical knowledge describes the understanding of a word's definition in isolation and without context (McGrew, LaForte, Schrank, 2014). Therefore, the dimension of lexical knowledge was taken directly from the definition of the construct; namely, the understanding of the meaning of a word. The understanding of the meaning of a word referred to two things: breadth and depth of vocabulary knowledge. According to Haastrup and Henriksen (2000), the depth of vocabulary knowledge can be described by paradigmatic (antonyms, synonyms, hyponym, gradation) and syntagmatic relations (word class differences). Based on this, the TPL was then divided into two parts: synonyms and antonyms, and then divided again by the word class of the items. An early outline and samples of TPL items can be seen in Table 1.

Table 1
Early Outline.

Construct	Definition	Parts	Word Class	Items	Sample of Items				
					A	B	C	D	E
Lexical Knowledge	Understanding of definitions of a word	Synonym	Adjective	8	Labil	Stabil	Kokoh	Goyah	Mantap	Getar
			Verb	6	Kandas	Sukses	Jatuh	Buntu	Gagal	Terseret
			Noun	6	Cakrawala	Horizon	Langit	Matahari	Awan	Angkasa
		Antonym	Adjective	8	Capat	Lama	Pendek	Kencang	Pelan	Segera
			Verb	6	Tumbuh	Susut	Berkembang	Kecil	Kurang	Besar
			Noun	6	Integrasi	Satu	Damai	Pisah	Perang	Gabung

The second instrument used was Subtest 3 (Word Relations) from the Tes Inteligensi Kolektip Indonesia-Menengah (TIKI-M). This test is intended to measure verbal comprehension, one of the factors of intelligence based on the Factor-Model Theory from French et al. and the Structure-of-Intellect Model from Guilford (Rachmawati & Andriani, 2014). According to Perfetti (2007), verbal comprehension has a strong relationship with lexical knowledge. Based on

that argument, Subtest 3 TIKI-M was chosen as a criterion for correlation with the other test validation technique.

TIKI-M has a reliability coefficient of r = .918 and its subtest 3 has a standardized loading factor of 76 (n = 377). This showed that subtest 3 on TIKI-M was reliable and valid in measuring verbal comprehension as one factor of intelligence.

Procedure

The procedure of constructing the TPL test was to follow the test construction steps from Cohen, Swerdlik and Sturman (2013). It began with a conceptualization phase; followed by a test construction phase, try out, items analysis, and item revision.

Conceptualization

Since objective measures for determining students' majors in senior high school in Indonesia have been lackluster, there needs to be a better way to determine criteria and their respective standardized measurements. From the results of the literature review, it was decided to use the CHC theory as the basis for the construction of intelligence measurement. The CHC theory was chosen because it is the most comprehensive taxonomy of cognitive capacity in explaining specific cognitive skills (McGrew & Wendling, 2010) that are useful in the process of selecting a major.

The specific cognitive ability chosen was lexical knowledge. This choice was based on the importance of the lexical knowledge for students, especially those in the Social Science and Language majors (Ackerman, 2003; Qian, 2002; Nassaji, 2006; Zareva, Schwanenflugel, & Nikolova, 2005; Laufer & Nation, 1995).

Test Construction

The determination of the item format in the TPL was based on the construct domain of lexical knowledge. For the TPL, the item format selected was paper-and-pencil. The selected-response format type to be used was a multiple-choice format, with five alternative options. A dichotomous scoring method was used, with 1 as the correct answer and 0 given for the wrong answers.

The next step is item pooling. The number of items targeted for the final version on the TPL was 20 (10 items for synonyms and 10 for antonyms). This number was based on the fact that the TPL would be one part of a battery of intelligence tests, so the number of items needed to take participant resilience into account. For that purpose, 40 items were pooled.

Try Out

To receive feedback about the quality of the items, the test itself was tested with a sample of participants. Samples for try out were collected using a non-probability sampling method with a convenience sampling technique. Participants came from two senior high schools in Depok. In total, there were 82 participants in this try out (n_{female} = 48, M_{age} = 16,5; SD_{age}= 4,71),). Participants were first year students in high school. The maximum total score that could be obtained from this was 40 (Total $score_{min}$=0).

Item Analysis

The item analysis was performed on the following try out results: item difficulty index, item discrimination index, expected distractor power, and actual distractor power. The discrimination index items were analyzed using corrected item-total correlations. The standard used as a reference is $cr_{it} \geq .20$ (Nunnally & Bernstein, 1994). Distractor power analysis was performed using EDP and ADP comparisons. The ADP for each choice of wrong answer is predicted to be not much different from the EDP value of the item.

Item Revision

From the try out phase, 19 items out of 40 needed to be revised or eliminated based on the analysis results for each item. For the purpose of keeping back-up items for field study, only 4 of the 19 items that did not pass the standard in item analysis were eliminated. The remaining 15 items were revised.

Data Analysis

Cronbach's Alpha was used in measuring inter-item correlation in try out and field study. For validation purposes, Pearson Product Moment (1-tailed) was used to measure the correlation between the TPL score and the criterion (TIKI-M Subtest 3).

Results

Item Analysis

The item analysis process involved an item difficulty analysis, an item discrimination analysis, and a distractor power analysis. Besides a quantitative item analysis, there was also a qualitative item analysis. This analysis examined content, design, and item representation within the theoretical framework.

Item Difficulty Analysis

As seen in Table 2, the item difficulty index has five categories. However, there are some items whose difficulty index was not categorized as designed. Overall, these item tests have a high item difficulty index, which means item tests are relatively easy for participants of this research to answer. The result of the item difficulty analysis is shown in Table 2.

Table 2
Item Difficulty Index.

Difficulty Index Range	Category	Item Number	Total Amount
1,0–0,8	Very Easy	Part 1: 1, 11, 16, 17 Part 2: 2, 3, 5, 6, 9, 11, 12, 17	12 (35.30%)
0,8–0,6	Easy	Part 1: 2, 4, 9, 10, 14, 18 Part 2: 10, 13, 14, 16, 18	11 (32.35%)
0,6–0,4	Medium	Part 1: 3, 13 Part 2: 1, 4, 7	5 (14.71%)
0,4–0,2	Hard	Part 1: 5, 12, 15 Part 2: -	3 (8.82%)
0,2–0,0	Very Hard	Part 1: 6, 7, 8 Part 2: -	3 (8.82%)

Item Discrimination Analysis

According to statistical analysis methods, the item discrimination index is completed using the corrected item-total correlation technique. The correlation coefficient on each item is then compared to the criteria according to Nunally and Bernstein (1994). Based on those criteria, two categories are created for the item discrimination index: good discrimination ability and poor discrimination ability. "Good discrimination ability" is the category for items with a discrimination index above criteria 0,2, whereas "Poor discrimination ability" is the category for items with a discrimination index below criteria 0,2.

As seen in Table 3, there are 19 test items that are categorized as "good discrimination ability," which means have a discrimination index above 0,2. Items with an index above 0,2 have the capability of differentiating whether people have or do have not the ability to answer correctly. However, 15 items are categorized as "poor discrimination ability." Those items have an insufficient capability of differentiating people's ability. Items with "poor discrimination ability" are evaluated or deleted from the test.

The result of the corrected item-total correlation measure is shown in Table 3.

Table 3.
Discrimination Item Index.

Indicator cr_{it}	Description	Item Number	Amount
>0,2	Good discrimination ability	Part 1: 2, 6, 8, 11, 14, 18 Part 2: 1, 2, 3, 5, 6, 9, 10, 11, 12, 13, 16, 17, 18	19 (55.88%)
<0,2	Poor discrimination ability	Part 1: 1, 3, 4, 5, 7, 9, 10, 12, 13, 15, 16, 17 Part 2: 4, 7, 14	15 (44.12%)

Distractor Power Analysis

The distractor power analysis result is divided into two categories: functioned and not functioned. An item is categorized as "functioned" if the Actual Distractor Power (ADP) is on every distractor and the Expected Distractor Power (EDP) is equivalent. As seen in Table 4, there are 20 items categorized within the "functioned" category, which means that each distractor on those items can optimally function to outwit people without ability as they are being measured with this test. Items in the "not functioned" category are evaluated or deleted from this test.

The results of distractor power analysis and its functionality are shown in Table 4.

Table 4
Distractor Power Analysis.

Distractor Power	Item Number	Amount
Functioned	Part 1: 1, 4, 6, 7, 9, 10, 11, 16, 17 Part 2: 3, 4, 5, 9, 10, 11, 12, 13, 14, 17, 18	20 (58.82%)
Not Functioned	Part 1: 2, 3, 5, 8, 12, 13, 14, 15, 18 Part 2: 2, 6, 7, 16	14 (41.18%)

Integrative Items Analysis

Various items analysis results (quantitative and qualitative) will serve as the basis for integrative analysis. The integrative analysis was done to select the 20 best items from the available item-pooling. This analysis also considered the representation of items within each dimension proportionately. Through this analysis, 14 items were eliminated because they did not meet the criteria of good items (quantitative and qualitative). The selected 20 items (part 1: item number of 2, 5, 6, 8, 11, 12, 13, 14, 16, and 18; and part 2: 1, 3, 5, 9, 10, 12, 13, 16, 17, and 18) will be tested for reliability and validity to measure the psychometric aspects of this test.

Reliability

We employed the inter-item consistency test of the TPL for the reliability analysis. With 20 test items through a single trial test using an alpha coefficient, it yielded result $\alpha = .65$ ($M = 12.53$; $SD = 2.91$). This result showed that 65% of observed score came from true variance and 35% was error variance derived from content sampling and item heterogeneity. According Cohen, Swerdlik, & Sturman (2013), the lowest acceptable category of reliability is $.65 \leq \alpha < .80$. Based on these references, TPL items were consistent in measuring the same constructs and could be said to be homogeneous, although they still required some revisions. The SEM value obtained was 1.72 ($LoC = 95\%$).

Validity

The TPL validity test used a construct validity method with correlations with other test techniques. This was done using the Pearson Product Moment (1-tailed) correlation model. Result of validity test of TPL with subtest 3 "Words Relations" in TIKI-M as criterion, yielded correlation index $r = .313$ ($n = 104$; $p < .001$). This showed that the TPL could explain 9.8% variance on the subtest 3 "Word Relations" in TIKI-M. Thus, the TPL was valid in measuring the construct of lexical knowledge because it had significant positive correlation with the subtest 3 "Word Relation" in TIKI-M that measures verbal comprehension.

Norms

The norms used in the TPL were group norms. The norming process used standard score techniques by changing the value of the raw scores into the new standard scores. The standard scores used were $M = 10$ and $SD = 3$. With the new scaled score, the TPL had a new range of scores from 0 to 18. This means that participants who answered all the items wrong will have scores of 0, and those who answer all items correctly receive scores of 18. The overall norming scores can be seen in Table 5.

Table 5
Norms

Scaled-score	Raw-score
18	20
17	19
16	18
15	17
14	16
13	15
12	14
10	13
9	12
8	11
7	10

Table 5, cont.
Norms

Scaled-score	Raw-score
6	9
5	8
4	7
3	6
2	5
1	4
0	3–0

Discussion

In the try out phase, the TPL had 40 items consisting of two parts: synonyms and antonyms. Both sections had items that represent three classes of words: adjectives, verbs, and nouns. The results of reliability testing in the try out phase showed that the TPL was not reliable enough. Based on the item analysis, the variation of the difficulty index on the TPL was good. In terms of the item discrimination index, there were still quite a lot of items with poor discrimination indices. Based on the analysis of integrative tables, it was found that there were false-choice options on some items chosen by more participants than the correct answer.

The low reliability coefficients gave the participants' scores a large percentage of errors. This led to a large range of participant true-scores. Thus, it was possible to have participants true-scores intersect among themselves. The low reliability coefficients in the TPL are still considered acceptable because, as Nunnally and Bernstein (1994) argued, the TPL used construct validation and was still in the early stages of development. It also shows that improvements in the maximum-performance test are not as easy as in typical performance types (Nunnally & Bernstein, 1994). Improvements in the answer options and the elimination of items after the try out phase did not necessarily increase the reliability coefficient significantly.

One possible cause of the low reliability coefficient of the TPL is the variability of the sample (Anastasi & Urbina, 1997). The sample selection using the convenience sampling technique does not take into account the variation of sample characteristics, so the probability of obtaining samples with high homogeneity increases. It is possible that one of them is caused by similarities in school characteristics used as data sources, such as location and passing grades (Pemerintah Provinsi DKI Jakarta, 2015). Another possible reason for this issue was the low number of samples tested. Increasing the number of samples would increase the probability of a better reliability coefficient.

According to Nunnally and Bernstein (1994), improving a reliability coefficient can be done by sorting all items based on cr_{it} value, then selecting the item to use from the highest cr_{it} value. They also said that if the method has not been able to increase the reliability coefficient to the desired number, then items with the value of cr_{it} >.20 can be added (Nunnally & Bernstein, 1994). This is difficult to apply in the TPL context. The seemingly-unchangeable reliability coefficient is due to the unevenness of the cr_{it} score in the two test sections. In Part I (synonyms), many items had low cr_{it} values. The opposite thing happened in Part II (antonyms). On the basis of the equality of each part and the representation of each word class in each part, there were items with a high cr_{it} value (3) in Part II that had to be removed. Meanwhile, in Part I, there were

items with low cr_{it} value (4 items) that had to be maintained. Ultimately, both of these mutually eliminated each other, so the reliability coefficient did not change.

The results of item analysis in field study showed considerable changes compared to the results in the try out phase. The most obvious change was apparent from the distribution of the item difficulty level. At the try out phase, the difficulty level of the item was distributed evenly, with the "moderate" difficulty level having the highest amount. The TPL version for field study showed that the items had difficulty levels that tended to be easy. From here, we can conclude that changes were made in the answer choices, which was followed by a decrease in difficulty level. An example was on the "Cepat" item in part II (antonyms). In the try out phase, the answer options on this item were: A. Panjang; B. Pendek; C. Kencang; D. Lamban (the right answer); and E. Segera. In the try out, the "A" answer option became the most-preferred choice. After being revised by changing the "D" answer option to "Lambat," the item difficulty level changed drastically (*p*.220 to.577; cr_{it}.082 to.223). Since item selection is preferred for prioritizing the cr_{it} value (Anastasi & Urbina, 1997; Nunnally & Bernstein, 1994), the difficulty level of TPL items is sacrificed, even though the content consideration still retained some items that were difficult (with low cr_{it} values).

For the validity coefficient of correlations, there was a slight decrease after the elimination process based on the results of field study. The correlation coefficient obtained was below the generally-accepted value (.30 -.40) (Nunnally & Bernstein, 1994), but according to Anastasia and Urbina (1997) the correlation coefficient value of.20 -.30 is still acceptable for use in the selection process, under certain conditions. It does not vary much with the main purpose of the TPL. This low coefficient of correlation showed that even though both test kits had a significant relationship, there were considerable differences between the two tests. This might be due to the different theories of intelligence on which both tests are based. In addition, differences in both test formats also may have contributed to low coefficients of correlation.

The results of the reliability test using Cronbach's Alpha method showed that the TPL has consistent items in measuring a construct. However, the reliability coefficients obtained belong to the low category, indicating that the items on the TPL still needed to be revised or tested again on different samples.

The validity test using the subtest 3 "Words Relations" of TIKI-M as a criterion showed a significant positive relationship; it indicates that the TPL was valid in measuring the construct of lexical knowledge. However, the correlation coefficient obtained (r = .286) fell below the commonly used standard, which is around.30 -.40 (Nunnally & Bernstein, 1994).

The item analysis performed on the TPL shows that TPL items are less able to differentiate the lexical knowledge level among individuals. The results of the item difficulty index calculation also showed that the TPL's items have varying degrees of difficulty, less suitable for use as a recommendation for the purpose of selecting a major. The distractor power analysis showed that the answer choices for TPL items worked quite well and were distributed evenly. The norms suitable for use in the TPL were group norms, with standard score techniques using a mean of 10 and a standard deviation of three.

For future research and development of similar or the same measurements, more items are recommended for the item pooling phase. This is because the nature of the lexical knowledge construct makes it nearly impossible to revise only the poor items. A change in one word for an item or answer choice converts that item into a completely different item than the researcher first intended. With a pool of items that consist of a lot more components, the possibility of getting more good items without a need to revise the poor ones will increase. This would make the study much more effective and efficient. Secondly, for a construct of verbal nature such as this, content validity has the same importance as construct validity. For example, in the case of lexical knowledge construct, expert judgment from linguists or prior study about vocabulary development could serve as a good evaluation of an item aside from the "conventional" item analysis and cr_{it}.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, *121*(2), 219.

Ackerman, P. L. (2003). Cognitive ability and non-ability trait determinants of expertise. *Educational Researcher*, *32*(8), 15-20. Doi: 10.3102/0013189X032008015

Anastasi, A., & Urbina, U. (1997). *Psychological testing* (7th Ed). New Jersey: Prentice Hall.

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

Cohen, R. J., & Swerdlik, M. (2009). *Psychological testing and assessment: An introduction to test and measurement*. New York: McGraw-Hill.

Cohen, R. J., Swerdlik, M. E., & Sturman, E. D. (2013). *Psychological testing and assessment: An introduction to tests and measurement* (8th Ed). New York: McGraw-Hill.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando: Holt, Rinehart, and Winston.

Ebel, R. L., & Friesbie, D. A. (1991). *Essentials of educational measurement*. New Jersey: Prentice Hall.

Fan, X., & Chen, M. (2001). Parental involvement and students' academic achievement: A meta-analysis. *Educational psychology review*, *13*(1), 1-22.

Flanagan, D. P., & Harrison, P. L. (2012). *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.). New York: Guilford Press.

Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). *The Wechsler Intelligence Scale and Gf-Gc theory: a contemporary approach to interpretation*. Needham Heights: Allyn & Bacon.

Friendenberg, L. (1995). *Psychological testing: Design, analysis, and use*. Boston: Allyn & Bacon.

Haastrup, K., & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics*, *10*(2), 221-240. Doi: 10.1111/j.1473-4192.2000.tb00149.x

Howard, K. I., & Forehand, G. A. (1962). A method for correcting item-total correlations for the effect of relevant item inclusion. *Educational and Psychological Measurement*, *22*(4), 731-735. Doi: 10.1177/001316446202200407

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport: Praeger.

Kaplan, R. M., & Saccuzo, D. P. (2012). *Psychological testing: Principles, applications, & issues.* California: Thomson Wadsworth.

Kaufman, A. S. (2009). *IQ testing 101*. New York: Springer Publishing.

Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2014, Juli 11). *Peminatan pada pendidikan menengah*. Kementerian Pendidikan dan Kebudayaan. Diunduh dari http://jdih.kemdikbud.go.id/new/public/produkhukum

Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2014, Januari 14). *Pedoman peminatan peserta didik*. Kementerian Pendidikan dan Kebudayaan. Diunduh dari https://kemdikbud.go.id/kemdikbud/dokumen/Paparan/Paparan%20Warmendik.pdf

LaForte, E. M., McGrew, K. S., & Schrank, F. A. (2014). *Assessment Service Bulletin Number*.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, *16*(3), 307-322.

Lohman, D. F. (2006). Beliefs about differences between ability and accomplishment: From folk theories to cognitive science. *Roeper Review*, *29*(1), 32-40.

Lohman, D. F., & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted*, *29*(4), 451-484.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151–179). New York: Guilford.

McGrew, K. S. (2005). The Cattell–Horn–Carroll theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136–181). New York: Guilford Press.

McGrew, K. S. (2009). Editorial: CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37, 1–10.

McGrew, K. S., & Wendling, B. J. (2010). Cattell-Horn-Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in The Schools*, *47*(7). Doi: 10.1002/pits.20497

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation.

Nassaji, H. (2006). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success. *The Modern Language Journal*, *90*(3), 387-401.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill.Inc.

Pässler, K., Beinicke, A., & Hell, B. (2015). Interests and intelligence: A meta-analysis. *Intelligence*, *50*, 30-51. Doi: 10.1016/j.intell.2015.02.001

Pemerintah Provinsi DKI Jakarta. (2015). *Passing grade SMA DKI Jakarta*. Diakses dari ttp://ppdbdki.org/passinggrade/index.aspx/

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific studies of reading*, *11*(4), 357-383.

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. *Precursors of functional literacy*, *11*, 67-86.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513-536.

Rachmawati, F. A., & Andriani, F. (2014). Confirmatory factor analysis: Tes Inteligensi Kolektip Indonesia Tingkat Menengah (TIKI-M). *Jurnal Psikologi Pendidikan dan Perkembangan*, 3(1).

Salthouse, T. A. (1993). Speed and knowledge as determinants of adult age differences in verbal tasks. *Journal of Gerontology*, *48*(1), p 29-36.

Saputro, I. (2016). Jurusan IPA dianggap lebih bergengsi, sekolah bingung tentukan penjurusan siswanya. *Tribun Solo.com*. Diakses dari http://solo.tribunnews.com/2016/06/24/jurusan-ipa-dianggap-lebih-bergengsi-sekolah-bingung-tentukan-penjurusan-siswanya

Schneider, W. J., & McGrew, K. (2012). *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed., pp. 99–139). New York: Guilford Press.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, *75*(3), 417-453.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, *27*(04), 567-595.