Development of an Indonesian Figural-Inductive Reasoning Test for High School Students Based on the Cattell-Horn-Carroll Theory

Anita Dwinata Lubis[a] and Dewi Maulina[b]*

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

*Corresponding author:
Dewi Maulina
Psychology Research Method Department
Faculty of Psychology, Universitas Indonesia
Jl. Lkr. Kampus Raya, Depok, Jawa Barat
Indonesia, 16424
Tel.: +62 217270004
email address: dewi_maulina@yahoo.com

Development of an Indonesian Figural-Inductive Reasoning Test for High School Students Based on Cattell-Horn-Carroll Theory

The changing of Indonesia's school curriculum in 2013 brought new regulations to divide students into specializations once they reach tenth grade. This study was conducted to develop a new test called Figural-Inductive Reasoning Test (*FIR* test) as a subtest of the new intelligence battery test, *Tes Inteligensi Siswa SMA* (*TISS*) or High School Student Intelligence Test, which is used as a method of helping students select a specialization. The *FIR* test was developed based on Schneider and McGrew's (2012) Cattell-Horn-Carroll Theory of Human Intelligence (CHC). The *FIR* test was constructed in a figural spatial format and tested in order to evaluate its psychometric properties. Data were collected from tenth grade students in three different high schools in Jakarta (N = 97). Reliability testing using *Cronbach's Alpha* method showed that the *FIR* test had adequate internal consistency in measuring a construct. The *FIR* test was considered valid in measuring inductive reasoning since it was shown to correlate significantly with *Standard Progressive Matrices* (*SPM*), which measures the same construct. Item difficulty analysis indicated that most items had a low difficulty level. Item discrimination analysis indicated that half of the items had a sufficient item discrimination index. Distractors power analysis revealed that some distractors needed revision in order to increase the item qualities. From the research results, it can be concluded that the *FIR* test is a promising subtest for the new *TISS* test in Indonesia. However, items in the *FIR* test still need to be revised and retested with a larger sample so that it can be applied and used for making decisions about high school students' specialization.

Keywords: CHC intelligence test; high school students; inductive reasoning

Introduction

In Indonesian high schools, students are allowed to choose one of several provided specializations: Natural Science, Social Science, or Language. In the prior curriculum, students' grades from tenth grade were used to split students among specializations to begin in eleventh grade (Riandari, 2007). The legalization of a new curriculum in 2013 brought new regulations to all Indonesian high schools; selecting student specialization would now begin in the tenth grade instead (Kemendikbud, 2013). This means that specialization would now be chosen before students' evaluation by school examination. Because schools could no longer base the specialization selection on grades, they could no longer simply split students mechanically. The new regulation caused schools to have to find a new effective procedure to assign students into the appropriate specialization; this was important because the majors that they select determine their later careers. This period of high school was an important phase because this is when students explored their interests before planning the next stage of their careers (Super & Hall, 1978).

Based on several alternatives for selecting specialization from the 2013 curriculum, placement tests were considered the best method. Placement tests, especially psychology tests, could be most effective for use in placement because of their ability to predict future performance, describe particular psychological traits, and even explain an individual's covert behavior (Anastasi & Urbina, 1997; Cohen, Swerdlik, & Sturman, 2013; Murphy & Davidshofer, 2005).

One of psychological variables that can predict students' ability in school is intelligence (Malykh, 2017; Koller, Baumert, & Schnabel, 2011; Weinert, Helke, & Schneider, 1989). Intelligence tests were used in several schools to provide recommendations for student specialization; for example, in Madrasah Aliyah (MA), Citra Cendikia, and SMA Kristen Petra 3 Surabaya (Madrasah Aliyah Citra Cendikia, 2015; sma3.pppkpetra.or.id). However, there were several disadvantages of using the current Indonesian intelligence test. First, many online sites had posted the intelligence test publicly, such as one publication about the Tes Inteligensi Umum (General Intelligence Test) (Sehertian, n.d.). In addition, the current intelligence test, *Tes Kemampuan Differensial* or TKD (Differential Ability Test), was still based on the classic intelligence theory of primary mental ability, rather than on a more recently developed intelligence theory (Manual TKD Lengkap, n.d.). The use of a more recent, contemporary theory of intelligence would better predict students' achievement.

Intelligence is defined as a set of individual abilities involving a thinking process to solve problems and adapt to the environment (Gardner, 2011; Sternberg, 1997; Wechsler, 1944). Hunt (2011) stated that intelligence could be explained in three approaches: psychometric, cognitive, and biological. Among these three, the psychometric approach was the most influential and the most developed approach to describing an individual's intelligence (Neisser et al., 1996). One of the best psychometric approach theories in defining human cognitive structure is the Cattell-Horn-Carroll theory of human intelligence or CHC (McGrew, 2009).

CHC integrated two intelligence theories: the Cattell-Horn Gf-Gc and Carroll Three Stratum Cognitive Abilities (Schneider & McGrew, 2012). CHC sought to explain human cognitive ability in three layers or strata. These consisted of general ability (*g*), broad ability, and narrow ability. Currently CHC has about 16 broad abilities and 81 narrow abilities (Schneider & McGrew, 2012). The CHC theory explained human cognitive ability in terms of taxonomy or categorization. This theory was built on a strong and comprehensive psychometric foundation (McGrew, 2009). Because of its acuity in explaining individual-specific abilities, intelligence tests based on this theory can serve as an effective tool for determining high school students' specialization.

Unfortunately, intelligence tests based on CHC from abroad are difficult to implement for selecting Indonesian high school students' specialization. The differences of language and culture make it impossible to directly apply the tests in Indonesia. Furthermore, some of the CHC tests are individually administered; for example, the Woodcock Johnson III Cognitive Abilities test (Blackwell, 2009). If the test were used for selecting specialization in high school, the test would not be effective due to large number of students needing testing. Existing tests had also not been tailored to specifically measure several aspects of high school students' cognitive ability. Therefore, it is necessary to create a new CHC-based intelligence battery test that measures various cognitive abilities in various specializations in high school.

Fluid reasoning is one of the various cognitive abilities in CHC theory proven to be needed by students in high school. Fluid reasoning is correlated significantly with academic achievement.

Fluid reasoning (*Gf*) refers to the ability to solve novel problems without depending on previous education or experience (Schneider & McGrew, 2012). *Gf* significantly predicted student academic achievement in mathematics and reading (Floyd, Evans, & McGrew, 2003; Green, et al., 2017; Taub, Keith, Floyd, & McGrew, 200_). As an ability to solve novel problems, *Gf* is considered core intelligence because the factor analytic found that *Gf* had the greatest factor loading to intelligence (Carroll, 1993).

As a broad ability, *Gf* consists of three narrow abilities: induction (*I*), general sequential reasoning (*RG*), and quantitative reasoning (*RQ*). Induction, commonly referred to as inductive reasoning, is the core and most important narrow ability of Gf. Inductive reasoning is one of the key abilities needed for success in high school because it requires students to comprehend a principle and then make conclusions based on the information (Kemdikbud, 2013). Students with high inductive reasoning ability are more likely to succeed in natural science subjects that rely heavily on logic such as mathematics (Roth, et al., 2015) and biology (Lawson, Banks, & Logvin, 2006). Therefore, measuring inductive reasoning ability is important because it is necessary to succeed in high school.

Induction or inductive reasoning is defined as *"*the ability to observe a phenomenon and discover the underlying principles or rules that determine its behavior*"* (Schneider & McGrew, 2012, p. 112). Inductive reasoning describes a mental process of categorizing, applying, and inferring rules or principles (LaForte, McGrew, and Schrank, 2014; Schrank, 2010). This means being able to understand the rules of observed phenomena, make hypotheses about relationships between those phenomena, and then be able to draw accurate inferences about the relationships (Wilhelm, 2005).

Inductive reasoning can be measured in the form of figural spatial stimuli, such as geometric shapes. This is beneficial because figural-spatial stimuli tend to be free from the influence of other constructs such as memory (Wilhelm, 2005). Matrices are cognitive tasks that can be used to measure inductive reasoning (Carroll, 1993; Wilhelm, 2005). Rules or principles that are commonly used in matrices tasks are constant in a row, quantitative progressive pairwise, figure addition or subtraction, and distribution of three values (Carpenter, Just, & Shell, 1990; Primi, 2014). Stimuli complexity, type, and number of rules are factors that affect an item's level of difficulty (Primi, 2014). This study will use figural-spatial stimuli because they are considered cultural-fair stimuli; because the target test takers come from various Indonesian high schools, it was necessary to make sure that no one particular group was harmed by the method of administration for this inductive reasoning test. Cattell (1987) recommended that stimuli used in inductive reasoning tests worked best if either all familiar or all unfamiliar to all test takers.

Therefore, inductive reasoning is one of several abilities that the intelligence battery test will measure to assist high school students in selecting a specialization. The aim of this study was to develop an inductive reasoning subtest as part of the intelligence battery test (*TISS*, or *Tes Inteligensi Siswa SMA*). The new subtest will be named Figural Inductive Reasoning Test (*FIR*). This study will also test the psychometric properties of *FIR* so that it will meet the requirements for a good psychological test. Thus, this study will answer the following research questions:

1. Is the *FIR* test is reliable; that is, does it have good internal consistency in measuring a particular construct?
2. Is the *FIR* test valid in measuring inductive reasoning by correlation with another inductive reasoning test?

3. Do items in the *FIR* test have a suitable degree of difficulty? 4. Are items in the *FIR* test able to distinguish between students with high and low inductive reasoning ability?
5. Do items in the *FIR* test have satisfying distractors?

Methods

Participants

Participants were obtained from high school students in different class specializations. Participants were comprised of 97 tenth grade students ranging from 14 to 17 years old (M = 16.04, *SD* = 0.44) from three Jakarta public high schools. Participants' genders were equally distributed (female = 56.70 %). Most participants (72.16 %) were from the natural science specialization. Participants were recruited using *convenience sampling*. School selection was based on perceived level of school quality by Jakarta citizens and categorized as high, medium, and low quality.

Measures

The FIR test is an optimum performance test that either has a time limit or is administered as a timed-power test. The *FIR* test is a paper and pencil test that is administered to a group. The FIR test items were developed based on three indicators. According to Schneider and McGrew (2012), individuals who had high inductive reasoning ability were able to: (1) identify rules, characteristics, or patterns from particular materials or stimuli; (2) categorize materials or stimuli based on rules, characteristics, or patterns and determine which materials or stimuli do not conform with particular rules, characteristics, and patterns; and (3) state the underlying principles of particular materials or stimuli explicitly. All the three indicators of inductive reasoning were measured using a multiple choice format with five alternatives. A score of +1 was given for every right answer and zero was given for every wrong answer. A total score was obtained by adding together all true answers. Therefore, a higher score indicated an individual's higher inductive reasoning ability. The item specifications for the *FIR* test are seen in Table 1.

Table 1
Test Specifications of the *FIR* Test

Item category	Numbers of rules	Type of stimuli	Target item	Item pooling
Easy	1	Familiar, matrices 2 × 3	7	10
Medium	1 or 2	Familiar, matrices 3 × 3	14	21
Difficult	2	Non-familiar matrices 3 × 3	9	14
Total			30	45

The items were divided into three categories: easy, intermediate, and difficult. The proportion of medium items was made the largest in order to obtain the optimal item discrimination index. The degree of item difficulty was based on numbers of rules and types of stimuli represented in the item. The rules were constant in a row, quantitative progressive pair ways, figure addition or subtraction, and distribution of three values. The types of stimuli were divided into two categories: familiar and non-familiar. The familiar stimuli consisted of geometrical shapes such as triangles and rectangles, while the non-familiar stimuli consisted of random patterns and were not associated with everyday objects. Matrices were divided into two forms: 2 × 3 matrices constructed of 2 rows and 3 columns to make a total of six cells; and 3 × 3 matrices constructed

of 3 rows and 3 columns for a total of nine cells. The last cell of each matrix was left empty so that the test-taker could apply deduced rules to fill in the missing cell.

Testing Method Data Analysis

Reliability testing in the *FIR* test was internal consistency with *Cronbach's Alpha*. Reliable test criteria were based on Kaplan and Saccuzzo's work (2005).The construct validity testing used the *Raven's Standard Progressive Matrices (SPM)* as the criteria. *SPM* was proven to be a valid test in measuring inductive reasoning (Wilhemn, 2005). Item difficulty analysis was conducted to ensure that each item had a satisfactory difficulty level. Easy items were expected to have $p >$ 0.6; intermediate items were expected to have p between 0.4 and > 0.6; and difficult items were expected to have $p \leq 0.2$. Item discrimination analysis used the *corrected item-total correlation* (cr_{it}) method, where a good discriminating item had $cr_{it} > 0.2$ (Nunnally and Bernstein, 1994). Distractor power analysis was carried out to ensure that each distractor had a satisfactory ability to deceive the test taker. Good distractors were those with a small ratio between the *expected distractor power* (EDP) and *actual distractor power* (ADP) (Friedenberg, 1995).

Procedure and Data Analysis

Expert Judgment and Readability Process

After test conceptualization, item pooling for as many as 45 items was collected. *Expert judgment* was conducted by two psychometric and intelligence experts. The experts evaluated every item based on the appropriate indicator and item specifications. Some items were found to be too difficult because they contained stimuli that were too complex. There were also ambiguous items that could be solved by applying more than one rule. The results of the expert judgment were used to revise and evaluate items in terms of item difficulty and quality, as well as test instruction.

A readability process was conducted on eight tenth grade students from Jakarta and Bogor and three ninth grade students from Jakarta and Depok. The ninth grade students were involved in order to enrich item feedback. Based on the readability process, ambiguous and confusing items were revised.

Try Out

In the try-out process, participants came from two public high schools in Depok. The first high school was still using the KTSP curriculum, so the students had not yet been divided into different majors. The total number of students involved in the try-out process was 62 (male = 37). The results are listed in Table 2.

Table 2
Try-out results

Description	
Mean	22.68
Standard deviation	5.48
Variance	30.1
Mean of duration	21 minutes 33 seconds
Maximum score	33
Minimum score	9

The percentage of participants who completed the *FIR* test on minute-25 was 75.62%. This illustrated that 25 minutes was the optimal amount of time in which to complete the test (N = 45). The reliability testing of the *FIR* test with *Cronbach's Alpha* showed $\alpha = 0.716$. According

to Kaplan and Saccuzo (2005), a reliability index of 0.70 can be categorized as reliable. Hence, we can infer that the *FIR* test had good internal consistency in measuring a particular construct.

Item difficulty analysis showed that 16 items were relatively easy; 10 items were relatively medium; and 19 items were relatively difficult. Item discrimination analysis showed that 22 items had $cr_{IT} \geq 0{,}2$. This means that 22 out of 45 items already had a good discriminating power to distinguish between students with high and low inductive reasoning ability. Item difficulty index, item discrimination level, and distractor power were analyzed, and due to the low quality in item discrimination index and item difficulty index, it was decided to reduce the total by five items. Furthermore, since the *FIR* test was part of *TISS*, the number of items should be considered in order to achieve the most efficient intelligence test battery.

Results

The total number of items tested was 40, with a 25-minute time limit. Descriptions of *FIR* test are listed in Table 3.

Table 3
The Description of the *FIR* Test

Description	
Mean Score	26.95
Standard deviation	4.79
Maximum score	35
Minimum score	9
Mean time	22 minutes 15 seconds
Minimum time	16 minutes 1 second
Maximum time	32 minutes 46 seconds
N completed in 25 minutes	62 (63.94%)
N completed in 27 minutes	75.62%

According to Crocker and Algina (2004), based on mean score obtained, the *FIR* test could be categorized as a test with a medium to easy level of difficulty since participants were able to answer an average of 27 out of 40 items (*mean* = 26.95, *SD* = 4.79). Observation during the testing showed that there were still many test takers who did not complete their test within the given time limit. Only 65% of participants completed the test in 25 minutes. This result indicated that the time limit of 25 minutes was still not the optimal time for the *FIR* test. Participants continued to finish the test until the 30-minute mark. Further analysis shows that the optimal time for completing the *FIR* test was 27 minutes; 75% of the respondents finished the test in 27 minutes.

Reliability testing shows that *FIR* had a coefficient *Cronbach's Alpha* = 0.72. According to Kaplan and Saccuzo (2005), this indicated that the *FIR* test had an adequate internal consistency in measuring the same construct. This also meant that 71.60% of observed score variance came from true scores and 28.40% came from content heterogeneity and content sampling error. In correlation with the *SPM* test, the construct validity testing showed $r = 0.63$, $n = 97$, $p < 0.01$. Thus, the *FIR* test was valid for measuring inductive reasoning ability because it correlated significantly with the *SPM* test that also measured the same construct. There was a 39.69% variance of the *FIR* test as associated with the *SPM* test.

Item difficulty analysis showed that 28 items were categorized as easy; 9 items were categorized as medium difficulty; and 3 items were categorized as difficult items (see Table 4). This shows

that most of the items in the *FIR* test were relatively easy. Item discrimination analysis showed that 23 items had a good discriminating power, while the other 17 items still had a low discriminating ability. The low item discrimination index was influenced by the item difficulty index. Many participants could answer an easy item correctly, whereas only a few participants could answer difficult items correctly. Both conditions affect the item's ability to differentiate ability. Distractor power analysis showed that 80% of items had enormous distance between ADP and EDP. This indicates that most items in the *FIR* test still had low distractor power. This affects item difficulty because participants could correctly answer the test by guessing from the given distractors.

Based on reliability testing, validity testing, and item analysis, we conducted an integrative analysis to select the best items for the final *FIR* test. The criteria used to select items were item difficulty index (*p*), item discrimination index ($cr_{it} \geq 0.2$), reliability test if item was deleted, and distractor power with least discrepancy between ADP and EDP. From the integrative analysis, there were 30 most satisfactory items. The items had good quality in measuring inductive reasoning but still had to be revised to increase quality. Revision had to be conducted in order to increase difficulty level and quality of stimuli, such as adding more rules and discarding some stimuli features. Finally, the 30 selected items were sorted according to level of difficulty, as seen in Table 4.

Table 4
Final Result of Item Selection

Item difficulty	Item order	Item total
Easy	2, 6, 9, 8, 5, 10, 3	7
Medium	25, 11, 17, 14, 16, 21, 13, 24, 23, 22, 27, 15, 12, 18	14
Difficult	38, 34, 37, 30, 36, 33, 40, 39, 32	9
Total		30

Note: Easy (p above .6); Medium (.4 ≤ p ≤ .6); Difficult (p below .4)

Reliability analysis was retested on the 30 selected items and showed that the coefficient α was increased to 0.727. Furthermore, validity testing showed that there was also enhancement in the validity coefficient ($r = 0.67$, $p < 0.01$) between the *FIR* test and *SPM*. Thus, it can be inferred that the *FIR* test has a high internal consistency and is valid for measuring inductive reasoning.

Discussion

The new *FIR* test was developed and examined to meet the standard of a good test according to psychometrics. The *FIR* test was found reliable as well as valid in measuring inductive reasoning but still needed several improvements in item difficulty, item discrimination, and distractor power. Future retests for the revised items were also needed to examine psychometric properties.

The high *FIR* score from samples was related to the characteristics of participants tested in the psychometric testing. The large number of items categorized as easy (n = 28) was influenced by participants' overall high inductive reasoning ability. This is supported by their *SPM* scores. Participants' *SPM* scores were also high, which meant they had a high level of inductive reasoning ability. It is possible that this also resulted from the school characteristics. The differences between high, intermediate, and low quality schools were not stated clearly, so selected schools might have shared the same qualities. In further examination, there is a need to create a school quality indicator before selecting which schools will participate.

Another finding was related to time limit of the test. It seemed that the previous time limit was not optimal because the number of participants who completed the test did not achieve 75% yet. On the other hand, the number of correct answers in the field data was higher than it was during the try-out process. This means that participants spent more time to correctly answer more items. Nunnally and Bernstein (1994) stated that in a timed-power test, a huge duration should be given to complete the test so that at least 90% of the test takers finish it. As a result, in this case, the time limit given was not sufficient to complete the test, so further examination is needed to determine the *FIR* test's optimal duration.

The reliability testing showed that the test has internal consistency in measuring inductive reasoning. This means that the constructs measured in test were homogenous. This result is consistent with de Koning, Sijtsma, and Hamers (2003) and Wilhelm (2005) who stated that figural-spatial stimuli were the best way to measure inductive reasoning because they are free from the influences of other constructs. On the other hand, there was a 28.4 % error variance from content heterogeneity and content sampling error. The error might have originated from the gestalt effect from stimuli (Van der Ven & Ellis, 2000). This means that perhaps the test taker is able to solve items by applying gestalt rather than the rules they found in the problems given. Therefore, further improvements are needed to enhance the *FIR* test reliability by reducing the gestalt effect. Little and Akin-Little (2014) also argued that the test ought to have a reliability coefficient of at least 0.8 to be used for any purpose in school. To increase the test's reliability index, we could revise the item, especially to increase cr_{IT}. It is better for an intelligence test to have more than one reliability score because some tests usually have more than one method of reliability, such as WJ-III COG (LaForte, Schrank, & McGrew, 2014). Test-retest reliability could be conducted to enhance test quality and to indicate that the test was successfully developed and could be used for several purposes.

The construct validity testing showed that the *FIR* test and *SPM* correlate highly with each other and have high amounts of shared variance. The high correlation is related to the similarity of their formats. Another factor was that the rules used in both tests were also similar (Carpenter, Just, & Shell, 1990; Primi, 2014). However, another construct validity testing could be conducted to confirm the *FIR* test validity; for example, using a convergent and discriminant method. In the previous examination, the restricted amount of time restrained the usage of any other validity method. Therefore, the duration of the test in the next study should be extended to allow another validity testing method. *Cattell's Culture Fair Intelligence Test* could be used as a convergent criterion because it measures reasoning ability that aligns with *FIR*. On the other hand, a vocabulary test could be used as a discriminant criterion because theoretically, vocabulary ability is not related to FIR. Factor analysis could also be conducted among all sub-tests in *TISS*, especially to other narrow abilities in *Gf*.o test criteria validity, the *FIR* test could be correlated with a particular achievement test such as a math test because the *FIR* test had proven capable of predicting students' academic achievement. Finally, construct validity could be measured with a contrasted group method, such as comparing the *FIR* score from high and low quality schools or comparing two different school specializations (e.g., natural science vs. social science).

Item difficulty analysis showed that some items already met item target qualifications but that some others still needed to be improved although they had been made carefully based on the item complexity and number and type of rules (Carpenter, Just, & Shell, 1990; Primi, 2014). Further review showed that some of an item's features affected its difficulty, especially making it relatively easy. First, some items had parts that could be grouped so that the information

processing load was reduced; for example, similar colors and geometric shapes (Primi, 2014). Second, figural-spatial stimuli used were not quite abstract, so that made the item easier as well (Schulze, Beauducel, & Brocke, 2004). As a consequence, revisions in the stimuli color, shape, and abstraction were needed to enhance the item's difficulty level.

Item discrimination analysis showed that 50% of items had satisfying discrimination levels. Item discrimination was important because it contributed to test reliability (Guilford & Frutcher, 1978). Thus, revisions were still needed in order to increase the *FIR* test's coefficient of reliability. Lots of distractors were not working to deceive the test takers, increasing the possibility of guessing the correct answer randomly. This could especially affect the item difficulty index (Guildford & Frutcher, 1978). In other words, unselected distractors completely reduced alternatives, making the correct answer more predictable. Hence, revisions toward distractors were also needed.

There were some limitations of the study. The first is related to the sample size. A restricted sample number caused unstable results, especially in item analysis. Nunnally and Bernstein (1994) stated that the minimum sample size required to obtain more stable results is at least five to ten times the number of items, or about 200-300 participants. Thus, a large sample size is recommended for further study. Another limitation was in terms of participant characteristics. Participants came mostly from natural science specializations in their schools, so characteristics in this study tended to be similar. This is likely to have resulted in a selected group whose common characteristics might affect the test score variability. High-medium-low quality school categories were also not clearly determined and were still limited to Jabodetabek. Meanwhile, since the test would be used by all Indonesian high schools, a wider sample of student and school characteristics is needed to better represent all Indonesia high school students in any further research.

Based on development and psychometric testing, it can be concluded that the *FIR* test as part of *TISS* has the potential to be a new way to measure inductive reasoning in high school students. The reliability test showed that the *FIR* test was reliable and had a good internal consistency. Furthermore, the *FIR* test was valid for measuring inductive reasoning because it correlated significantly with the *SPM* test, which also measures inductive reasoning. Item analysis showed that despite the fact that some items had inadequate difficulty levels, they already had a satisfying item discrimination index. However, most of the distractors were still poorly deceiving participants.

Finally, to be used as one of the devices to test for student specialization placement, it would be necessary to revise and retest the items with a larger sample. Hence, the *FIR* test quality could be improved and an adequate norm could be developed as guidance for test score interpretation.

References

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

Blackwell, T.L. (2001). Test review. *Rehabilitation Counseling Bulletin*, *44*, 232-235.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measure: a theoretical account of processing in the raven progressive matrices. *Psych. Rev, 97*(3), 404-431 .

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

Cattell, R. B. (1987). *Intelligence: its structure, growth, and action*. New York: Elsevier Science Publishing Company. Inc.

Cohen, R. J., Swerdlik, M. E., & Sturman, E.D. (2013). *Psychological testing and assessment: An introduction to test and measurement*. NY: McGraw-Hill.
Croker, L. & Algina, J. (2008). *Introduction to classical and modern test theory* (2nd Ed.). Mason, Ohio: Cengage Learning.
Floyd, R. G., Evans J. J., & McGrew, K. S. (2003). Relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and mathematics achievement across the school age-years. *Psychology in the School*, *40*, 155-171, DOI: 10.1002/pits.10083
Friedenberg, L. (1995). *Psychological testing: Design, analysis, and use*. USA: Allyn and Bacon
Green, C. T., Bunge, S. A., Chiongbian, V. B., Barrow, M., & Ferrer, E. (2017). Fluid reasoning predicts future mathematical performance among children and adolescents. *Journal of Experimental Child Psychology*, *157*, 125-143.
Gardner, H. (2011). *Frames of mind* (3rd Ed.) New York: Basic Books.
Guilford, J. P. & Fruchter, B. (1978). *Fundamental statistics in psychology and education*. (6th Ed.). Singapore: McGrew-Hill Book. Co
Hunt, E. B (2011). *Human intelligence*. New York: Cambridge University Press.
Kaplan, R. M. & Saccuzzo, D. P. (2005) *Psychological testing: principles, applications, and issues.* (6th Ed.). Belmont: Wadsworth
Koller, O., Baumert, J., & Schnabel, K. (2001). Does interest matter? The relationship between academic interest and achievement in mathematics. *Journal for Research in Mathematics Education*, *32*, 448-470.
LaForte, E.M., McGrew, K.S., & Schrank, F.A. (2014). *WJ IV technical abstract (Woodcock-Johnson IV Assessment service bulletin no. 2)*. Rolling Meadows, IL: Riverside.
Laidra, K., Pullman, H., & Allik, J. (2006). Personality and intelligence as predictors of academic achievement: A cross-sectional study form elementary to secondary school. *Personality and Individual Differences*, *42*, 441-451.
Lawson, A. L., Banks, D. L., & Logvin, M. (2007). Self-efficacy, reasoning ability, and achievement in college biology. *Journal of Research in Science Teaching*, *44*, 706-724.
Little & Akin-Little. (2014). *Methods of academic assessment*. In Little & Akin-Little (Eds.) Academic assessment and intervention. Oxon: Taylor & Francis.
Madrasah Aliyah Citra Cendikia. (2015). *Placement tes for Specialization and Psychology (Tes Penempatan Jurusan dan Psikologi)*. Macitracendikia.sch.id/tes-penempatan-jurusan-dan-psikologi.
Malykh, S. (2017). The role of personality traits and intelligence in academic achievement of Russian high school students. *Procedia–Social and Behavioral Sciences*, *237*, 1304-1309.
Manual of Differential Ability Test (TKD) (n.d.). Retrieved from http://www.lpsp3.com/Manual-TKD-Lengkap.html?o=default
McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*, 1-10. doi:10.1016/j.intell.2008.08.004
Murphy, K. R. & Davidshofer, C. O. (2005). *Psychological testing: principles and applications.* (6th Ed.). New Jersey: Pearson Education
Neisser, U., et al., (1996). Intelligence: knowns and unknowns. *American Psychologist*, *51*, 77-108
Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory* (3rd Ed.) USA: McGraw-Hill, Inc.
Riandari, H. (2007). *Curriculum Model For High School Students (Model kurikulum tingkat satuan pendidikan (KTSP) SMA dan MA)*. Solo: PT. Tiga Serangkai Pustaka
Roth, B., Becker, N., Romeyke, S., Schafer, S., Domnik, F., & Spinath, F. M. Intelligence and good grades: A meta-analysis. *Intelligence*, *53*, 118-137.
Primi, R. (2014). Developing a fluid intelligence scale through a combination of rasch modeling and cognitive psychology. *Psychological Assessment*, *26*, 774-788. http://dx.doi.org/10.1037/a0036712
Sattler, J. M. (1992). *Assessment of children.* (3rd Ed.). California: Publisher. Inc.
Sehertian, E. (n.d). *The example of test item and answering key from general intelligence test* (*Contoh soal dan jawaban tes intelegensi umum)*. Retrieved from https://www.academia.edu/8497619/ Contoh_Soal_dan_ Jawaban_Tes_Intelegensi _Umum
Schneider, J. W. & McGrew, K. S (2012). *The Cattell–Horn–Carroll model of Intelligence*. In Flanagan, D. P. & Harrison, P. L. (Eds.). Contemporary intellectual assessment: theories, tests, and issues. (3rd Ed.). New York: Spring Street
Schulze, D., Beauducel, A. & Brocke, B. (2005) Semantically meaningful and abstract figural reasoning in the context of fluid and crystallized intelligence. Intelligence, *33*, 143-159. doi:10.1016/j.intell.2004.07.011
Sternberg, R. J. (1997). The concept of intelligence and its role in lifelong learning and success. *American Psychologist, 52*, 1030-1037. DOI: 10.1037/0003-066X.52.10.1030.
Super, D. E. & Hall, D. T. (1978). Career development: Exploration and planning. *Annual Review Psychology*, *29*, 333-372.
Taub, G. E., Keith, T. J.,, Floyd, R. G., & McGrew, K. S. (2008). Effects of general and broad cognitive abilities in mathematics achievement. *School Psychology Quarterly*, *23*, 187-198. DOI: 10.1037/1045-3830.23.2.187.
Wechsler, D. (1944). *The measurement of adult intelligence*. Baltimore: Waverly Press. INC,
Weinert, F. E., Helmke, A. & Schneider, W. (1989). *Individual differences in performance and in school achievement: Plausible parallels and explained discrepancies.* In Mandl, E. de Corte, N, Bennett, & H.F. Friedrich (Eds.) An instruction. Oxford: Pergamon Press.
Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R.W., Engle (Eds.). *Handbook of understanding and measuring intelligence*, 373-392.
Van der Ven, A.H.G.S, & Ellis, J. H. (2000). A rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences*, *29*, 45-64.