Advances in Social Science, Education and Humanities Research (ASSEHR), volume 135
1st International Conference on Intervention and Applied Psychology (ICIAP 2017)

Developing a New Quantitative Reasoning Test for Indonesian High School Students using the Cattell-Horn-Carroll (CHC) Theory of Intelligence

Khairunisa Damayanti[a] and 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

Developing a New Quantitative Reasoning Test for Indonesian High School Students using the Cattell-Horn-Carroll (CHC) Theory of Intelligence

This study aimed to develop a new intelligence measurement based on the CHC theory of intelligence; specifically, a quantitative reasoning test for tapping into high school students' interests. The test consists of two subtests: deductive quantitative reasoning (QR-VA) and inductive quantitative reasoning (QR-DA). There are 25 items of verbal arithmetic questions for QR-VA and 30 items of number series questions for QR-DA. Reliability analysis using Cronbach Alpha showed that QR-VA did not have good internal consistency, while QR-DA had good internal consistency. Validity analysis using correlation with *Tes Kemampuan Diferensial* (TKD) showed that both quantitative reasoning subtests are valid for measuring quantitative reasoning abilities. Furthermore, results from item analysis showed that the quantitative reasoning test has varying degrees of difficulty level and is able to discriminate quantitative reasoning abilities of high school students. For the final version of the test, 15 items were selected for QR-VA and 20 items were selected for QR-DA with appropriate degrees of difficulty. The within norm group with standard score of M = 10, SD = 3 were used as QR test norms.

Keywords: quantitative reasoning, CHC theory of intelligence, verbal arithmetic, number series

Introduction

One purpose of high school education in Indonesia is to prepare the learners to be able to develop their intelligence, talents, and interests that they can use graduating from the high school. Choosing a specialization in high school will help the students be more focused on developing their own interests and abilities (Minister of Education and Culture, 2013). Based on the Regulation of Education and Culture Minister Number 64 in 2014 about specialization in secondary education, specializing during high school is based on the student junior high school report card, a national exam, and a recommendation letter from a school counselor. However, these references may not be objective enough because they may encourage dishonest behavior that can affect the score results, both on the student report card in the national exam. Lestari and Asyanti (2015) performed research with some students via a questionnaire and found that junior high school students still displayed dishonest behavior. The questionnaire provided a case study and asked students for their responses. Then they classified the responses as either honest or dishonest behavior. The dishonest behavior reported was possible to do during daily tests or the national exam. Anies Baswedan, as the Minister of Education and Culture, said that there 80% deception still occurs by junior high school students in Indonesia (Alamsyah, 2015).

To place students accordingly requires the appropriate tools to successfully tap students' interests and abilities. Intelligence measurement is known to be a functional tool for placement in education (Aiken, 1997). It is also able to predict one's likelihood of academic success (Friendenberg, 1995). Unfortunately, the existing Indonesian intelligence measurement test has some weaknesses that can affect the inferences drawn from the test results. One of these is that the test version is out of date. In addition, the answers for the out-of-date version are available on the Internet because they have been leaked. To overcome this weakness, a new intelligence

measurement test based on contemporary theory needs to be developed. One famous theory of intelligence is the Cattell-Horn-Carroll (CHC) theory of cognitive abilities. The CHC theory is considered to be the most comprehensive theory (McGrew, et al., 1997 as cited in Nicholson, 2009) because it represents intelligence in three ways: g factor, broad abilities, and narrow abilities). Broad abilities are defined as persistent and basic rules and characteristics that affect much behavior in a particular domain (Carroll, 1993). Each broad ability has several narrow abilities. Narrow ability represents the specialization of ability that reflects one's experience or education (Carroll, 1993). We could say that broad ability is more common than narrow ability. Both broad and narrow abilities contribute to academic ability (Alfonso, Flanagan & Radwan, 2005).

There are 16 broad abilities in CHC theory, and one of them is fluid intelligence (Gf). McGrew (2014) defined fluid intelligence as "the use of deliberate and controlled focused attention to solving novel 'on the spot' problems that cannot be solved using prior knowledge." Fluid intelligence usually involves some combination of inductive and deductive reasoning to solve problems (McGrew, 2009). According to Flanagan and Dixon (2013), fluid intelligence is a mental operation used when a person faces a new task that cannot be solved automatically. A person with good fluid intelligence will be able to think, reason, and act quickly to solve various problems and use short-term memory successfully (Willis, Dumont & Kaufman, 2011).

In fluid reasoning, there are three narrow abilities: induction, general sequential reasoning, and quantitative reasoning. McGrew (2014) defined quantitative reasoning as "the ability to reason, either with induction or deduction, with a number or mathematical relations, operations, and algorithms." Agustin, Agustin, Brunkow, and Thomas (2012) add that the focus of quantitative reasoning is the practical application of mathematics to construct a quantitative argument in various contexts. This reasoning focuses on the application of numbers and mathematical concepts in daily problems. Flanagan and Dixon (2013) also report that quantitative reasoning is consistently related to mathematical ability applicable in everyday life. Although it depends on mathematical concepts, quantitative reasoning is related to other fields in everyday life besides mathematics. Deborah (2001, as cited in Elrod, 2014) stated that quantitative reasoning is an interdisciplinary ability that helps solve a problem. Based on these descriptions, this study aimed to construct a new quantitative reasoning ability test as a comprehensive battery of intelligence measurement. The goal of this study is to develop a quantitative reasoning test that has good psychometric qualities and includes good internal consistency, validity, a varying degree of difficulty, and is able to discriminate students' quantitative abilities. It also needs to have the appropriate norms to be used as an instrument to describe the quantitative reasoning ability in 10^{th} grade high school students.

Methods

Participant

The population of this study is 10th-grade high school students from all over Indonesia. However, in constructing the instrument, we used a sample based on accessibility. The sample was selected through non-probability sampling with an incidental sampling technique. In the try-out, 85 10^{th}-grade students from the public high school in Depok were used as a sample. The sample consisted of 50 female students (58.8%) and 35 male students (41.2%). The mean age

was 15 years 10 months. Each participant would take the whole QR subtest. In the field study, the sample consisted of 10th-grade high school students from the Jakarta public high school and had 98 students for the QR-VA subtest and 101 students for the QR-DA subtest. Participants who took the QR-VA subtest consisted of 55 female students (56.12%) and 43 male students (43.88%). Participants who took the QR-DA subtest consisted of 56 female students (55.45%) and 45 male students (44.55%). The mean age for QR-VA participants was 16 years old and for QR-DA participants was 15 years 11 months.

Measures

The quantitative reasoning instrument has two test types based on McGrew's definition (2014). The test type is based on what kind of reasoning is involved; that is, deductive or inductive reasoning. For each subtest, we used a number as a stimulus. In deductive reasoning, we provided facts and elicited an inference that could be considered to be true based on logic (Jacobs, 1982 as cited in Shadiq, 2007). So, for the deductive subtest, the participant is given a word problem and must solve it based on available information using mathematical operations and concepts. Here is an example of the test:

A truck carries 12 boxes of apples. Each box contains 30 apples. In the first shop, 5 boxes of apples are unloaded. How many apples remain in the truck? ("*Sebuah truk mengangkut 12 kotak apel. Setiap kotaknya berisi 30 buah apel. Di toko pertama, ada 5 kotak apel yang diturunkan. Berapa jumlah apel yang tersisa di mobil truk*?")

In inductive reasoning, the student must make a general inference based on specific statements (Shadiq, 2007). This thinking process uses an assumption and tests that assumption to learn the truth. In the inductive subtest, the participant is shown a series of numbers and participants must complete the series based on a pattern. Here is an example of the test:

3 5 9 15 23 33 …. ….

Because this is a test of maximum performance, it has various difficulty levels. Based on Loftus and Suppes (1972, as cited in Mayer, 1982), in deductive reasoning, the level of difficulty increases depending on whether the previous problem is solved correctly. This allows coverage of many different operations, complex sentence structures and relational propositions. Nesher (1986, as cited in Asrigantini, 1989) discussed three schemes: dynamic, static and relational, of which the relational scheme is the most difficult. Meanwhile, for the inductive reasoning test, Holzman, Pellegrino, and Glaser (1983, as cited in Lee & Worth, 2000) said that the difficulty level of the number series is based on the processing dimension and content knowledge dimension. The type of number, such as fraction or decimal, also can influence the difficulty. Every correct answer receives a score of one and incorrect answers receive a score of zero. For the number series problems, the answer will be considered correct if the participant answers all of the blank numbers correctly. The final score for each subtest equals to the sum of all he correct answers.

Procedure

The procedure to construct these QR subtests is based on Cohen, Swerdlik, and Sturman (2013). Their method consists of test conceptualization, test construction, expert judgment and legibility testing, test try-out, item analysis, and test revision. In the test conceptualization, we determined the instrument's objective, which is to use a quantitative reasoning intelligence subtest to measure 10th-grade high school students' ability to successfully perform quantitative reasoning. Quantitative reasoning can be used to indicate one of the recommendations during the selection process for specialization. Later, the quantitative reasoning intelligence subtest will be administered in groups. This quantitative reasoning intelligence subtest is divided into two formats: QR-VA for quantitative reasoning, which involves deductive reasoning and QR-DA for quantitative reasoning, which involves inductive reasoning.

In the next step, test construction, we start by selecting the type of test. This instrument is classified as a maximum performance test for which the objective is to see how well a person can perform (Cronbach, 1960 as cited in Klehe & Latham, 2008). Next, we select the item types. Each subtest uses numbers as the stimulus with a free response item type; that is, the participant has only to write the answer in a blank column. All answers are written in numerical form. To construct the instrument, we use McGrew's (2014) definition, which defines quantitative reasoning as the ability to reason, either with induction or deduction, with numbers or mathematical relations, operations, and algorithms. There are no indicators or dimensions stated in the definition. The target item for each subtest is different; there are 15 target items for deductive reasoning (QR-VA) and 20 for inductive reasoning (QR-DA).

After the items are created, they are all judged by experts and also by some target sample students through legibility testing. The experts include the educational psychologist, who provides suggestions about the number and sentence selection as related to face validity; and a psychometrist, who provided feedback about the construction, layout, and instruction of the items. Based on the results of the expert judgment, there were some items that were recommended for revision or even deletion (Anastasi & Urbina, 1997). For the QR-VA subtest, the expert judges provided advice about the word selection. Some items did not use the correct word, making them harder to understand. The revision for the QR-DA subtest was focused on the face validity, meaning means the number selection. Some sequences were using too large or complicated numbers, making the sequences too difficult. Along with this, there was also legibility testing with six junior high school students and 15 senior high school students from any Jakarta-Bogor-Depok-Bekasi city areas in the Indonesia Detabek area. The objective of legibility testing is to ensure that participants understand the instructions and all the questions. The results showed that all participants understood the instructions, but that many items were still too difficult for them and took a long time to answer. What caused the subtest to be too difficult was the numbers that were too large or too complicated. There were only a few sentences that some participants found confusing.

The items were revised based on the result of expert judgment and legibility testing. One of the examples of a QR-VA item is

A restaurant needs 45 liters of oil to fry 300 chickens. To fulfill an order of 125 chickens, how many liters of oil are needed? ("*Suatu restoran membutuhkan 45 liter minyak untuk menggoreng*

300 ekor ayam. Untuk memenuhi pesanan 125 ekor ayam, berapa liter minyak yang diperlukan?")

Most participants understood the case, but they felt this item was too difficult because of the large number selection. The expert judges did not have any advice about the word selection, so the revisions just focused on the number selection, and after these revisions, the number of participants who could answer the questions increased. The item was revised as follows:

A restaurant needs 40 liters of oils to fry 200 chickens. To fulfill an order of 125 chickens, how many liters of oil are needed? ("*Suatu restoran membutuhkan 40 liter minyak untuk menggoreng 200 ekor ayam. Untuk memenuhi pesanan 125 ekor ayam, berapa liter minyak yang digunakan?*")

For the QR-DA test, the revisions suggested were for the number selection and pattern of sequences. For example:

34 33 31 28 24 19 ….. …..

The pattern of this sequence is $U_{n+1} = U_n-(n+1)$, where n is the term order and U_n is the number based on the term order. The number is subtracted from the previous number with an increase in a pattern. The expert judges advised revision because the pattern was not properly identifiable. So, the item was revised with a similar but more identifiable pattern $U_{n+1} = U_n-(7-(n-1))$.

50 43 37 32 28 25 ….. …..

The pattern could be translated by subtracting by a number that decreased in order, starting from 7. After the revision, both of the subtests were tested with a group sample of high school students. There were 85 10th-grade students from two senior high schools in Depok. Based on the data gathered, we calculated the reliability and revised the items.

Data Analysis

In constructing the quantitative reasoning instrument, some psychometric testing was used. Cronbach's alpha is a reliability testing technique used to find an item's internal consistency. Another psychometry testing technique is various validity testing; this can include content sampling, face validity, or last construct validity. In construct validity, we examined the correlation between QR subtests and some of the Tes Kemampuan Differensial (TKD) subtests. TKD is one of the intelligence tests developed in Indonesia in 1969. TKD is still used today and it is known to be valid in measuring differential abilities (Widiawati, n.d). All validity testing techniques basically aim to ensure that the instrument really measures the subject's quantitative reasoning ability. So, in testing the validity of the quantitative reasoning subtest, we only used TKD subtests that were related to quantitative reasoning ability, which were TKD 5-R and TKD 6-R. Both subtests could represent deductive and inductive quantitative reasoning ability. The

revision of the items was based on both quantitative and qualitative item analysis. The last test constructs the norm based on group norms using a standard score technique.

Results

The data were collected at three high schools in Jakarta at different times in May 2017. All participants were 10th grade high school students. From the data collection, there were 98 students for the QR-VA test and 101 students for the QR-DA test. The students were also tested using the validation tests TKD 5-R and TKD 6-R. The duration for the entire test was divided into 20 minutes for QR-DA, 25 minutes for QR-VA, 7 minutes for TKD 5-R, and 10 minutes for TKD 6-R.

The result of the tests performed on 98 participants for the QR-VA subtest can be seen in Table 1. The mean values obtained from all participants were 15.918 and the deviation value from mean was 3.088. The variance was 9.54, indicating that the values obtained from all samples were quite diverse. During the process, participants were given a time limit to answer 25 verbal arithmetic questions. Sixty out of 98 participants were able to answer all of the questions completely. The fastest working time was 10 minutes 56 seconds, while the optimum time limit was 25 minutes, based on calculating the 75th percentile.

Table 1
The Description of QR-VA Participant

N	*Mean*	*SD*	*Variance*	*Max.*	*Min.*
98	15.918	3.088	9.54	24	6

Table 2 shows the test result for the QR-DA subtest. A total of 101 participants were tested for the QR-DA subtest, and they achieved an average value of 18.475. The amount of deviation between the data obtained and the average value was 3.6 SD. And the distribution of the value obtained by participants was quite diverse, with a variance value of 12.97. All participants in the QR-DA test were given time limits to complete 30 number series questions. A total of 36 out of 101 participants were able to complete all questions within the given time limit. Based on the time recording, the fastest working time was 13 minutes 30 seconds. The optimum time limit for answering all the questions as based on the 75th percentile was 20 minutes.

Table 2
The Description of QR-DA Participant

N	*Mean*	*SD*	*Variance*	*Max.*	*Min.*
101	18.475	3.601	12.972	27	9

Reliability testing using Cronbach Alpha revealed that the Alpha coefficient for QR-VA was 0.632. This means that 63.2% of observed scores were obtained from the true variance score and 36.8% was variance error, like content sampling and content heterogeneity. Meanwhile, for the QR-DA subtest, the reliability coefficient was 0.742, which means that 74.2% of observed scores were from the true score and 25.8% was variance error, like content sampling and content heterogeneity. Based on Kaplan and Sacuzzo (2013), only the items in the QR-DA subtest were consistent enough to measure a similar construct. SEM was calculated for both subtests as 1.136 for QR-VA and 0.994 for QR-DA.

Next, validity was tested using correlation with another test method. It was found that the QR-VA subtest correlated with TKD 5-R (r=0.360, p<0.01). It can be concluded that the QR-VA subtest provided a valid measurement for assessing quantitative reasoning ability involving deductive reasoning. It was also found that the QR-DA subtest correlated significantly with TKD 6-R (r=0.573, p<0.01). Therefore, it can be concluded that the QR-DA subtest provided a valid measurement for assessing quantitative reasoning ability involving inductive reasoning.

The item analysis was performed by analyzing item difficulty and item discrimination. This quantitative reasoning test was classified as a maximum performance test, so the item difficulty is important because there difficulty levels exist. The result from analysis of item difficulty revealed that the difficulty level for both quantitative reasoning subtests was relatively distributed from easy to moderate to difficult. The results for the QR-VA item difficulty analysis can be seen in Table 3. The results showed that some items had difficulty levels that were not distributed proportionately. That is, some items that were constructed as moderate difficulty level in fact had quite a lot of participants that were able to solve the items. Also, there were items that were constructed to have a high difficulty level that were classified as moderate level difficulty items.

Table 3
The Item Difficulty Analysis of QR-VA subtest

Item Difficulty Index	**Difficulty Level**	**Item Number**	**∑Item**	**%**
$p \geq 0.8$	Very Easy	1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 15	11 item	44
$0.6 \leq p < 0.8$	Easy	9, 10, 14, 20	4 item	16
$0.4 \leq p < 0.6$	Moderate	13, 14, 16, 17, 18, 21, 23	7 item	28
$0.2 \leq p < 0.4$	Difficult	19	1 item	4
$p < 0.2$	Very Difficult	22, 24, 25	3 item	12
	TOTAL		**25 item**	**100**

The item difficulty analysis was also completed for the QR-DA subtest, as seen in Table 4. Similar to the result from the QR-VA subtest, the item difficulty level of QR-DA was not distributed evenly. There were items that were initially constructed to have moderate level difficulty, but after trying it out with participants, the items were too easy to answer. Similar problems occurred with items at the moderate and high difficulty levels.

Table 4
The Item Difficulty Analysis of QR-DA subtest

Item Difficulty Index	Difficulty Level	Item Number	∑Item	%
$p \geq 0.8$	Very easy	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	14 item	46.67
$0.6 \leq p < 0.8$	Easy	16, 17, 18, 19	4 item	13.33
$0.4 \leq p < 0.6$	Moderate	15, 22, 23	3 item	1
$0.2 \leq p < 0.4$	Difficult	20, 24, 25, 28	4 item	13.33
$p < 0.2$	Very difficult	21, 26, 27, 29, 30	5 item	16.67
	TOTAL		**30 item**	**100**

Next, the item discrimination was tested to find out whether the quantitative reasoning subtest was able to discriminate between participants with high quantitative reasoning abilities and participants with low quantitative reasoning abilities. A correlation indices method is commonly used in item selection (Anastasi & Urbina, 1997). According to Nunnally and Bernstein (1994), items with a correlation index (Crit) above 0.2 classify the item as having good discrimination ability. Table 5 in the Appendix shows the results of the discrimination analysis for the QR-VA subtest. There were 12 items that had a good ability to discriminate between participants with high and low abilities.

Table 5
The Item Discrimination Analysis of QR-VA subtest

Cr_{IT} Index	Explanation	Item Number	%	
>0,2	Item has a good discrimination ability	10, 12, 13, 14, 15, 16, 17, 19, 21, 22, 23, 25	12 item	48
<0,2	Item has a poor discrimination ability	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 18, 20, 24	13 item	52
	TOTAL		**25 items**	**100**

Meanwhile, Table 6 describes the results for the QR-DA subtest. For the QR-DA subtest, there were 15 items identified as having good discrimination ability. The other items were classified as having poor discrimination ability.

Table 6
The Item Discrimination Analysis of QR-DA subtest

Cr_{IT} Index	Explanation	Item Number		%
>0,2	Item has a good discrimination ability	13, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 27, 28, 29, 30	15 item	50
<0,2	Item has a poor discrimination ability	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 21, 26	15 item	50
	TOTAL		**30 item**	**100**

To improve the test quality, some items then underwent revision and/or deletion. The QR-VA subtest became the focus of revision due to its low value of reliability. The revision included sentence or number revision and their relationship to the everyday life problems. Based on the analysis of item discrimination and item difficulty, 15 QR-VA items and 20 QR-DA items were obtained with the appropriate distribution of difficulty level. Some revisions included sentence or number revision for a particular item (see the detail in Table 7 of the Appendix). Below is the example of a revised item:

The total number of Reza's and Budi's toys is 23 pieces. The number of Budi's toys is 7 more toys than Reza's. How many toys does Reza have? (*"Total jumlah mainan mobil-mobilan Reza dan Budi adalah 23 buah. Jumlah mobil-mobilan yang dimiliki Budi lebih banyak 7 buah jika dibandingkan Reza. Berapakah jumlah mainan mobil-mobilan milik Reza?"*)

This item should be revised because the word "Reza" was repeated too often and confused the participants. Also, this item was considered to be a moderate level item and needed to be revised as more difficult. It was decided that having a larger number would increase the level of difficulty. The item was revised as follows:

The total number of Reza's and Budi's toys is 31 pieces. If compared with Reza's, Budi has 13 more toys. How many toys does Reza have? (*"Total jumlah mainan mobil-mobilan Reza dan Budi adalah 31 buah. Jika dibandingkan dengan milik Reza, jumlah mainan mobil-mobilan Budi lebih banyak 13 buah. Berapakan jumlah mainan mobil-mobilan milik Reza?"*)

After we obtained a number of items according to the target, we did the final testing using the same methods as in the first testing.

The reliability testing for the QR-VA subtest was done using Cronbach Alpha and the *alpha coefficient* of QR-VA was 0.644. This means that a 64.4% observed score was obtained from the true variance score and 35.6% was a variance error such as content sampling or content heterogeneity. For the QR-DA subtest, the reliability coefficient was 0.732. This means that a 73.2% observed score was obtained from the true variance score and 26.8% was a variance error such as content sampling or content heterogeneity. There was an increase in reliability coefficient for both subtests, but only the QR-DA subtest items were consistent in measuring one

construct. The SEM value for each subtest was 0.919 for the QR-VA subtest and 0.865 for the QR-DA subtest.

The validation testing was done by using the same method as in the first testing; that is, correlation with another test. The QR-VA subtest was found to correlate with the TKD 5-R (r=0.388, p<0.01). From this, we can infer that the QR-VA subtest was valid for measuring quantitative reasoning ability involving deductive reasoning. For the QR-DA subtest, a significant correlation was found with TKD 6-R (r=0.065, p<0.01), from which we can conclude that the QR-DA subtest was valid for measuring quantitative reasoning ability involving inductive reasoning.

After finalizing the number of items, the norm was created using the within group norms method with a standard score technique. A normality test was completed using the Kolmogorov-Smirnov technique for both subtests. The results from normality testing showed that the data was not normally distributed, so linear transformation was performed. The standard score of mean = 10 and SD = 3 were used to construct the norm due to the distribution pattern of the standard score that did not differ greatly from the original data. The selection of mean and SD are based on the spreading out of a standard score that does not differ greatly from the original data.

Table 7
The Norm of Quantitative Reasoning Subtest

Standard Score **QR-VA (M = 10, SD = 3)**	**Raw Score QR-VA**	**Raw Score QR-DA**	***Standard Score*** **QR-DA (M = 10, SD = 3)**
20	15	19–20	20
19	-	-	19
18	-	-	18
17	-	-	17
16	13–14	-	16
15	12	18	15
14	11	16–17	14
13	10	15	13
12	-	14	12
11	9	13	11
10	8	12	10
9	7	11	9
8	6	10	8
7	5	9	7
6	-	8	6
5	4	7	5
4	3	6	4
3	1–2	4–5	3
2	-	-	2
1	-	-	1
0	0	0-3	0

Discussion

The reliability was tested using Cronbach Alpha and showed that the QR-VA subtest was not reliable. The low-reliability coefficient indicated that the variance error in QR-VA was higher than in QR-DA. The possible error variance when using Cronbach Alpha is due to content sampling and heterogeneity content. The QR-VA subtest included verbal arithmetic questions; solving that kind of question requires not only quantitative reasoning ability, but also verbal reasoning ability (Purwanto, 1997). Therefore, the item in the QR-VA subtest may indeed not fully measure a particular ability, but there is a possibility that it measured other abilities.

Another factor that may affect the reliability is the variability of the sample group (Anastasi & Urbina, 1997). The more heterogeneous a group sample, the greater the variability obtained. In both quantitative reasoning subtests, the sample group was the 10th-grade high school students in South Jakarta. Sampling was done in just one area, so it can be said that the sample was homogenous. Anastasi and Urbina (1997) also mention that using a greater sample number will be more likely to obtain a variance. Another thing that affects reliability is the test administration (Cohen & Swerdlik, 2009). The quantitative reasoning subtest was tested together with another subtest, so there can be an impact in fatigue, especially before break hours or after school.

The validity test was done using construct validity and correlation with the TKD test technique. Results obtained show that the two subtests are valid measurements for assessing or measuring the same construct with the validation test. The correlation coefficient was also moderate, which indicated an appropriate value for the newly developed test. A high coefficient might indicate that the new test was measuring the same construct with the previous instrument; thus it did not need a new instrument (Anastasi & Urbina, 1997).

The norm was done using a standard score technique. Based on normality testing, the data was not normally distributed, so linear transformation was performed first. In constructing an instrument that is administered in a group, it usually based on testing 100.000 participants (Aiken, 1997). This number is significantly higher than the number of samples obtained in this study. Anastasi and Urbina (1997) suggest that the sample used should be represented by making a distinct criterion of the sample. In the quantitative reasoning instrument testing, the sample is 10th-grade high school students. Samples still need to have distinct criteria; for example, coming from public or private school, or a school with a particular accreditation.

There is another limitation in the testing for this study. First, the sample was less heterogeneous due to time constraints and accessibility. The sample try-out had participants from only two public schools in Depok; it should be done in other cities like Bogor, Tangerang, and Bekasi. The same thing also happened in field testing, with only participants from Jakarta included in the sample. This result cannot be generalized into a large population and affect to reliability, item difficulty index, and norm. Second, the test administration was less standardized than it should be. In field testing, different conditions are encountered, such as how the instructor presents the directions, and the length of the test time. Furthermore, there were other subtests performed at the same time as the quantitative reasoning subtests, which could affect participant fatigue.

For further research and development, the testing should obtain a more heterogeneous sample so that it can represent the population more accurately. The sampling technique should be

considered as it will help to obtain a more representative sample. Another important thing would be to standardize the test administration. This could be accomplished by providing standardized instructions and standardized sequences for the test administration. Lastly, the test construction should be evaluated and revised based on the error source, especially for the QR-VA. These question types involve not only the quantitative reasoning ability, but also the verbal reasoning ability. Carroll (1993) argues that to learn one's ability for quantitative reasoning, the item should be simple and easily understood. So, in constructing the items for the QR-VA, it needs to involve other experts, such as a mathematics teacher and a linguist to gain various perspectives.

References

Agustin, M. Z., Agustin, M., Brunkow, P., & Thomas, S. (2012). Developing quantitative reasoning: Will taking traditional math courses suffice? An empirical study. *The Journal of General Education*, *61*(4), 305-313.

Aiken, L.R. (1997). *Psychological testing and assessment*. (9th ed.). Boston: Allyn and Bacon.

Alamsyah, I. E (2015). 80 persen SMP negeri di Indonesia lakukan kecurangan UN. Diperoleh dari http://www.republika.co.id/berita/pendidikan/eduaction/15/06/11/nprw27-80-persen-smp-negeri-di-indonesia-lakukan-kecurangan-un. Diakses pada 17 Mei 2017.

Alfonso, V. C., Flanagan, D. P., & Radwan, S. (2005). The impact of the Cattell-Horn-Carroll theory on test development and interpretation of cognitive and academic abilities. *Contemporary intellectual assessment: Theories, tests, and*, (2nd), 185-202.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, N.J.: Pearson Prentice Hall.

Asrigantini. (1989). *Peranan pemahaman verbal dan jenis soal dalam pemahaman aritmetika*. Depok: Universitas Indonesia.

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

Cohen, R. J., & Swerdlik, M. E. (2009). Psychological testing and assessment: An introduction to tests and measurement (7th ed.). 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.

Elrod, S. (2014). Quantitative reasoning: The next "across the curriculum" movement. Retrieved from https://www.aacu.org/peerreview/2014/summer/elrod

Flanagan, D. P. & Dixon, S. G. (2013). The Cattell-Horn-Carroll Theory of Cognitive Abilities. *Encyclopedia of special education*.

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

Friyatmi, F. (2011). *Faktor-faktor Penentu Perilaku Mencontek di Kalangan Mahasiswa Fakultas Ekonomi UNP. TINGKAP*, *7*(2).

Minister of Education and Culture. (2013). Pedoman peminatan peserta didik. [PDF document]. Retrieved April 12, 2017, from http://bk.fip.uny.ac.id/sites/psikologi-pendidikan-bimbingan.fip.uny.ac.id/files/PEDOMAN%20PEMINATAN%20SMA-SMK.pdf

Klehe, U. C., & Latham, G. (2008). Predicting typical and maximum performance with measures of motivation and abilities. *Psychologica Belgica*, *48*(2-3), 67-91.

Lee, F. L., & Heyworth, R. (2000). Problem complexity: A measure of problem difficulty in algebra by using computer. *EDUCATION JOURNAL-HONG KONG-CHINESE UNIVERSITY OF HONG KONG-*, *28*(1), 85-108.

Lestari, S., & Asyanti, S. (2015). *APAKAH SISWA SMP BERPERILAKU JUJUR DALAM SITUASI ULANGAN?*. In *PROSIDING SEMINAR NASIONAL & INTERNASIONAL.*

Mayer, R. E. (1982). Memory for algebra story problems. *Journal of educational psychology*, *74*(2), 199.

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

McGrew, K. S. (2014). Cattell-Horn-Carroll (CHC) theory of cognitive abilities definition. Retrieved March 13, 2017, from http://www.iapsych.com/chcdefsbrief.pdf.

Minister of Education and Culture Republic of Indonesia. (2014). Peraturan Menteri Pendidikan dan kebudayaan Republik Indonesia nomor 64 Tahun 2014. [PDF document]. Retrieved June 4, 2017, from http://pendis.kemenag.go.id/pai/file/dokumen/SisdiknasUUNo.20Tahun2003.pdf.

Nicholson, K. J. (2009). Use of Cattell-Horn-Carroll Specific Cognitive Abilities to Enhance Prediction of Reading on the Third Grade Pennsylvania System of State Assessment (Doctoral dissertation, Indiana University of Pennsylvania).

Purwanto, B. E. (1997). *Pengaruh kemampuan bernalar secara verbal terhadap kemampuan memecahkan masalah hitungan cerita.* Depok: Universitas Indonesia.

Shadiq, F. (2007). *Apa dan mengapa matematika begitu penting.* Pusat Departemen Pendidikan Nasional. Direktorat Jenderal Peningkatan Mutu Pendidik dan Tenaga Kependidikan. Yogyakarta: Pengembangan dan Pemberdayaan Pendidik dan Tenaga Kependidikan (P4TK) Matematika.

Widiawati, D. (n.d.). Tes kemampuan diferensial. [PDF document]. Retrieved June 3, 2017, from http://modul.mercubuana.ac.id/files/pbael/pbaelmercubuanaacid/Modul%20Backlink/Modul%20Genap%202010-2011/Fakultas%20Psikologi/Diah%20Widiawati%20 Psikodiagnostik%20V/ModulPsikodiagnostik5GP1011TM10.pdf.