Development and Validation of a Working Memory Capacity Test for High School Students

Debby Mardianti[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 and Validation of a Working Memory Capacity Test for High School Students

The aim of this study is to develop a working memory capacity (*WM*) test as one of the subtests for the new intelligence test (*TISS*, or *Tes Inteligensi Siswa SMA*). The *WM* test would be used as a reference in consideration of selecting a specialization for high school students in Indonesia. The construction of the WM test is based on the Cattell-Horn-Carroll (CHC) theory, which is the most comprehensive and contemporary theory of intelligence. The testing of the *WM* test was conducted on 97 first-year high school students in the Jakarta region, including students from natural science and social science specializations. Based on the item analyses, the *WM* test had good items with a varying item difficulty and good item discrimination. Reliability testing using the Cronbach Alpha method showed that the *WM* test had a good internal consistency. Validity testing using correlations with other test methods showed that WM test was a valid test for measuring working memory capacity. Based on the results, the WM test is a promising new test in Indonesia for measuring one type of intelligence: that is, working memory capacity based on the CHC theory.

Keywords: CHC theory of intelligence, high school, intelligence test, working memory capacity

Introduction

The major in SMA/MA, currently known as specialization, is one of the determinants for students' careers, since the selected specialization becomes a stepping stone for the direction students will take in further study and in their future careers. Generally, students' specialization in high school influences their major or faculty that they will pursue in university. Therefore, specialization in high school is an important decision that must be made carefully. Selecting a misplaced specialization in high school can cause difficulties if followed through to studying in college. Specialization also provides students with the opportunity to determine the skills they want to improve on in certain subjects based on their basic capabilities, including intelligence, talent, interests, and choices (Kementerian Pendidikan dan Kebudayaan, 2013).

Ideally, the selection process for specialization is completed by considering several factors, including academic and non-academic achievements. These include report cards, the national exam (*UN*), student interest, and student ability (Kementerian Pendidikan dan Kebudayaan, 2013). Evaluation of student ability can be assessed through psychological tests. However, the selection of specialization is usually only based on the student report cards and student request (Widowati, 2015). Both of these aspects do not sufficiently represent student interests, talents, and abilities. The report card is determined by teacher or school using different standards at each school. In addition, using of the report card for the basis of specialization is inappropriate because report cards include reference to the complete learning system applied to all schools, so it includes reporting on remedial learning (Direktorat Pembinaan SMA, 2010). Remedial results listed on a student's report card obviously cannot be claimed as the students' true abilities.

Student specialization should rely on results of actual student abilities. Therefore, objective

reports on students' ability were not only based on student reports. The use of psychological tests; namely, intelligence tests was one good alternative for getting a more detailed description of student ability. The intelligence test is one of the commonly used tests for evaluating learning development and making decisions related to educational placement (Friedenberg, 1995). Intelligence can be said to be a clue about students' learning potential, so the results of the intelligence test can be used to refer students into suitable specializations. In addition, decision-making on the specialization selection would be more accurate if taking student intelligence into consideration. This is intended so that students proceed through the learning process without barriers and obtain at least a satisfactory achievement (Rufaidah, 2015).

Since 1905, the intelligence test has been widely developed, and it contributed to the emergence of intelligence tests from many other experts (Cohen, Sturman & Swerdlik, 2013). The development of other intelligence tests was based on several rationales from different intelligence theories. Broadly, there were two basic ideas about the theory of intelligence. One theory believes that intelligence consists of a single capacity, while others believe that intelligence consists of many abilities (Pal, Pal, & Tourani, 2004). The theory of contemporary intelligence that has maintained the strongest and most significant influence on the measurement of human intelligence these days is the Cattel-Horn-Carroll (CHC) theory (Flanagan & Harrison, 2005). The CHC theory was known as an influential theory because it is the most comprehensive one; it is supported by psychometric theory as it relates to cognitive structure and empirical academic ability. The CHC theory was supported by empirical data (Flanagan & Harrison, 2005). In addition, CHC theory is comprehensive, summarizing and integrating human cognitive abilities (McGrew, 2005). CHC theory also provided the basis for the formation of many other intelligence tests developed today, both new and revised from earlier versions (Flanagan & Harrison, 2005; Dehn, 2008). Therefore, CHC theory was summarized as the basis for measurement of human intelligence capability (Schneider & McGrew, 2012).

Many intelligence tests have been developed abroad based on CHC theory, including WJ-R (Woodcock Johnson-Revised) and WPPSI-III (Wechsler Preschool and Primary Scale of Intelligence-III). In Indonesia, the development of intelligence tests based on CHC theory is still limited. The current study is being conducted to develop one subtest of the new intelligence test based on CHC theory that will be used for the purpose of selecting a specialization in high school. The development of this intelligence test will consist of measuring several abilities that are required for high school learning and will be associated with Broad ability and Narrow ability as they are known in CHC theory (Flanagan & Dixon, 2013).

One broad ability in CHC theory that is relevant to learning needs in high school is short-term working memory (Gwm). Gwm describes the ability to receive and process information; most of the skills required in SMA require good information processing. There are three narrow abilities in Gwm: memory span, working memory capacity, and attention control (Flanagan & Dixon, 2013). However, one narrow ability with an important role that is required in the student learning process is working memory capacity. Working memory capacity is an important aspect because information processing involves the capability of accessing working memory (Dhen, 2008). Working memory is a dynamic memory storage system, functioning to manipulate and retain new incoming information (McGrew, 2005). The learning process in individuals is highly dependent on working memory (Dhen, 2008). Everything one learns and remembers uses

working memory to do so (Davis, 2011). In addition, working memory affects classroom performance in specific subjects, such as reading comprehension and math (Dhen, as cited in Davis, 2011). Working memory is also a good predictor of national assessment results in English, mathematics and science subjects (Tariq & Noor, 2012). Therefore, working memory is recognized as having a strong relationship with academic learning and higher level cognitive function (Dehn, 2008). Working memory is also a predictor of individual academic performance (La lopa & Holich, 2014). Thus, we can conclude that working memory is important for high school students to possess, and that its use will be required during the learning process. The measurement of working memory capacity during the specialization process could be the predictor of students' success in the interest that he will choose to pursue. Therefore, the measurement of working memory to assist students in determining their specialization is very important and should be taken into account.

In the current study, we developed a working memory capacity (*WM*) test as one of the subtests for the new intelligence tests developed based on CHC theory, which are called *TISS*, or *Tes Inteligensi Siswa SMA*. The development of the *TISS* and *WM* tests should be able to trace high school students talent and used for selection specialization in high school students. Working memory can also be measured using two types of information: verbal and nonverbal (visuospatial) (Diamond, as cited in Filgueiras, 2016). In this study, the *WM* test will be administered in visual form with nonverbal responses and administered in groups. This form was chosen to meet the needs of the test purpose in the school setting.

To ensure that the *WM* test meets the requirements of a good psychological measurement tool, psychometric testing will be performed through reliability testing, validity testing, and item analysis. Reliability testing was conducted to ensure that the *WM* test is a reliable test in the sense of having good internal consistency. Validity testing was conducted to ensure that the *WM* test is a valid test for measuring the memory construct. Finally, item analysis was conducted to ensure that the WM test has varying degrees of difficulty and a good discriminating index to be able to distinguish individuals with high and low levels of *WM* ability.

Methods

Participants

The participants of this study were 97 high school students of class X in Jakarta, Bogor, Depok, and Tangerang (Jabodetabek) areas. The samples were taken from the leading and non-superior school category in Jabodetabek. Participants came from natural science and social science specializations. The ages of participants ranged between 15 and 16 years (M = 15.75, SD = 0.43). The proportion of male and female participants was quite balanced. In addition, most of the participants were from the natural science specialization (64.6%). The sampling method used was non-probability sampling or convenience sampling.

Measures

The *WM* test is categorized as a maximum performance test, administered in groups, and in paper and pencil. Based on the time limit, the WM test is categorized as a timed-power test. The *WM* test consists of three indicators that showed to what extent students were able to maintain

attention on the provided stimulus, ignore distractions, not be disturbed by any interference, and able to recall information stored in secondary memory. Based on these indicators, the final items in the WM test would be up to 15 items and 25 items decided at the pooling items stage. Each item consisted of a stimulus in the form of numbers and words. Participants were asked to arrange the stimulus starting with the smallest array of numbers and then the words in alphabetical order. The degree of difficulty in the WM test was determined by the number of stimuli and syllables in each item and the initial letter on the word stimulus in each item. Time limits were differentiated by two stages: the presentation of the item and recall of the item. The item presentation describes when the item was displayed in sequence to the participants. The recall item was the stage in which participants remembered the stimulus and wrote it on the answer sheet. The time limit for the item presentation was determined by the amount of stimulus within each item.

The scoring technique used in the WM test was to provide a score for each item that was successfully remembered and written down by participants. Each item would be scored 1 for all stimulus numbers and words properly written and score 0 for stimulus numbers and words that were not written down successfully. The total score earned by the test taker was the total number of item scores that were answered correctly. The range of scores obtained by test participants in the WM tests ranged from 0 to 15. The final scores showed that the higher the score obtained, the higher the capability of the individual's working memory.

Procedure

The process for developing the *WM* test included item pooling, expert judgment, readability, and try-out. Expert judgment was performed to ensure that all of the items represented the indicator and met the criteria of good items. Based on the expert judgment, some revisions were made to the item content, especially on the word stimulus used. After that, we also conducted a readability process with some high school students in the Jabodetabek area to ensure that items were understood and to determine the time limit for the *WM* test. Then, the try-out test was conducted with 83 participants at three schools in the Jabodetabek area. The results showed that the WM test had a good internal consistency to measure a construct. In addition, most of the items in the WM test were classified as easy items and were still not distributed in accordance with the item difficulty. The analysis of the item discrimination showed that most of the items on the WM test had good item discrimination. Some revisions were made to the *WM* item test before we conducted a field testing.

Data Analysis

The quantitative item analysis was conducted by analyzing the item difficulty level (p) and item discrimination with the corrected total correlation (cr_{IT}) method. The best discriminating item was the item with $cr_{IT} \geq 0.3$ (Nunnally& Bernstein, 1994). Reliability testing of the *WM* will be conducted using the Cronbach Alpha method. Reliability coefficients that are considered good are equal to 0.70–0.80 (Kaplan and Saccuzzo, 2013). In addition, the Standard Error of Measurement (*SEM*) will be calculated to find out how far the obtained score deviated from the true score. Furthermore, a construct validity testing for *WM* tests would be conducted by using correlation with other test methods. The *WM* test would be correlated with *IST* (Intelligenz Structure Test) as the criterion.

Results

Table 1 displays the description of working memory capacity in the *WM* test.

Table 1
Description of *Working Memory Capacity*

Descriptive	
Minimum Score	5
Maximum Score	25
Mean	17.44
Standard Deviation	4.79
Time limit of item presentation	
Easy items	5”
Moderate items	7”
Difficult items	10”
Time limit of item recall	
Easy items	10”
Moderate items	20”
Difficult items	25”

Note: Σitem=25; N=96

The result showed that the *WM* test was classified as an easy test (M = 17.44, SD = 4.79). The average value obtained indicated that about 70% of participants were able to answer the items correctly in the *WM* test. In addition, the time limit for item presentation and the time limit of item recall in the specified *WM t*est represented optimum time limits. From observations made during the test, almost all participants were able to complete each item in the *WM* test within the given time limit. Seventy-five percent of participants completed all the items within the time limit.

The reliability testing for the *WM* test showed a reliability coefficient of $\alpha = 0.844$. This result indicates that the *WM* test had a good internal consistency for measuring a particular construct. The reliability coefficient of 0.844 indicates that the 84.4% variance of the observed score was the true score variance, and 15.6% was the error variance derived from the content sampling and content heterogeneity errors. Furthermore, the SEM value of the *WM* test was obtained at 1.86. The magnitude of the SEM value shows the average index of the number of measurement errors in the test score.

Validity testing of the *WM* test using the IST test subtest ME as criteria yield a Pearson correlation index of $r = 0.281$, $n = 97$, $p < 0.05$. Thus it can be stated that the *WM* test was valid for measuring the working memory capacity construct because it correlated significantly with IST test subtest ME which also measured the memory construct. There was a 7.9% shared variance between the *WM* test and the IST test.

Item difficulty analysis for the *WM* test showed that most of the items were categorized as easy items. Based on the test results, several items were not in accordance with the item specification. The final test should have consisted of seven easy items (numbers 1–7), 11 moderate items (numbers 8–18), and seven difficult items (numbers 19–25). Unfortunately, the results showed that most of the items fell into the easy category.

Analysis of the item discrimination index for the *WM* test indicated that most items in the *WM* test had a good discriminatory power. There were19 items (76%) that already had values of $cr_{IT}>$

0.3. However, the low discriminating power of these items was related to their difficulty level. Item that had a cr_{IT} value below 0.3 were relatively easy items. For those items, most participants were able to answer them, so the items will not have a good discriminating power. It can be concluded that poor discrimination items were related to their level of difficulty.

Furthermore, an integrative analysis was conducted to select the final items to meet the item specification on the *WM* test. The criteria used in selecting the items were qualitative and quantitative item analysis. Qualitative item analysis was based on the content and form of the item, while the quantitative item analysis was based on item difficulty (p) and item discrimination (cr_{IT}). Based on integrative item analysis, we selected the 15 best items for the *WM* test. Of the selected items, eight met the criteria for a good item, with seven items that should still be revised. Item revision in item content was needed primarily to increase the level of item difficulty.

After we selected the 15 items, a re-calculation of reliability and validity testing was conducted. Reliability test using Cronbach Alpha returned an increasing reliability coefficient of the *WM* test to 0.794. Furthermore, validity testing of the *WM* test also showed an increasing correlation coefficient with IST test subtest ME ($r = 0.303$, $n = 97$, $p < 0.05$). Thus, it can be concluded that the *WM* test had a good internal consistency for measuring a particular construct and validity for measuring the working memory capacity construct.

Discussion

The current study aimed to develop and validate a new subtest of intelligence test, called the *WM* test, to assist in making recommendations for high school student specialization selection. The result showed that the *WM* test is a promising subtest in measuring working memory, which is needed to learn successfully in the high school setting.

The field testing showed that the *WM* test was classified as an easy test. This may have been related to the sample characteristic and the item content. The sample used in the *WM* test came from the Jakarta area with a good learning quality. Thus, it was assumed that most of the participants in this study had a good working memory capacity, so that affected the degree of difficulty in the *WM* test. In terms of item content, the difficulty of the *WM* test was made based on the length of the items list. However, based on the tests performed it, it did not seem to affect the level of difficulty of the *WM* tests. The result showed that the difference in a difficult and easy item lies in the use of words that are encountered often in everyday life. Thus, it can be concluded that there is an influence of familiarity on the item difficulty level. Moreover, concrete and abstract words also affected the item difficulty; students recalled concrete words more easily than abstract words (Yui, Ng & A, 2017). Thus, it is necessary to consider word familiarity as well as the use of abstract words when creating the degree of difficulty for a *WM* test.

In addition, the degree of item difficulty was also influenced by the variation of the numbers used. In this study, we only used numbers 1–11, and for moderate items, only number 8–11. This meant that the number variations were smaller, so there were items with the same number that made it easier for the individual to remember the stimuli. Therefore, scrambling the numbers used should be considered, as well as involving numbers 1–20 randomly without taking the

number of syllables in the numbers into account. In addition, in relation to the distraction of assigned tasks, recalling items as related to alphabetic and numerical arrangements was an integral process of task and stimulus provided. Therefore, the distraction could be not effective because it was part of the information processing. Creating a distraction task for the *WM* test that is not part of the information processing on the main task is needed, such as performing a countdown operation before the item for recall is provided (Cowan, 2010). From this result, we also found that both the time limit for item presentation and item recall were optimum for the *WM* test. Our observations during the test showed that almost all test takers were able to complete the *WM* test within a specified time limit.

Based on the results of the reliability test using the Cronbach Alpha method, the *WM* test had a good internal consistency. However, the purpose of the *WM* test was to provide recommendations to high school students for determining the most suitable specialization for each student based on his/her ability. For purpose of the test, we needed a higher reliability coefficient (Little & Akin-little, 2014). Thus, it was necessary to improve the *WM* test reliability. Reliability of the *WM* test could be enhanced through the use of more heterogeneous samples drawn from multiple regions and more diverse schools to get a more varied distribution of students' ability.

Based on the result of validity test, we found that the *WM* test was valid for measuring working memory because it correlated significantly with the IST test, which measures the same construct. The relatively low degree of correlation between the *WM* test and the ME subtest could be due to the type of memory measured by both tests. Although both tests measured the same construct (memory), the types of memory measured were different. The *WM* test measured working memory, whereas the ME subtests measured long-term memory, which is related to verbal memory. Thus, the correlation index obtained was not very high.

Item analysis on the *WM* test showed that there weresome items that needed to be revised. The revision were needed to increase the item difficulty level, because the WM test still did not have the appropriate proportion in terms of its level of item difficulty. Most of the items were easy items that can lead to a low discriminating power, because it could be answered by all participants. The degree of item difficulty was also influenced by the capability of the samples used in the *WM* test.

Furthermore, an important factor affecting the overall test results was related to the process of item presentation. The items on the *WM* test were presented visually by displaying a number of simultaneous stimuli. The process of remembering the working memory began by including visual information through the visuospatial sketchpad role: first remembering, then processing the information in the working memory (Cowan, 2010). The process that occurred in the *WM* test was by preparing item stimulus. However, the presentation of a *WM* test item by displaying all the item stimuli in the same time could actually make the stimulus process possible immediately or at least before entering information to remember. This is because the individual could see the whole stimulus in the item. With the instructions delivered in the beginning to arrange the item, the individual had a tendency to enter the information with the correct stimulus arrangement. It certainly affected the overall working memory task and was not in accordance with the process of remembering the information referred to in the *WM* test. Therefore, it was necessary to change

the item presentation in the *WM* test. Changes in the item presentation can be done by displaying the stimulus on the *WM* test items one at a time. Thus, the individual would not see all the stimuli on the item and could not perform the stimulus at the same time. The process of information processing could only be done if each item was displayed and then remembered by the individual.

Based on psychometric tests that were conducted on the *WM* test, it can be concluded that the *WM* test was a reliable test, which means it had a good internal consistency for measuring a particular construct. Furthermore, the *WM* test was valid for measuring the working memory capacity construct because it correlated significantly with the IST test subtest ME, which also measures the memory construct. Based on the analysis of the degree of item difficulty, it can be concluded that most of the items in the *WM* test were found to be relatively easy. This is still not in accordance with the degree of difficulty levels that should consist of a combination of easy, moderate, and difficult items. Based on the analysis of item discrimination using the corrected-item total correlation (cr_{IT}) method, it can be concluded that most of the items in the *WM* test already had good differentiation. This meant that item was able to distinguish between students with high and low working memory capacity.

This study had some limitations. First, there were a number of suggestions for advanced research to improve the quality of the *WM* tests. Some items should be revised and retested with a larger sample, as well as considering the quality of the SMA to obtain heterogeneous samples. Then, re-testing of the reliability, validity, and analysis of items would be needed. Furthermore, norms could be compiled with the higher number of samples so that norms would produce a representative samples of high school students in Indonesia. In addition, other validity testing methods such as convergent and discriminant methods, factor analysis, and criterion validity testing by using academic achievement as criteria were required. In conclusion, the quality of the WM tests could be improved to incorporate better psychometric properties.

References

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

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*. Doi: 10.1177/0963721409359277

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.

Davis, D. (2011). Identifying working memory capacity: A study of two working memory assessment tools.

Dehn, M. J. (2008). *Working memory and academic learning: Assessment and intervention.* Wiley.

Direktorat Pembinaan SMA. (2010). Juknis Pembelajaran Tuntas, Remedial dan Pengayaan di SMA.

Filgueiras, A. (2016). Neural basis of phonological working memory: testing theoretical models using fmRI meta-analysis (Tesis, Pontifical Chatolic University of Rio de Janeiro, Rio de Janeiro, Brasil). Retrieved from https://www.maxwell.vrac.pucrio.br/Busca_etds.php?strSecao=resultado&nrSeq=26568@2. doi: 10.17771/PUCRio.acad.26568

Flanagan, D. P., & Dixon, S. G. (2013). The Cattell-Horn-Carroll Theory of cognitive abilities. In D. P. Flanagan (Ed.), *Encyclopedia of Special Education* (pp. 368-382). John Wiley & Sons.

Friedenberg, L. (1995). Psychological testing: Design, analysis, and use. Needham Heights, MA: Allyn & Bacon.

Kaplan, R. M., & Saccuzzo, D.P. (2005). *Psychological testing: Principles, aplications, and issues* (6th ed.). Belmont: Thomson Wadsworth.

Kemendikbud, R. I. (2013). Pedoman Peminatan Peserta Didik.

La lopa, J. M.,& Holich, G. (2014). The critical role of working memory in academic achievement. *Journal of Culinary Science & Technology, 12*, 258-278. doi: 10.1080/15428052.2014.913952

Little, S. G., & Akin-Little, A. (2014). *Academic assessment and intervention*. New York, NY: Routledge.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitiveabilities: Past, present, and future. In D. P. Flanagan &P. L. Harrison (Eds.), *Contemporary intellectual assessment:Theories, tests, and issues* (2nd ed., pp. 136–182). New York,NY: Guilford.

McGrew, K. S. (2014). Cattell-horn-carroll (CHC) theory of cognitive abilities definitions.

Nunnally, J. C., & Bernstein, I. H. (1994). The assessment of reliability.*Psychometric theory*, *3*(1), 248-292.

Pal, H. R., Pal, A.,& Tourani, P.(2004). Theories of intelligence. *Everyman's Science*, 3.

Rufaidah, A. (2015). Pengaruh inteligensi dan minat siswa terhadap putusan pemilihan jurusan. *Faktor Jurnal Ilmiah Kependidikan,2*(2).

Schneider, W. J., & McGrew, K. (2012). The Cattell-Horn-Carroll model of intelligence. In, D. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues (3rd ed.)*. New York: Guilford.

Tariq, S., & Noor, S. (2012). Impact of working memory on academic achievementof university science students in Punjab, Pakistan. *Journal of Education and Practice, 3(2).*

Weiss, L. G., Saklofske, D. H., Coalson, D., & Raiford, S. E. (Eds.). (2010). *WAIS-IV clinical use and interpretation: Scientist-practitioner perspectives*. Academic Press.

Widowati, V. N. (2015). Studi kasus tentang proses penjurusan beberapa SMA di Yogyakarta (Skripsi, Universitas Sanata Dharma, Yogyakarta, Indonesia).

Yui, L., Ng, R. J., & A, H. P. (2017). Concrete vs abstract words-what do you recall better?a study on dual coding theory. *Peer J*.

Table 1. Description of *Working Memory Capacity*.