CPM JP3I

by Yonathan Natanael

Submission date: 27-Aug-2022 01:52PM (UTC+0700)
Submission ID: 1887801063
File name: Template_JP3I_2020.docx (207.89K)
Word count: 6611
Character count: 35711

JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: http://dx.doi.org/10.15408/jp3i.vxix.xxxx
http://journal.uinjkt.ac.id/index.php/jp3i

Psychometric Evaluation of Raven Coloured Progressive Matrices Test in Indonesian Early Chilhood Sample using A Rasch Measurement Model

Yonathan Natanael[1], Irfan Fahmi[2], Dini Utami Mulyaningsih[3]

[1,2] Department of Psychology, Universitas Islam Negeri (UIN) Sunan Gunung Djati Bandung, Indonesia.
[3] Department of Tarbiya and Education, Universitas Islam Negeri (UIN) Sunan Gunung Djati Bandung, Indonesia.
E-mail: yonathan@uinsgd.ac.id

Abstract

Coloured Progressive Matrices (CPM) is a psychological test very well known among Indonesian psychologists as a test that functions to measure intelligence. Some researchers who use CPM in their research reveal that CPM has weakness in the principle of measurement equivalence. Therefore, the focus of this research is to evaluate the details of the psychometric properties of CPM by using the Rasch model. The research uses a secondary data analysis approach, where the primary data sets available in Psychological Service are collected into a single file by the researcher for further analysis. A total of 748 input data contains 371 boys and 377 girls with an age range from fifth to seventh years old who have taken an intelligence test to prepare children's school readiness. The analysis of the Rasch model shows that CPM is proven to have unidimensionality, has a fairly good reliability value, and eight-item are not suitable for testing intelligence. Only twenty-eight items of CPM are suitable for measuring the intelligence of children in Indonesia.

Keywords: intelligence, psychological testing, early chilhood, psychometric evaluation, Rasch measurement.

Introduction

Individual intelligence difference has become a fascinating study topic to be discussed from the past until now (Hülür et al., 2011). Raven (1983) defines intelligence as an ability that reflects individual differences in capturing information, experiences, and a condition that has been experienced. The wider community knows intelligence as one of the most important factors for children to determine their educational success (Ardini & Handini, 2018). Furthermore, according to Gorey (2001), intelligence also closely relates to academic achievement in the early childhood development stage. Therefore, it is inevitable that intelligence and education in children at the early childhood stage are two things that are inseparable from being prepared children to get an education in school until they are implementing school learning.

For example, children in the early childhood stage from ages five to seven in the United States are carefully prepared to enter elementary school since it is the primary age for a child to enter elementary school (Ziol-Guest & McKenna, 2014). Similar to parents in Indonesia, parents prepare their children to enter elementary school by taking a series of psychological assessments in the first place. The actual evidence can be seen in the data on participants in the Psychological assessment carried out by a Psychological Institute in Yogyakarta in 2021. There are 321 children aged six to seven taking a series of psychological assessment processes to determine whether the child is ready to learn in elementary school (Gusniarti et al., 2021). One of the instruments used in the series of assessments is Colored Progressive Matrices, which measure children's intelligence.

In principle, Coloured Progressive Matrices (CPM) are instruments to determine the level of intelligence of a child without any previous learning process (Sanz-Cervera et al., 2015). It is also used to measure abstract thinking skills in children (van Schoor et al., 2016). CPM can measure a child's abstract thinking ability since all of the items contain patterned, colored, and various shapes, hoping that children will be able to analyze various patterns and provide answers to the pictures one by one (Yoshizawa et al., 2014).

CPM is a psychological test popularly used among Indonesian psychologists in assessment activities. The popularity can not be separated from its time-saving procedure, its effortlessly administering, and the work pattern is quite simple only by responding to images that need to be adjusted to the shape of the pattern Besides being widely used for assessment activities, CPM has advantages in its use. It is not only used to measure the intelligence of children with a normal brain function (Khan, 2015), but also used for special needs children such as children with autism and attention-deficit/hyperactive disorder(Sanz-Cervera et al., 2015).

Bass (2000) emphasized that CPM is a psychological instrument used by many psychology researchers in various countries. In 2018, CPM was studied on children in Sardinian, Italy. This study used a large research sample; 1626 children aged five to thirteen from Elementary School and Junior High School became the research sample (Nicotra et al., 2018). The results of the research show descriptive results of the CPM calibration of items for each of the sets, as found that the most difficult items in CPM were item A11 (for Set A), item Ab12 (for Set Ab), and item B12 (for Set B). The study did not provide more detailed information regarding assumptions, fit statistics, instrument reliability, item bias detection on the instrument, and other indices that are very important to be presented when conducting analysis using the Rasch model.

Referring to other studies, CPM is strongly suspected of having weaknesses in terms of psychometric properties. The main weakness in CPM is the gap in the research results using CPM as a research instrument, especially in the analysis of instrument bias. The CPM in the two previous studies was described as a highly consistent instrument and free from bias (Agnoli et al., 2012; Antoniou et al., 2022). Both studies support CPM as an instrument that prioritizes the principle of measurement equivalence or fairness for anyone who uses it. In contrast to the two studies above, Sigmon (1983) and Lúcio et al. (2019) revealed that the CPM indicated a slight gender bias in its items. Other researchers

http://journal.uinjkt.ac.id/index.php/jp3i

also support that if CPM shows significant differences in boys, CPM is considered more beneficial for boys in its use (Lynn & Irwing, 2004). Therefore, in his research, Balboni et al. (2010) suggested a study to check measurement invariance in CPM.

This research aims to evaluate CPM in more detail from a psychometric point of view and prove the research suggestions made by Balboni and his colleagues (to examine measurement invariance in CPM). This study examines measurement invariance using the Differential Item Functioning (DIF) technique of the Rasch model analysis. It is used since it has the same logic in testing measurement invariance. In principle, DIF and measurement invariance are both used to identify bias or not in an instrument being analyzed.

Rasch Measurement Model

In 1960 a mathematician named George Rasch introduced mathematical modeling known as the Rasch model. The Rasch model was first introduced to analyze dichotomy data (Kreiner, 2013). The data is obtained from correct or incorrect answers on a test. The Rasch model, in its application, provides interpretation results from information-rich data, such as providing in-depth information about individual abilities and item difficulty levels (Khairani & Razak, 2015).

The individual ability parameter is obtained from the ratio calculation of the number of individuals answering correctly compared to the number of incorrect answers. The ratio value is then transformed to a range of interval sizes using logarithms or log odd; the final result is a logit value. The logit value is the final value that can be precise and accurate in measuring an individual's ability to be compared with other individuals (Khairani & Razak, 2015).

Similar to the difficulty level parameter, the value obtained is also in the form of logit. The value generated is also from the proportion of incorrect answers divided by correct answers on the items tested, which are then transformed using log odd. If the two parameters have the same unit size in the form of logit, then the two parameters can be compared equally (Khairani & Razak, 2015). Logit resulting from individual abilities and item difficulty levels can be explained more meaningfully (Wright & Stone, 1979).

The principal of Rasch model analysis has mandatory assumptions that need to be fulfilled, such as the assumption of unidimensionality and local independence (Mair, 2018), and added goodness of fit indices infit Mean-Square (MNSQ) and outfit MNSQ. The MNSQ outfit is a reasonable limit in determining the level of difficulty of an item. If the calculation shows that the MNSQ outfit value is less or more than the limit, it can be ascertained that the analyzed items indicate that they are unsuitable for use. At the same time, the MNSQ infit is very sensitive to the obtained responses (Khairani & Razak, 2015).

The privilege of analyzing the Rasch model is the detailed information describing the analysis results, such as fit statistics, reliability values, separation index, and comparison of individual abilities with item difficulty levels (Clements et al., 2008). The statistical fit index serves to see how the Rasch model meets the right expectations for the analytical model. It shows scores comparison from the overall individual ability (logit person mean) that can be compared to the item difficulty level (logit item mean) and displays item-person separation and item-person reliability to show the item's suitability and person in the tests carried out. According to previous research, the Rasch model is a very accurate method to see the quality of an instrument because there are item parameters with precise persons (Jong et al., 2015).

Methods

Research Design

The secondary data analysis method approach is used in this research (Johnston, 2014). It is a technique for analyzing primary data collected by other people or institutions with other purposes. Secondary data analysis includes an empirical research approach, noting that the research follows research principles when using direct data collection. Secondary data analysis can be carried out for systematic investigation processes in various fields. The mandatory steps in conducting secondary data analysis are (1) developing research questions, (2) identifying the data set as a whole to be carried out, and (3) evaluating the data set obtained (this stage will be described in the analysis procedure).

Sample

The primary data of this research was obtained from the results of intelligence testing conducted by Darunnisa Psychological Service on children in twenty schools in Bandung (Kindergarten and Elementary School) from 2017 to 2021. The primary data consists of hundreds of data files, which are merged into a set of files in the form of xls files (now referred to as secondary data). Based on secondary data, it is known that the total sample of this study was 748 children with a composition of 371 females (49.6%) and 377 males (50.4%). The age range of the research sample based on secondary data is five to seven years (M = 5.67 & SD = 0.54), which is based on the developmental task of the children. They are at the school readiness stage (Williams & Lerner, 2019).

Instrument

Raven's Colored Progressive Matrices is an instrument to measure children's intelligence, especially children aged five to eleven years old (Muniz et al., 2016). There are three sets of questions (set A, set Ab, set B), each consists of twelve items with different difficulty levels, so the total number of CPM items is 36. The difficulty level of item B is higher than that of item Ab, while item Ab's difficulty level is more difficult than item A. All items consist of blank picture sections that need to be filled in; children are expected to choose one of the six available answer options. There is only one correct answer for each question, indicating the minimum CPM value is 0 while the maximum CPM value is 36. Practically, CPM can be done individually or in groups since there is no time limit for the process. The test-retest reliability value of CPM in previous studies was .90, which means that CPM is an instrument that consistently measures intelligence in children (Lehmann et al., 2014).

Procedure

The research procedure uses steps based on the secondary data analysis approach. The first step regarding research questions was obtained after the researchers conducted a literature review and found problems with CPM, which psychologists in Indonesia widely use. A study of the literature review found that there were gaps in the results of previous research. These researchers had carried out the basis of obtaining the research question. This paper questioned whether CPM is suitable for measuring the intelligence of Indonesian children. Therefore, this study wanted to evaluate the psychometric properties of CPM with a sample of children in the early childhood stage in Indonesia.

The researcher took the second step to identify the data set as a whole. Before identifying the data set, the researcher asked permission to conduct the CPM research by using a request letter. After approval, Darunnisa Psychological Service was willing to provide data in hundreds of xls file data sets. Data set identification was carried out by researchers from the data given, and then it was combined into one xls file. Merging hundreds of data sets into one xls file is the next step to make statistical analysis easier since the statistical programs generally analyze only one data file. The last step is to analyze the xls file using a statistical program specifically analyzing the Rasch model. The analysis results will be presented in the research results section of this paper.

Statistical Analysis

The research data analysis was carried out using the Winstep 3.65 program. Some of the limitations of the ideal reference value used in this study include: (a) the benchmark value of an instrument proven

to be unidimensional is when the raw variance value explained by measured > 40% (Holster & Lake, 2016); (b) the criteria for an instrument that does not have local independence between items is by looking at the critical value (Q3) < .30 (Christensen et al., 2017); (c) the ideal limit of the person-item separation index > 3 (Duncan et al., 2003); (d) item fit testing with the Rasch model is the MNSQ outfit value in the range of .5 to 1.5 logit (Boone et al., 2014); (e) categorization of results from item calibration to determine item difficulty level using the range -.30 to .30 logit (Wicaksono et al., 2021); and (f) the item is indicated to be biased if the DIF construct value is > .40 (Rogers & Swaminathan, 1990).

Results and Discussion

Result

Unidimensionality and Local Independence

This study can also prove two mandatory assumptions in analyzing the Rasch model. The first assumption regarding unidimensionality in the Rasch research model is that the raw variance value explained by measures on CPM is 50.3%, exceeding the limit set by the previous research of > 40% (Holster & Lake, 2016). It is certain that the CPM in this study only measures one aspect, namely intelligence. The second assumption regarding local independence, Bond and Fox (2015), in their writings, suggests that local independence shows that there is no link between one item and another in terms of the response given to an instrument. The occurrence of linkages between items can be seen from the results of the largest standardized residual correlation between items. The largest standardized residual CPM value correlation was found between item A2 and item A3 at .23. The largest standardized residual correlation value obtained is below the standard critical value (Q3) < .30 (Christensen et al., 2017), which means that each CPM item is independent.

Fit Statistics and Reliability

Overall (see table 1), this study found the person mean = .55 logit, for the item mean = .00 logit. If the person means value is greater than the item means value on the cognitive scale, the sample does not feel difficulty when working. The average intelligence value of the sample who does CPM is in the high category. The results of this analysis are in line with research in measuring achievement tests (Othman et al., 2015). It can also be interpreted that CPM is too easy to do for children in the early childhood stage in Indonesia.

The person's standard deviation value is 1.09, below the item's standard deviation value of 2.60. It indicates that the research sample is a sample with almost similar intelligence levels (uniform to one another). Standard deviation items have a greater value, indicating that CPM items have varying problem difficulty levels. The person separation index = 1.9 < 3 strengthens the evidence from the small value of the person standard deviation, which means that the sample of this study is homogeneous, both in terms of age criteria, developmental stages, and in terms of intelligence in almost the same quality. The item separation index of 13.12 > 3 indicates that the difficulty level grouping on CPM items is appropriate according to the Winstep program. Direct evidence of the standard deviation and item separation item analysis will be described more clearly in the item fit section of this analysis.

The next result is to discuss item-person reliability. The value of person reliability is .78, while item reliability is .99. The reliability value here is not showing the constancy of an instrument. In this research, item reliability measures how good the CPM items tested are. Person reliability is used to measure the appropriateness of research participants in this study. Mohd et al. (2017) state that item-person reliability in this study is included in the good category, namely, if item reliability is > .70 and person reliability > .80. Item-person reliability criteria are met for this study, which shows no problems were found in items or persons (items and participants are correct in measuring this intelligence).

For the reliability value of CPM, as indicated by the Cronbach Alpha value (KR-20), it was found that the value of = .79 > .70 was in the acceptable range for an instrument (Sharma, 2016). CPM is an instrument that has been proven to measure intelligence consistently. Besides that, in testing the fit model, it is also discussed whether the Rasch modeling used is appropriate or not. This is showed by the Chi-square value = 20493.52 and p-value = 1, which proves that the test model is fit (p-value > .01) and acceptable (Renny et al., 2013).

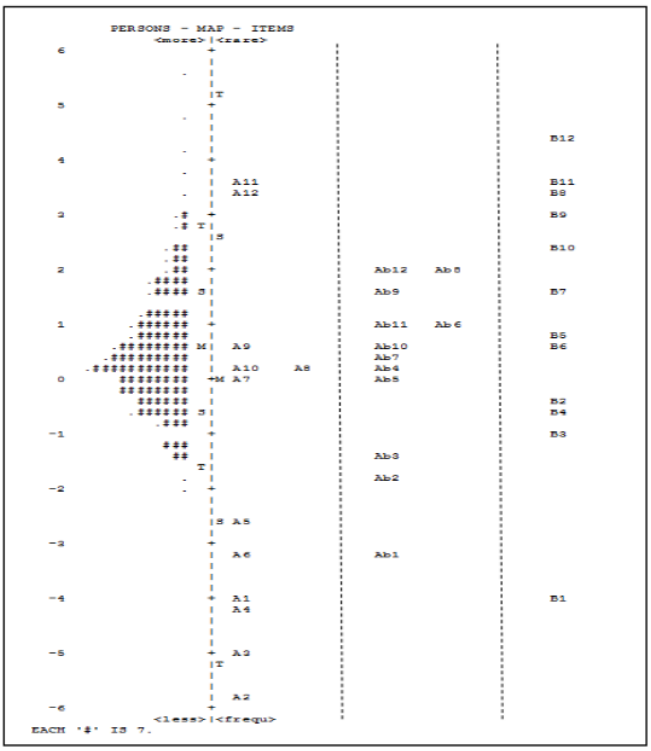
Table 1. Summary Fit Statistics

	Person	***Item***
N	748	36
Measure		
Mean	.55	.00
Standard Deviation	1.09	2.60
Standard Error	.50	.15
Outfit Mean Square		
Mean	1.01	1.00
Standard Deviation	.26	.10
Separation	1.90	13.12
Reliability	.78	.99
Alpha Cronbach (KR-20)	.79	
Chi-square	20493.52	
p-value	1	

Wright Map, Calibration, and Item Fit

The Wright map (see figure 1) shows the distribution of research sample ability level with the level of item difficulty. The researcher modified the Wright map for each test set to make it easier to understand. The Wright map, in general, shows that the ability of the research sample is higher than the level of the questions' difficulty. It can be seen from the pattern of distribution of the ability of the research sample that higher than the distribution question difficulty level. It can be observed that there are 10 of the 36 items are at the bottom of the diagram, proving that the CPM is a psychological instrument that is easy for the research sample to work on.

http://journal.uinjkt.ac.id/index.php/jp3i



Sources: Winstep output.

Figure 1. Wright Map Coloured Progressive Matrices

The evidence in further examination has shown that CPM is an easy instrument. It can be seen from the calibration results of the CPM items in table 2. The item calibration divides each set into several test categories, Wicaksono et al. (2021) categorize the level of difficulty of the test based on the range -.30 to +.30 logit value item (hard level if logit value item > .3, medium level in logit value range -.30 to .30, and easy level if logit value item < -.30). The CPM for Set A consisted of 12, categorized as hard, medium, and easy levels. The most difficult items in Set A are items A11, A12, A9.

Table 2. Item Fit & Calibration Coloured Progressive Matrices

Set Test	Category Level	Item Number	Logit Value Item	Standard Error	Outfit MNSQ	Point Mass. Corr
A	Hard	A11	3.56	.15	1.08	.23
		A12	3.45	.14	1.17	.20
		A9	.56	.08	1.11	.35
	Medium	A10	.23	.08	1.10	.36
		A8	.21	.08	1.40	.18
		A7	.04	.08	1.11	.33
	Easy	A5	-2.52	.15	1.04	.19
		A6	-3.16	.20	.61	.23
		A1	-3.97	.28	.53	.16
		A4	-4.15	.31	.83	.10
		A3	-4.95	.45	.29	.14
		A2	-5.88	.71	.17	.11

Ab	Hard	Ab12	2.10	.10	1.67	.13
		Ab8	1.98	.10	1.00	.43
		Ab9	1.57	.09	.95	.47
		Ab6	.99	.08	.79	.57
		Ab11	.95	.08	1.05	.39
		Ab10	.54	.08	1.11	.35
		Ab7	.42	.08	.96	.46
	Medium	Ab4	.19	.08	.90	.47
		Ab5	.08	.08	.92	.45
	Easy	Ab3	-1.30	.10	1.07	.29
		Ab2	-1.79	.12	.85	.30
		Ab1	-3.20	.20	.66	.21
B	Hard	B12	4.43	.21	.42	.36
		B11	3.54	.15	.69	.44
		B8	3.47	.14	.66	.42
		B9	3.08	.13	.77	.43
		B10	2.40	.11	1.18	.35
		B7	1.68	.09	.99	.43
		B5	.80	.08	.79	.55
		B6	.69	.08	.99	.43
	Easy	B2	-.43	.09	.96	.38
		B4	-.65	.09	.99	.41
		B3	-.90	.09	.79	.47
		B1	-4.06	.29	.52	.16

Note. Item's marked in gray lack of item fit limitation.

For Set Ab, the 12 items are divided into only three categories of difficulty levels, namely hard, medium and easy. Based on the results of item calibration, seventh difficult items were found for Set Ab. Set B consists of two levels of difficulty based on the calibration results. It was found that Set B was the set with the most difficult items. It can be seen from the eight items (B12, B11, B8, B9, B10, B7, B5, and B6) that are included in the hard level, while there are four easy item in Set B, namely item B2, B4, B3, and B1.

The results of item calibration in this study are closely related to the test fit item. Fit Item was analyzed using the Rasch model (see table 2) shows that three PCM items do not fit, which is indicated by the MNSQ outfit value less than the range .5 – 1.5 (Boone et al., 2014). The three items are items A3, A2, and B12. The researcher gives a gray mark on the value of the MNSQ outfit items that do not fit in table 2. In Rasch model analysis, items that do not meet the standard limits ideal should be discarded in the study. It can also be interpreted that the three items do not accurately measure intelligence or problematic items to be tested on the CPM.

Item Category

A misconception was also found between the person's ability and an item's difficulty level on one CPM item (item Ab12). In general, a child with high ability can answer correctly every item that is done. In item Ab12, it was found that 13 children (about 2% of the total sample) with high ability (average measure = 1.14 logit > .55 person mean logit) answered incorrectly on the questions with medium difficulty level. This item Ab12, when studied further, is suspected by the researcher to have a good answer distractor. For the other 35 items included in the general item category, if a child is smart, the child will be able to answer the questions asked.

DIF Analysis

This study also found something important to be described in the research results, namely the principle of measurement equivalence. Based on the results of the DIF analysis, CPM in this study violates the principle of measurement equivalence, which means that CPM is biased or favorable to one of the groups tested on several of its items (see table 3). DIF in CPM only occurs in Set A and Set B,

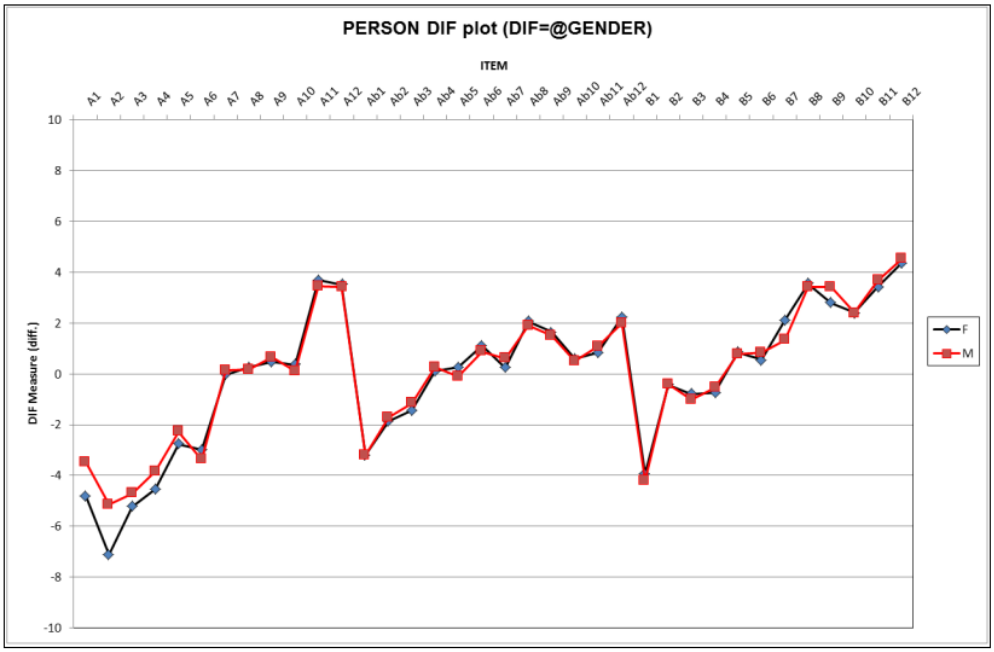
http://journal.uinjkt.ac.id/index.php/jp3i

while for Set Ab, it does not. It is confirmed that Set Ab is an ideal Set for measuring intelligence. In Set A, five items were detected to have DIF (A1, A2, A3, A4, and A5), and for Set B, it was found that two items had DIF (B7 and B9). The seven CPM items that experienced DIF in Set A and Set B were at moderate (.40 to .60) and high DIF (> .60) levels, following previous research, which stated that an item occurred DIF if the difference in DIF values between groups reached the referenced values (Rogers & Swaminathan, 1990). It can be seen from the graph (see figure 2) that the difference in DIF is visible from the distance range between the blue line (female) and the red line (male). Items A1, A2, A3, A4, A5, and B9 showed a higher DIF value in the male group (boys), which indicated that the six items benefited the male group (boys). Only 1 item, namely item B7, benefits the female group (girls).

Table 3. DIF Analysis

Item	DIF		DIF Contrast	t	Prob	Result DIF Criterion
	Male	Female				
A1	-3.47	-4.81	1.35	2.02	.043	High DIF
A2	-5.13	-7.11	1.99	1.03	.305	High DIF
A3	-4.71	-5.23	.51	.56	.577	Moderate DIF
A4	-3.84	-4.52	.68	1.07	.286	High DIF
A5	-2.29	-2.76	.47	1.54	.123	Moderate DIF
B7	1.33	2.11	.79	-4.24	.000	High DIF
B9	3.41	2.78	.63	2.43	.015	High DIF

Note. t-value positive is item tends to benefit the male group, and vice versa.



Sources: Picture originality from excel output (F = for Girls and M = for Boys).

Figure 2. Differential Item Functioning Coloured Progressive Matrices

Discussion

The assumption of unidimensionality and local independence was also proven in the analysis. Unidimensionality assumption shows strong evidence that CPM is a test that measures intelligence; it can be observed with the raw variance value limit explained by measures > 40%. The unidimensional assumption testing results support previous research, which obtained a test model that fits CPM with a unidimensional model using confirmatory factor analysis (Lúcio et al., 2019). The assumption of local independence shows that the items in the CPM do not have a close relationship with one another. All CPM items are ensured to be independent.

In the fit statistics test, only the person separation index was found lacking; less than 3 (Duncan et al., 2003), because the study sample was at the same stage of development, namely early childhood, as a not-too-distant age range, it can be said that the participants were homogeneous. Moreover, even when they were examined from the educational age of the research sample, they were included in the ideal age category for children to prepare for elementary school entry (Ziol-Guest & McKenna, 2014).

Regarding the instrument's consistency, the CPM reliability value obtained in this study was = .79. In previous studies using CPM, the test-retest reliability value = .90, which means that the reliability value obtained was smaller than in previous studies. It can be interpreted that CPM is less consistent in measuring intelligence in Indonesia. The results of this study are contradictive to the previous research, which stated that CPM is a consistent instrument (Agnoli et al., 2012). Another thing that allowed CPM to get a smaller reliability value than previous studies was that this study found four items that were not good at calibrating CPM items. From a psychometric point of view, the researcher's strong suspicion is that the items that are not good automatically have a tremendous impact on reliability calculations.

In addition, researchers also considered the research approach as an important factor in determining the consistency of the instrument. This study utilizes secondary data for analysis (one measurement), while Lehmann et al. (2014) conducted experimental research (twice measurements). The research approach is also strongly suspected as a factor that affects the reliability value of an instrument. The research approach that uses two measurements allows it to have more value than the instrument's consistency and the research sample's consistency. The study's limitation is that the data collection is not directly collected since it will require a high cost and a very long time to get as many as 748 data from children who will prepare themselves for school.

Regarding the item calibration analysis, something very useful was found in new information about CPM. In addition to obtaining four items that did not meet the MNSQ outfit value limit, the researcher found that each set in the CPM had different criteria for difficulty levels. For the research sample consisting of children at the early childhood stage in Indonesia, especially Set Ab in the CPM, there is no single item that is difficult to do. Being not difficult does not mean completely easy; for example, in the CPM category item analyzer for Set Ab on item Ab12, the researchers strongly suspect that distractors can make children who have high abilities get stuck in choosing answers. On the one hand, CPM is an easy test, but some distractions can make a child think harder.

Another very strong factor that plays a role in why high abilities children answer incorrectly on item Ab12 is the complexity of the picture pattern being asked. Based on the time calculations data carried out by testers at Darunnisa Psychological Service, it was found that on average, children working on CPM item Ab12 took longer than other items (for example item Ab12 M = 10.67 seconds, item Ab11 M = 10.3 seconds, item Ab12 M = 10.3 seconds, item Ab12 Ab5 M = 9.3 seconds, while for other items in Set Ab takes about 5.8 to 7.5 seconds). This finding proves that item complexity is highly correlated with the processing time of the questions.

The indication that the Ab12 item has the highest difficulty level in the Ab CPM Set is in line with the calibration results of the previous research item conducted in Italy. Nicotra et al. (2018) found that the most difficult items in CPM were item A11 (for Set A), item Ab12 (for Set Ab), and item B12 (for

http://journal.uinjkt.ac.id/index.php/jp3i

Set B). The findings of this study resulted in similar findings regarding the most difficult items in each test set. It can be interpreted that children in Indonesia and Italy have similarities in the level of difficulty working on CPM items with complex patterns. However, this research, which provides more detailed information in terms of fit items, is also shown, which researchers consider the strengths of this research.

Regarding the DIF analysis, it was known that seven items indicated gender bias (more favorable for the male group); this is following previous research allegations (Lúcio et al., 2019; Lynn & Irwing, 2004; Sigmon, 1983). It was found that the number of bias items is not a small percentage but quite large. When calculated, the number of CPM items identified as biased is 7 out of 36, or 19.4% of CPM items are gender-biased. In addition, it is also proven that six of the seven CPM items that are proven to be gender-biased are more favorable for the male group in measuring intelligence. In general, men benefit more from tests or tests in the form of visuals or geometry in line with research that has been carried out in Indonesia (Ridho, 2014). Ridho (2014) also, in his research, gives his view that men will benefit more if they are tested in the cognitive realm.

25 Another limitation in this study is that the researcher obtained data from Psychological institutions, which means that the researcher could not analyze whether cultural differences also caused the CPM items to be biased. In the gender bias test, it was found that there was an advantage in the male group. From the data obtained, there are no demographics regarding the culture of the research sample. The important thing that researchers feel is if the studies control the demographics of the participants, the richer the demographics may be able to provide more useful information, such as gender, culture, and even from different countries of origin.

Conclusion

So, based on data analysis and discussion, it can be concluded that eight items should not be used in intelligence testing using CPM, namely items A1, A2, A3, A4, A5, B7, B9, and B12. For items A1, A4, A5, and A7, it was not good enough due to DIF detection, item B12 did not meet the item fit limit, and items A2 and A3 were detected in both. Therefore, there are only 28 items that are eligible to be used to measure intelligence in samples at the early childhood stage.

The researcher's suggestion for further research is that with 28 items that are eligible for the CPM, it is necessary to make new norms that are adapted for children in Indonesia. In addition, CPM can be tested for cultural bias because the results of studies in different countries find different things. Studies using CPM likely indicate cultural bias, especially for researchers and psychologists who are interested in measuring children's intelligence. To prove the level of consistency of CPM in Indonesia, a study using an experimental approach with two measurements (pre-test and post-test) needs to be carried out as a follow-up study to evaluate further the reliability of CPM.

References

Agnoli, S., Mancini, G., Pozzoli, T., Baldaro, B., Russo, P. M., & Surcinelli, P. (2012). The interaction between emotional intelligence and cognitive ability in predicting scholastic performance in school-aged children. *Personality and Individual Differences*, *53*(5), 660–665. https://doi.org/10.1016/j.paid.2012.05.020

Antoniou, F., Alkhadim, G., Mouzaki, A., & Simos, P. (2022). A Psychometric Analysis of Raven's Colored Progressive Matrices: Evaluating Guessing and Carelessness Using the 4PL Item Response Theory Model. *Journal of Intelligence*, *10*(1), 6. https://doi.org/10.3390/jintelligence10010006

Ardini, P., & Handini, M. (2018). The Influence of Instructional Method, Visual Spatial Intelligence, and School Readiness on Early Reading Abilities. *Journal of Scientific Research and Reports*, *17*(4), 1–22. https://doi.org/10.9734/jsrr/2017/38737

Balboni, G., Naglieri, J. A., & Cubelli, R. (2010). Concurrent and predictive validity of the Raven Progressive Matrices and the Naglieri Nonverbal Ability Test. *Journal of Psychoeducational Assessment*, *28*(3), 222–235. https://doi.org/10.1177/0734282909343763

Bass, N. (2000). *The Raven' s Coloured Progressive Matrices Test: A pilot study for the establishment of normative data for Xhosa- speaking Primary School pupils in the Grahamstown region* [Thesis, Rhodes University]. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.535.5516&rep=rep1&type=pdf

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model. In Applying the Rasch model: Fundamental Measurement in the Human Sciences* (3rd Ed). Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer Science & Bussiness Media Dordrecht.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-Based Early Maths Assessment. *Educational Psychology*, *28*(4), 457–482. https://doi.org/10.1080/01443410701777272

Duncan, P. W., Bode, R. K., Min Lai, S., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation*, *84*(7), 950–963. https://doi.org/10.1016/S0003-9993(03)00035-2

Gorey, K. M. (2001). Early childhood education: A meta-analytic affirmation of the short- and long-term benefits of educational opportunity. *School Psychology Quarterly*, *16*(1), 9–30. https://doi.org/10.1521/scpq.16.1.9.19163

Gusniarti, U., Rachmawati, M. A., Wibisono, S., Annatagia, L., Agustina, I., & Rumiani, R. (2021). Norming of Coloured Progressive Matrices test in elementary school children based on classical measurement theory and Rasch modeling. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia (JP3I)*, *10*(2), 172–183. https://doi.org/10.15408/jp3i.v10i2.18155

Holster, T. A., & Lake, J. (2016). Guessing and the Rasch model. *Language Assessment Quarterly*, *13*(2), 124–141. https://doi.org/10.1080/15434303.2016.1160096

Hülür, G., Wilhelm, O., & Robitzsch, A. (2011). Intelligence differentiation in early childhood. *Journal of Individual Differences*, *32*(3), 170–179. https://doi.org/10.1027/1614-0001/a000049

Johnston, M. P. (2014). Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries*, *3*, 619–626. http://www.qqml-journal.net/index.php/qqml/article/download/169/170

Jong, C., Hodges, T. E., Royal, K. D., & Welder, R. M. (2015). Instruments to measure elementary preservice teachers concepetions: An application of the Rasch rating scale model. *Educational Research Quarterly*, *39*(1), 21–48. https://files.eric.ed.gov/fulltext/EJ1166724.pdf

Khairani, A. Z. B., & Razak, N. B. A. (2015). Modeling a multiple choice mathematics test with the Rasch model. *Indian Journal of Science and Technology*, *8*(12), 1–6. https://doi.org/10.17485/ijst/2015/v8i12/70650

Khan, S. A. (2015). Relationship between dental fluorosis and intelligence quotient of school going children in and around lucknow district: A cross-sectional study. *Journal of Clinical and Diagnostic Research*, *9*(11), ZC10–ZC15. https://doi.org/10.7860/JCDR/2015/15518.6726

Kreiner, S. (2013). The Rasch model for dichotomous items. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch Models in Health* (pp. 5–26). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118574454.ch1

Lehmann, J., Quaiser-Pohl, C., & Jansen, P. (2014). Correlation of motor skill, mental rotation, and working memory in 3- to 6-year-old children. *European Journal of Developmental Psychology*, *11*(5), 560–573. https://doi.org/10.1080/17405629.2014.888995

Lúcio, P. S., Cogo-Moreira, H., Puglisi, M., Polanczyk, G. V., & Little, T. D. (2019). Psychometric investigation of the Raven's Colored Progressive Matrices test in a sample of preschool children. *Assessment*, *26*(7), 1399–1408. https://doi.org/10.1177/1073191117740205

Lynn, R., & Irwing, P. (2004). *Sex differences on the progressive matrices: A meta-analysis*. *32*(5), 481–498. https://doi.org/10.1016/j.intell.2004.06.008

Mair, P. (2018). *Modern psychometrics with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-93177-7

Mohd, R., Bakar, N. A., Hassan, S., & Hussain, A. H. (2017). Sustainable post-disaster recovery plan for flood victims in Gus Musang and Kuala Krai, Kelantan. *Pertanika Journal of Social Sciences & Humanities*, *25*(S), 1–12. http://psasir.upm.edu.my/id/eprint/58310/1/JSSH Vol. 25 %28S%29 Jan. 2017 %28View Full Journal%29.pdf#page=17

Muniz, M., Gomes, C. M. A., & Pasian, S. R. (2016). Factor structure of Raven's Coloured Progressive Matrices. *Psico-USF*, *21*(2), 259–272. https://doi.org/10.1590/1413-82712016210204

Nicotra, E. F., Grassi, P., Masala, C., & Petretto, D. R. (2018). A Rasch analysis of Raven's Colored Progressive Matrices to assess eductive intelligence: A study in a Sardinian sample. *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 1–6. https://doi.org/10.1109/MeMeA.2018.8438712

Othman, H., Ismail, N. A., Asshaari, I., Hamzah, F. M., & Nopiah, Z. M. (2015). Application Rasch measurement model for reliability measurement instrument in vector calculus course. *Journal of Engineering Science and Technology*, *10*(2), 77–83. http://jestec.taylors.edu.my/Special Issue UKM TLC 2013_2/UKMTLC 2013_6_2015_2_077_083.pdf

Raven, J. (1983). Raven Coloured Progressive Matrices Intelligence test in Thailand and in Denmark: A response. *School Psychology International*, *4*, 173–176. https://doi.org/10.1177/0143034383043007

Renny, Guritno, S., & Siringoringo, H. (2013). Perceived usefulness, ease of use, and attitude toward online shopping usefulness towards online airlines ticket purchase. *Procedia - Social and Behavioral Sciences*, *81*, 212–216. https://doi.org/10.1016/j.sbspro.2013.06.415

Ridho, A. (2014). *Differential item functioning pada tes multidimensi* [Dissertation, Universitas Gajah Mada]. https://www.google.co.id/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjQlrb7k_bvAhXafH0KHW_6C3wQFnoECAMQAA&url=http%3A%2F%2Frepository.uin-malang.ac.id%2F750%2F1%2FAli-Ridho.pdf&usg=AOvVaw0m21OYVtLMgiwu7rL4tuiJ

Rogers, H. J., & Swaminathan, H. (1990). A Comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105–116. https://doi.org/10.1177/014662169301700201

Sanz-Cervera, P., Pastor-Cerezuela, G., Fernández-Andrés, M. I., & Tárraga-Mínguez, R. (2015). Sensory processing in children with autism spectrum disorder: Relationship with non-verbal IQ, autism severity and attention deficit/hyperactivity disorder symptomatology. *Research in Developmental Disabilities*, *45–46*, 188–201. https://doi.org/10.1016/j.ridd.2015.07.031

Sharma, B. (2016). A focus on reliability in developmental research through Cronbach Alpha among medical, dental and paramedical professional. *Asian Pacific Journal of Health Sciences*, *3*(4), 271–278. https://doi.org/10.2466/pms.1983.56.2.484

Sigmon, S. B. (1983). Performance of American school children on Raven's Colored Progressive Matrices scale. *Perceptual and Motor Skills*, *56*(2), 484–486. https://doi.org/10.2466/pms.1983.56.2.484

van Schoor, N. M., Comijs, H. C., Llewellyn, D. J., & Lips, P. (2016). Cross-sectional and longitudinal associations between serum 25-hydroxyvitamin D and cognitive functioning. *International Psychogeriatrics*, *28*(5), 759–768. https://doi.org/10.1017/S1041610215002252

Wicaksono, D. A., Roebianto, A., & Sumintono, B. (2021). Internal validation of the Warwick-Edinburgh Mental Wellbeing Scale: Rasch analysis in the Indonesian context. *Journal of Educational, Health and Community Psychology*, *10*(2), 229–248. https://doi.org/10.12928/jehcp.v10i2.20260

Williams, P. G., & Lerner, M. A. (2019). School readiness. *Pediatrics*, *144*(2), 54–66. https://doi.org/10.1542/peds.2019-1766

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA PRESS.

Yoshizawa, K., Yasuda, N., Fukuda, M., Yukimoto, Y., Ogino, M., Hata, W., Ishizaka, I., & Higashikawa, M. (2014). Syntactic comprehension in patients with amyotrophic lateral sclerosis. *Behavioural Neurology*, *2014*, 1–8. https://doi.org/10.1155/2014/230578

Ziol-Guest, K. M., & McKenna, C. C. (2014). Early childhood housing instability and school readiness. *Child Development*, *85*(1), 103–113. https://doi.org/10.1111/cdev.12105

CPM JP3I

ORIGINALITY REPORT

12%	10%	8%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Diponegoro Student Paper	4%
2	journal.uinjkt.ac.id Internet Source	2%
3	pdfs.semanticscholar.org Internet Source	<1%
4	W S M Sanjaya, D Anggraeni, R Denya, N Ismail. "Numerical Simulation Bidirectional Chaotic Synchronization of Spiegel-Moore Circuit and Its Application for Secure Communication", IOP Conference Series: Materials Science and Engineering, 2017 Publication	<1%
5	"Rasch Measurement", Springer Science and Business Media LLC, 2020 Publication	<1%
6	www.tandfonline.com Internet Source	<1%
7	"Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference	<1%

Proceedings", Springer Science and Business Media LLC, 2018
Publication

8	"Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings", Springer Science and Business Media LLC, 2016 Publication	<1 %
9	Submitted to University of Malaya Student Paper	<1 %
10	PARK, MIHWA, and XIUFENG LIU. "Assessing Understanding of the Energy Concept in Different Science Disciplines : ASSESSING UNDERSTANDING OF THE ENERGY", Science Education, 2016. Publication	<1 %
11	journals.sagepub.com Internet Source	<1 %
12	Muntazhimah Muntazhimah, Reni Wahyuni. "The Development and Validation of Mathematical Reflective Thinking Test Instruments for Prospective Mathematics Teachers Using the Rasch Model", Jurnal Elemen, 2022 Publication	<1 %
13	documents.uow.edu.au Internet Source	<1 %

14	Bahrul Hayat. "Adjustment for guessing in a basic statistics test for Indonesian undergraduate psychology students using the rasch model", Cogent Education, 2022 Publication	<1 %
15	Kusen Kusen. "Rasch Test for Digital Learning Material Development Based on MOOCS", Tadbir : Jurnal Studi Manajemen Pendidikan, 2022 Publication	<1 %
16	flex.flinders.edu.au Internet Source	<1 %
17	hosted.jalt.org Internet Source	<1 %
18	journal.unnes.ac.id Internet Source	<1 %
19	obsesi.or.id Internet Source	<1 %
20	Almubarak, R Iriani. "Students' understanding of chemical bonding material through the rasch modelling data symptoms analysis", Journal of Physics: Conference Series, 2021 Publication	<1 %
21	Lois Finch, Johanne Higgins, Sharon Wood-Dauphinee, Nancy Mayo. "Development of a measure of functioning for stroke recovery:	<1 %

	The functional recovery measure", Disability and Rehabilitation, 2009 Publication	
22	dl.uswr.ac.ir Internet Source	<1 %
23	dr.library.brocku.ca Internet Source	<1 %
24	handwiki.org Internet Source	<1 %
25	helda.helsinki.fi Internet Source	<1 %
26	link.springer.com Internet Source	<1 %
27	psasir.upm.edu.my Internet Source	<1 %
28	pubs.asha.org Internet Source	<1 %
29	www.agba.us Internet Source	<1 %
30	www.psychologyandeducation.net Internet Source	<1 %
31	www.researchgate.net Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		