JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia), 12(2), 2023, 93-107
DOI: http://dx.doi.org/10.15408/jp3i.v12i2.27838
http://journal.uinjkt.ac.id/index.php/jp3i

# Evaluating Psychometric Properties of Raven's Coloured Progressive Matrices Test in Indonesian Sample using the Rasch Model

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

Department of Psychology, UIN Sunan Gunung Djati Bandung, Indonesia[1]
Department of Tarbiya and Education, UIN Sunan Gunung Djati Bandung, Indonesia[2]

yonathan@uinsgd.ac.id

## Abstract

Coloured Progressive Matrices (CPM) is a psychological test well known among Indonesian psychologists to measure intelligence. Some researchers who use CPM in their research reveal that CPM has weaknesses 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. This research used a secondary data analysis approach, where the primary data sets from a psychological service were collected into a single file for further analysis. Data of 371 boys and 377 girls with an age range of five to seven years old who took an intelligence test to assess their school readiness were collected. The Rasch model analysis showed that CPM showed unidimensionality and local independence, had a fairly good reliability value, and eight items were unsuitable for testing intelligence. Only twenty-eight items of CPM were suitable for measuring children's intelligence in Indonesia.

**Keywords**: coloured progressive matrices, psychometric properties, Rasch model

## *Abstrak*

Coloured Progressive Matrices (CPM) *merupakan salah satu tes psikologi yang sangat terkenal di kalangan psikolog Indonesia yang berfungsi untuk mengukur kecerdasan. Beberapa peneliti sebelumnya yang menggunakan CPM untuk keperluan penelitian mengungkapkan bahwa CPM memiliki kelemahan dalam prinsip kesetaraan dalam pengukuran. Oleh karena itu, fokus penelitian ini adalah mengevaluasi secara detail properti psikometri CPM dengan* Rasch model. *Penelitian ini menggunakan pendekatan analisis data sekunder, dimana kumpulan data primer yang tersedia berada di suatu biro psikologi yang dikumpulkan data satu file oleh peneliti untuk dianalisis lebih lanjut. Sebanyak 748 data input terdiri dari 371 anak laki-laki dan 377 anak perempuan dengan rentang usia lima sampai tujuh tahun, yang telah mengikuti seluruh rangkaian tes kecerdasan untuk persiapan sekolah. Analisis* Rasch model *menunjukkan bahwa CPM terbukti memenuhi asumsi unidimensional dan independensi lokal, memiliki nilai reliablitas yang cukup baik, dan diketahui delapan item kurang sesuai untuk mengukur inteligensi. Hanya dua puluh delapan item CPM yang cocok untuk mengukur kecerdasan anak di Indonesia.*

**Kata kunci**: coloured progressive matrices, *properti psikometri*, Rasch model

## Introduction

Individual differences in intelligence have become a fascinating study topic to be discussed (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 community in general are aware that intelligence is one of the most important factors to determine children's educational success (Ardini & Handini, 2018). Furthermore, according to Gorey (2001), intelligence is also closely related to academic achievement in the early childhood development stage. Therefore, it is inevitable that intelligence and education in children in the early childhood stage are inseparable from preparing children for school until they receive education with more formal curriculum.

For example, children in the early childhood stage from age five to seven in the United States are carefully prepared for elementary school since this age range 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. The actual evidence can be seen in the data on participants in a psychological assessment conducted by a psychological institute in Yogyakarta in 2021. A total of 321 children aged six to seven went through a series of psychological assessment processes to determine whether they were prepared for elementary school (Gusniarti et al., 2021). One of the instruments used in the series of assessments is Coloured Progressive Matrices, which measure children's intelligence.

In principle, Coloured Progressive Matrices (CPM) are instruments to determine a child's level of intelligence 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 the most widely used psychological test among Indonesian psychologists in assessment activities. The reason is because of its time-saving procedure, the test is easily administered, 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 is 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; 1,626 elementary school and junior high school children aged five to thirteen became the research sample (Nicotra et al., 2018). The results of the research showed descriptive results of the CPM calibration of items for each of the sections. This study found that the most difficult items in CPM were item A11 (for Section A), item Ab12 (for Section Ab), and item B12 (for Section 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 important to be presented when conducting analysis using the Rasch model.

Referring to other studies, CPM is highly likely to have weaknesses in terms of psychometric properties. The main weakness in CPM is the gap in the results of research using CPM as a research instrument, especially in the analysis of instrument bias. Two previous studies described CPM as a highly consistent instrument, and it is 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 CPM indicated a slight gender bias in its items. Other researchers 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 further study to check measurement invariance in CPM.

This research aims to evaluate CPM in a more detailed manner from a psychometric point of view and to examine measurement invariance in CPM as suggested by Balboni et al. 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 whether the instrument being analyzed has a bias.

**Rasch Measurement Theory**

In 1960 a mathematician named George Rasch introduced a mathematical modeling known as the Rasch model. The Rasch model was first introduced to analyze dichotomous data (Kreiner, 2013), in which the data was obtained from correct or incorrect answers on a test. In addition, the Rasch model can also be used to analyze polytomous data widely used to measure attitude known as the Rating Scale Model or Partial Credit Model (Andrich & Marais, 2019). In its application, the Rasch model provides interpretation results from information-rich data, such as providing in-depth information about individual abilities and item difficulty levels (Khairani & Razak, 2015).

To describe the relationship estimation of individual abilities and item difficulty, the Rasch model could be given by the following formula:

$$P(X_{ni} = 1|\beta_n, \delta_i, \alpha_i) = \frac{e^{\alpha(\beta_n - \delta_i)}}{1 + e^{\alpha(\beta_n - \delta_i)}}$$

where:

$X_{ni} = 1$ refers to the response obtained by subject *n* to item *i* (correct response of the item)

$\beta_n$ refers to the ability of subject n;

$\delta_i$ refers to the difficulty of the i;

$\alpha_i$ refers to an item's discrimination index;

*e* is the base of natural point logarithm (e = 2.718….)

The individual ability parameter is obtained from the ratio calculation of the number of individuals answering correctly 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, and it 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, they 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, such as the assumption of unidimensionality and local independence (Mair, 2018), and added goodness of fit indices of 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 is likely that the analyzed items are unsuitable for use. At the same time, the MNSQ infit is very sensitive to the obtained responses (Khairani & Razak, 2015).

The advantage of analysis using 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 highly 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 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).

### Participants

The primary data of this research was obtained from the results of intelligence testing conducted by Darunnisa Psychological Service on children in twenty kindergartens and elementary schools in Bandung from 2017 to 2021. The primary data consisted of hundreds of data files, which were merged into a set of files in the form of Excel files (now referred to as secondary data). Based on the secondary data, the total number of participants of this study was 748 children comprising 371 girls (49.6%) and 377 boys (50.4%). The age range of the research participants based on the secondary data is five to seven years (M = 5.67 & SD = 0.54), which is based on the developmental task of the children. These children 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 sections of questions (section A, section Ab, section B), and each consists of twelve items with different difficulty levels, and thus 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 score is 0 while the maximum CPM score is 36. Practically, CPM can be done individually or in groups since there is no time limit for the assessment. 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 adopted steps based on the secondary data analysis approach. The first step was regarding research questions. The research questions were developed after the researchers conducted a literature review and found problems with CPM. 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 the quality of psychometric properties of CPM. Therefore, this study aimed to evaluate the psychometric properties of CPM with participants of children in the early childhood stage in Indonesia.

The second step was to identify the data set as a whole. Before identifying the data set, the researchers sent a letter requesting permission to Darunnisa Psychological Service to conduct the CPM research there. After giving permission, Darunnisa Psychological Service was willing to provide data in hundreds of Excel file data sets. Data set identification was carried out by researchers from the data given, and then it was combined into one Excel file. Merging hundreds of data sets into one Excel file was the next step to make statistical analysis easier since the statistical programs generally analyze only one data file. The last step was to analyze the Excel file using a statistical program, specifically the Rasch model analysis. The results of the analysis will be presented in the 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 is proven to be unidimensional when the raw variance value is explained by a measures value of > 40% (Holster & Lake, 2016); (b) the criterion for an instrument that does not have local independence between items is by looking at the critical value (Q3) of < .30 (Christensen et al., 2017); (c) the ideal limit of the person-item separation index is > 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 uses 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

### Results

#### *Unidimensionality and Local Independence*

This study may also prove two mandatory assumptions in analyzing the Rasch model. The first assumption is regarding unidimensionality in the Rasch research model. The raw variance value explained by measures on CPM was 50.3%, exceeding the limit set by the previous research of > 40% (Holster & Lake, 2016). The CPM in this study only measured one aspect, namely intelligence. The second assumption is on local independence. In their book, Bond and Fox (2015) suggest that local independence shows that there is no link between one item and another in terms of the response given to an instrument. 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 was below the standard critical value (Q3) of < .30 (Christensen et al., 2017), which may indicate that each CPM item is independent.

#### *Fit Statistics and Reliability*

Overall (see Table 1), this study found the person mean = .55 logit, while for the item mean = .00 logit. If the person means value was greater than the item means value on the cognitive scale, the participant had no difficulty in doing the test. The average intelligence value of the participant who did CPM was in the high category. The results of this analysis are in line with those of research on measuring achievement tests (Othman et al., 2015). This suggests that CPM could be too easy for children in the early childhood stage in Indonesia to complete.

The person's standard deviation value was 1.09, below the item's standard deviation value of 2.60. This may indicate that the research participants had almost similar intelligence levels (uniform to one another). Standard deviation items had a greater value, indicating that CPM items had varying problem difficulty levels. The person separation index = 1.9 < 3 supported the evidence from the small value of the person standard deviation, which may indicate that the participants of this study were homogeneous, in terms of age criteria, developmental stages, and level of intelligence. The item separation index of 13.12 > 3 could indicate that the difficulty level grouping on CPM items was appropriate according to the Winstep program. Direct evidence of the standard deviation and item separation analysis will be further explored in the item fit section of this analysis.

The next result discusses item-person reliability. The value of person reliability was .78, while the value of item reliability was .99. The item reliability value here did not show the constancy of an instrument. In this research, item reliability measured how good the CPM items tested were. Person reliability was used to measure the appropriateness of research participants in this study. Item-person reliability in this study could be categorized as good, that is, item reliability is > .70 and person reliability is > .80 (Mohd et al., 2017). Item-person reliability criteria were met for this study, which shows that no problems were found in items or persons (items and participants were correct in measuring this intelligence).

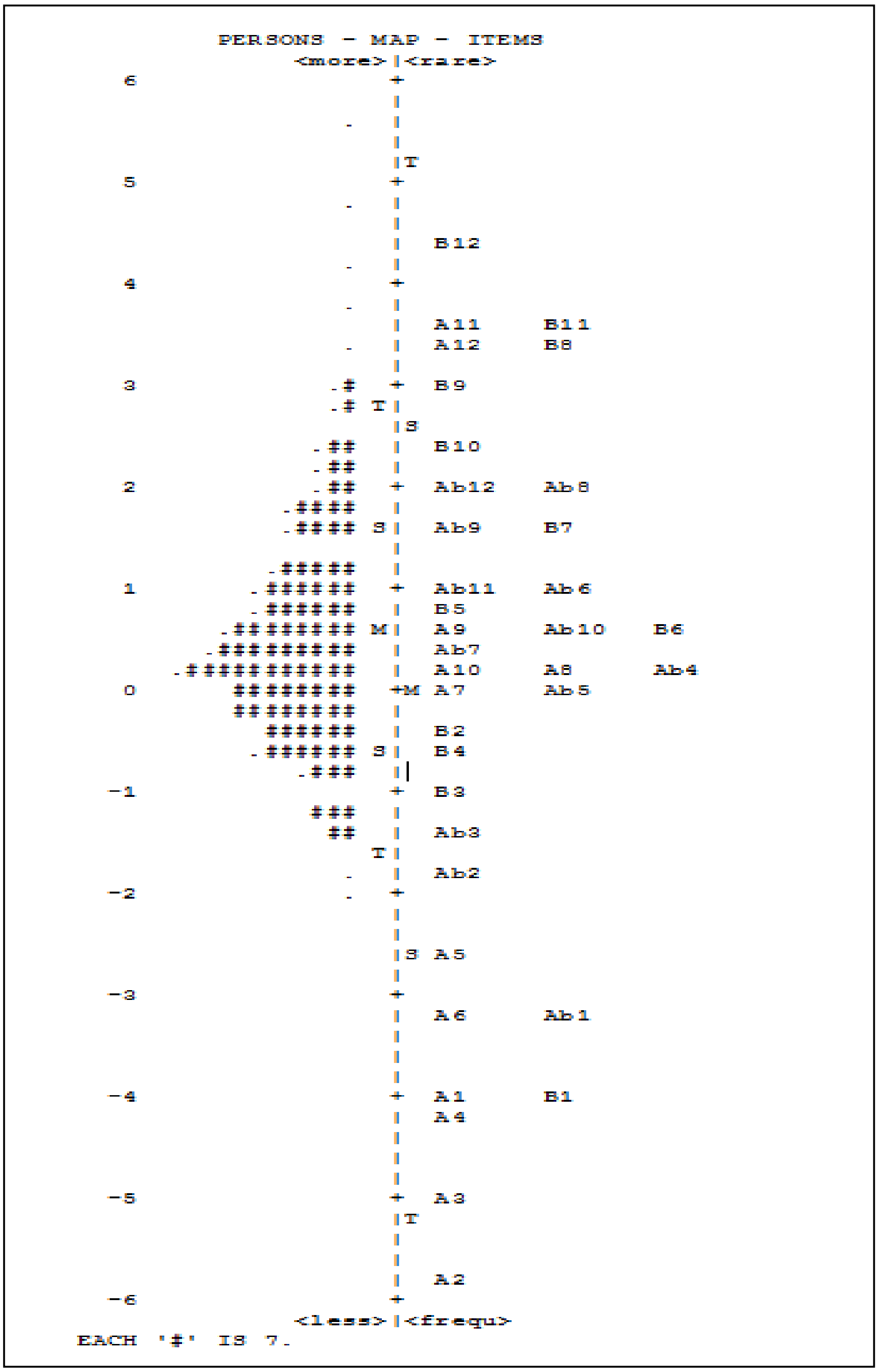
For the reliability value of CPM, as indicated by the Cronbach Alpha value (KR-20), it was found that the value of = .79 > .70 was in the acceptable range for an instrument (Sharma, 2016). CPM has been proven to measure intelligence consistently. In addition, in testing the fit model, a question arises is whether the Rasch modeling used is appropriate or not. This is shown 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 (Figure 1) shows the distribution of research participants' ability level with the level of item difficulty. Researchers modified the Wright map for each test section to make it easier to understand. In general, the Wright map shows that the ability of the research participants is higher than the questions' difficulty level. It can be seen from the pattern of distribution of the ability of the research participants that was higher than the distribution of questions' difficulty level. It is evident that 10 of the 36 items are at the bottom of the diagram, which may suggest that the CPM is a psychological instrument that could be slightly easy for the research participants to complete.



```
             PERSONS - MAP - ITEMS
                  <more>|<rare>
     6                  +
                        |
                    .   |
                        |
                        |T
     5                  +
                    .   |
                        |
                        |  B12
                    .   |
     4                  +
                    .   |
                        |  A11     B11
                    .   |  A12     B8
                        |
     3             .#   +  B9
                   .# T |
                        |S
                  .##   |  B10
                  .##   |
     2            .##   +  Ab12    Ab8
                .####   |
                .#### S |  Ab9     B7
                        |
               .#####   |
     1        .######   +  Ab11    Ab6
              .######   |  B5
            .######## M |  A9      Ab10    B6
           .#########   |  Ab7
         .###########   |  A10     A8      Ab4
     0      ########    +M A7      Ab5
            ########    |
              ######    |  B2
             .###### S |  B4
                 .###   ||
    -1                  +  B3
                  ###   |
                   ##   |  Ab3
                      T |
                    .   |  Ab2
    -2              .   +
                        |
                        |
                        |S A5
                        |
    -3                  +
                        |  A6      Ab1
                        |
                        |
                        |
    -4                  +  A1      B1
                        |  A4
                        |
                        |
                        |
    -5                  +  A3
                        |T
                        |
                        |
                        |  A2
    -6                  +
                  <less>|<frequ>
  EACH '#' IS 7.
```

Sources: Winstep output

**Figure 1.** Wright Map Coloured Progressive Matrices

The evidence from 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 section 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 > .30, medium level if logit value ranges from -.30 to .30, and easy level if logit value item < -.30). The CPM for Section A consisted of 12, categorized as hard, medium, and easy levels. The most difficult items in Section A are items A11, A12, A9.

**Table 2.** Item Fit & Calibration Coloured Progressive Matrices

| Section Test | Category Level | Item Number | Logit Value Item | Standard Error | Infit MNSQ | Outfit MNSQ | Point Mass. Corr |
|---|---|---|---|---|---|---|---|
| **Section A** | Hard | A11 | 3.56 | .15 | 1.13 | 1.08 | .23 |
| | | A12 | 3.45 | .14 | 1.18 | 1.17 | .20 |
| | | A9 | .56 | .08 | 1.10 | 1.11 | .35 |
| | Medium | A10 | .23 | .08 | 1.05 | 1.10 | .36 |
| | | A8 | .21 | .08 | 1.26 | 1.40 | .18 |
| | | A7 | .04 | .08 | 1.09 | 1.11 | .33 |
| | Easy | A5 | -2.52 | .15 | 1.01 | 1.04 | .19 |
| | | A6 | -3.16 | .20 | .95 | .61 | .23 |
| | | A1 | -3.97 | .28 | .99 | .53 | .16 |
| | | A4 | -4.15 | .31 | 1.01 | .83 | .10 |
| | | A3 | -4.95 | .45 | .98 | .29 | .14 |
| | | A2 | -5.88 | .71 | .98 | .17 | .11 |
| **Section Ab** | Hard | Ab12 | 2.10 | .10 | 1.31 | 1.67 | .13 |
| | | Ab8 | 1.98 | .10 | .98 | 1.00 | .43 |
| | | Ab9 | 1.57 | .09 | .94 | .95 | .47 |
| | | Ab6 | .99 | .08 | .83 | .79 | .57 |
| | | Ab11 | .95 | .08 | 1.04 | 1.05 | .39 |
| | | Ab10 | .54 | .08 | 1.08 | 1.11 | .35 |
| | | Ab7 | .42 | .08 | .96 | .96 | .46 |
| | Medium | Ab4 | .19 | .08 | .90 | .90 | .47 |
| | | Ab5 | .08 | .08 | .92 | .92 | .45 |
| | Easy | Ab3 | -1.30 | .10 | 1.02 | 1.07 | .29 |
| | | Ab2 | -1.79 | .12 | .98 | .85 | .30 |
| | | Ab1 | -3.20 | .20 | .97 | .66 | .21 |
| **Section B** | Hard | B12 | 4.43 | .21 | .89 | .42 | .36 |
| | | B11 | 3.54 | .15 | .85 | .69 | .44 |
| | | B8 | 3.47 | .14 | .87 | .66 | .42 |
| | | B9 | 3.08 | .13 | .90 | .77 | .43 |
| | | B10 | 2.40 | .11 | 1.02 | 1.18 | .35 |
| | | B7 | 1.68 | .09 | .99 | .99 | .43 |
| | | B5 | .80 | .08 | .86 | .79 | .55 |
| | | B6 | .69 | .08 | .99 | .99 | .43 |
| | Easy | B2 | -.43 | .09 | 1.01 | .96 | .38 |
| | | B4 | -.65 | .09 | .94 | .99 | .41 |
| | | B3 | -.90 | .09 | .86 | .79 | .47 |
| | | B1 | -4.06 | .29 | .98 | .52 | .16 |

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

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

The results of item calibration in this study were closely related to those of the test fit item. Fit item was analyzed using the Rasch model (see Table 2), and the results showed that three of the CPM items did not fit, as indicated by the MNSQ outfit value of less than the range of .5 – 1.5 (Boone et al., 2014). The three items were items A3, A2, and B12. A gray mark was given on the value of the MNSQ outfit items that did not fit in Table 2. In the Rasch model analysis, items that do not meet the standard limits ideal should be discarded. This may suggest that the three items do not accurately measure intelligence or problematic items to be tested on the CPM.

***Distractor Analysis***

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 every item correctly. In item Ab12, it was found that 13 children (about 2% of the total number of participants) with high ability (average measure = 1.14 logit > .55 person mean logit) incorrectly answered questions with medium difficulty level. Item Ab12, when studied further, may have a good answer distractor. For the other 35 items included in the general item category, the questions could be addressed if a child is smart.

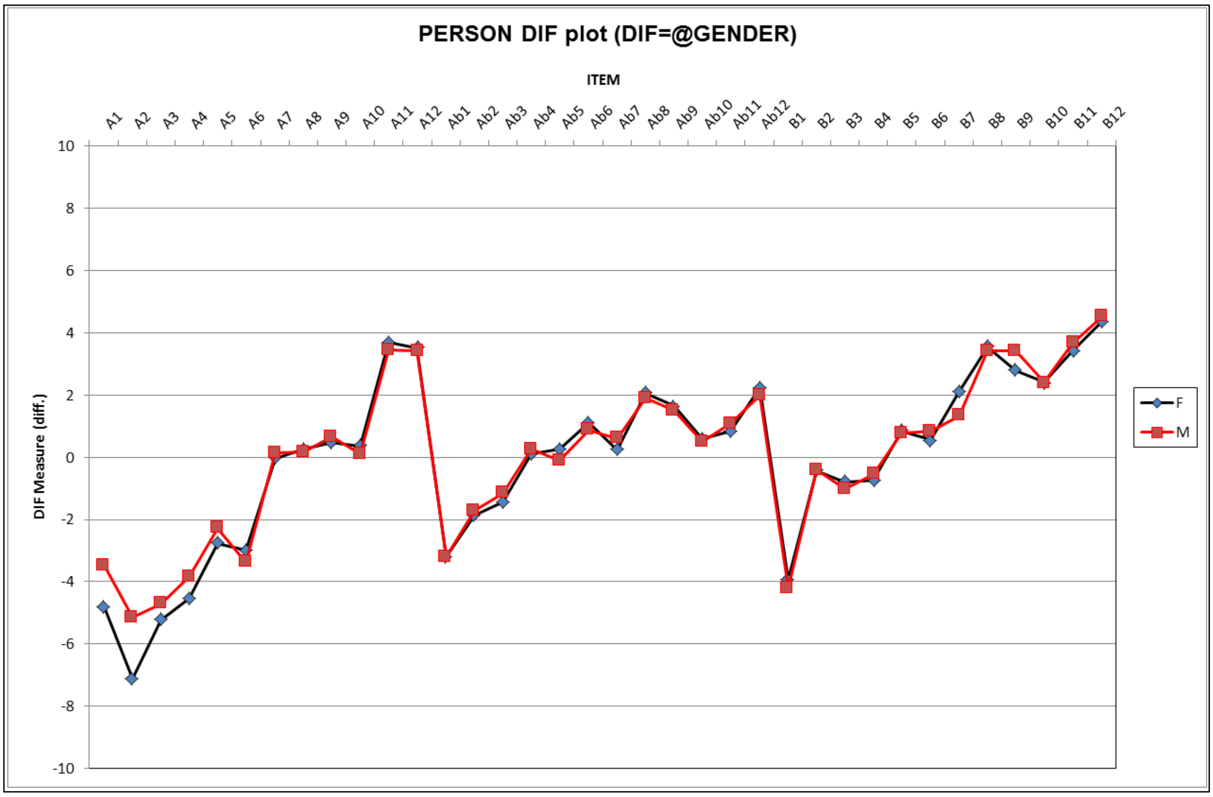
***DIF Analysis***

Another important finding in this study is the principle of measurement equivalence. Based on the results of the DIF analysis, CPM in this study seemed to violate the principle of measurement equivalence, which may suggest that CPM could be biased or favorable to one of the groups tested on several of its items (see Table 3). DIF in CPM only occurred in Section A and Section B, and not in Section Ab. This may suggest that Section Ab is an ideal section for measuring intelligence. In Section A, five items were detected to have DIF (A1, A2, A3, A4, and A5), and in Section B, it was found that two items had DIF (B7 and B9). The seven CPM items that experienced DIF in Section A and Section B were at moderate (.40 to .60) and high DIF (> .60) levels. This accords with previous research, which states that an item displays DIF if the difference in DIF values between groups reach 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, item B7, benefited 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 analysis also demonstrated the assumption of unidimensionality and local independence. 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 results of the unidimensional assumption testing 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 showed that the items in the CPM did not have a close relationship with one another. All CPM items were almost certain 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 participants were at the same stage of development, namely early childhood, with similar age range. Thus, the participants were homogeneous. Moreover, when they were examined on the basis of the educational age, they were included in the ideal age category of children to prepare for elementary school (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 in this study was smaller than that in previous studies. This may show that CPM has lower reliability score in one time measurement than in two-time measurement (test-retest) in measuring intelligence in Indonesia. The results of this study are contrary to those of the previous research, which stated that CPM is a consistent instrument (Agnoli et al., 2012). This study found four items that were not good at calibrating CPM items. These may have caused CPM to have a smaller reliability value than that of previous studies. From a psychometric point of view, there is a strong possibility that the items that are not good automatically have a tremendous impact on reliability calculations.

In addition, the research approach chosen was an important factor in determining the consistency of the instrument. This study used secondary data for analysis (one measurement), while Lehmann et al. (2014) conducted experimental research (twice measurements). It is highly likely that the research approach affects the reliability value of an instrument. The research approach that uses two measurements may have more value than the instrument's consistency and the research participant consistency. The limitation of this study is that the data was not directly collected since it could be costly and time consuming to collect data from 748 children preparing 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, this study found that each section in the CPM had different criteria for difficulty levels. For the research participants consisting of children at the early childhood stage in Indonesia, no single item was difficult to complete, especially Section Ab in the CPM. Although it may not be difficult to complete, it does not mean that it is completely easy; for example, in the CPM category item analyzer for item Ab12 of Section Ab, it is highly likely that distractors may cause children with high abilities to have difficulties in choosing answers. On the one hand, CPM is an easy test, but on the other, some distractors can make it harder for children to answer the questions.

Another 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 data on the amount of time test takers spent at Darunnisa Psychological Service, children on average spent more time on CPM item Ab12 than other items, for example item Ab12 (M = 10.67 seconds), item Ab11 (M = 10.3 seconds), item Ab12 (M = 10.3 seconds), and item Ab12 Ab5 (M = 9.3 seconds). They spent around 5.8 to 7.5 seconds for the other items in Section Ab. This finding suggests that item complexity is highly correlated with the processing time of the questions.

The indication that item Ab12 has the highest difficulty level in the Ab CPM Section is in line with the calibration results of previous research conducted in Italy. Nicotra et al. (2018) found that the most difficult items in CPM were item A11 (for Section A), item Ab12 (for Section Ab), and item B12 (for Section B). The findings of this study provide support for similar findings regarding the most difficult items in each test section. It may indicate that children in Indonesia and Italy have similarities in the level of difficulty in working on CPM items with complex patterns. This research provides more detailed information about fit items, and we believe that it is the strength of this research.

Based on the results of the DIF analysis, seven items indicated gender bias (more favorable for the male group). This finding corroborates the results of previous research (Lúcio et al., 2019; Lynn & Irwing, 2004; Sigmon, 1983). It was found that the number of bias items was quite large. When calculated, the number of CPM items identified as biased was 7 out of 36, or 19.4% of CPM items were gender-biased. In addition, six of the seven CPM items proven to be gender-biased were 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. This is in line with research that has been carried out in Indonesia (Ridho, 2014). Ridho (2014) states that men will benefit more if they are tested in the cognitive realm.

Another limitation in this study is that the data was obtained from psychological institutions. Consequently, it was difficult to analyze whether cultural differences also caused the gender bias. In the gender bias test, it was found that the test gave the male group an unfair advantage. From the demographic profile of the research participants, no variable was found on the culture of the research participants. Therefore, it is important for studies to have more control over the demographics. More diverse demographics may be able to provide more useful information, such as gender, culture, and even from different countries of origin.

## Conclusion

Based on the data analysis and discussion, eight items should not be used in intelligence testing using CPM. These are items A1, A2, A3, A4, A5, B7, B9, and B12. Items A1, A4, A5, and A7 may not be 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 participants at the early childhood stage. The issue of "norming" would be more crucial for psychological tests that are adapted from other countries or culture, such the CPM.

Further research could use the 28 items that are eligible for the CPM. It is necessary to make new norms that are adapted to children in Indonesia. In addition, CPM can be tested for cultural bias because the results of several studies in different countries find different results. Studies using CPM are likely to 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 further evaluate the reliability of CPM.

## Acknowledgement

Researchers would like to thank the support research data from Darunnisa Psychological Service.

## Conflict of Interest

Researchers declare that there are no conflict of interest regarding the publication of this paper.

## Author Contribution

YN, IF, DUM contributed to conceptualization, methodology, and writing of the manuscript; IF apply for permission to data retrieval from Darunnisa Psychological Service; YN data analysis and editing.

## 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

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory*. Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-13-7496-8

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* [PhD 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: 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' conceptions: 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. *Intelligence*, *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