Jurnal Pendidikan Dasar Nusantara Vol 11, No 1, July 2025
ISSN 2579-6461 (Online) ISSN 2460-6324 (Print)
DOI: https://doi.org/10.29407/jpdn.v11i1.26446

Validation of Dyscalculia Risk Detection Instrument in Grade 3 Elementary School Students Using the Rasch Model: Psychometric Analysis and Implications of Inclusive Education

Submitted:
July 19, 2025
Accepted:
July 28, 2025
Published:
July 31, 2025

Indra Praja Kusumah [1], M. Kusuma Wardhani [2], Dince Bunda [3]
indra.kusumah@uph.edu [1], kusuma.wardhani@uph.edu [2],
dince.bunda@uph.edu [3]
(Primary Teacher Education, Faculty of Education, Universitas Pelita Harapan) [1,2,3]

Abstract: This study aims to evaluate the psychometric quality of a dyscalculia risk screening instrument designed for grade 3 elementary school students. A total of 147 students from four elementary schools in Tangerang were involved as respondents. The instrument consisted of 10 questions and was analyzed using the Rasch model Item Response Theory (IRT) approach with jMetrics software. The results of the analysis show that the items have a fairly wide range of difficulty levels (logit –2.53 to +2.84), but the distribution is not evenly distributed. Some items show misfits for the Rasch model, as well as high error standards on extreme items. Estimated participant ability (theta) shows a logite distribution from –4.62 to +4.55, with measurement reliability at the item level very high (0.9659), but at the moderate participant level (0.5984). The correlation coefficient between the raw score and the theta estimate of 0.9861 supports the validity of the instrument construct. These findings recommend improvements to the distribution of item difficulty levels and revisions to non-model items so that the instrument can be used effectively in inclusion education for early screening of dyscalculia risk.

Keywords: Dyscalculia, Rasch Model, Early Detection, Elementary Students

PRELIMINARY

Inclusive education requires a comprehensive understanding of the learning needs of each student, including those who experience specific learning barriers such as dyscalculia. Dyscalculia is a developmental disorder in understanding numerical concepts, basic mathematical operations, and skills related to numbers and mathematical problem solving (APA, 2013; Nelson & Powell, 2018). These disorders not only impact children's academic performance, but also affect their confidence, emotion regulation, and learning motivation in the context of formal education (Peterson et al., 2017; Tang et al., 2020). Within the framework of inclusive education, it is important for teachers and educators to have accurate, objective, and reliable identification tools and strategies in detecting children with the possibility of developing dyscalculia.

The fundamental problem raised in this study is the low accuracy of identifying children with dyscalculia at the primary education level, especially in schools with limited inclusion management. Many teachers use curriculum-based assessments or informal observations that are subjective and non-standardized (Hutchison & Thomas, 2020; Skagerlund et al., 2019). This results in children with numerical learning difficulties often not being detected early and not receiving educational services that suit their needs. Therefore, an assessment instrument is needed that has high validity and reliability, and is able to measure children's numeracy skills carefully and deeply.

The approach used in this study is the modeling of Item Response Theory (IRT), which is a modern approach in psychometric measurement that focuses on the characteristics of the question item and the students' response to the item. Unlike the classical approach, IRT provides detailed information on the difficulty, differentiation, and guessing possibilities of each item, as well as allows for more precise estimation of learners' abilities at various skill levels (DeMars, 2018; Boone et al., 2020). Using IRT, the developed numeracy test instruments can be analyzed to ensure that each item is able to accurately measure the child's numerical ability, as well as detect dyscalculia indicators more specifically. The purpose of this study is to develop and analyze the quality of relevant numeracy ability identification instruments for children with dyscalculia in elementary school using the Item Response Theory approach. This study specifically aims to: (1) compile numeracy question items based on indicators of mathematics learning difficulties commonly experienced by children with dyscalculia, (2) test the quality of the question items using the IRT model, and (3) obtain a mapping of students' numeracy skills that can be used as a basis for more appropriate handling of education.

Theoretically, this research is based on two main foundations, namely the theory of dyscalculia and the theory of psychometric measurement based on IRT. Dyscalculia is classified as a part of specific learning difficulties that have diverse manifestations, such as difficulty understanding place values, number concepts, number sequences, as well as basic operations such as addition and subtraction (Butterworth & Laurillard, 2016; Nelson & Powell, 2018). Several theoretical models explain dyscalculia as a result of deficits in visuo-spatial working memory, inability to understand symbolic representations of numbers, as well as weaknesses in the automation of arithmetic facts (Tang et al., 2020). On the other hand, the theory of item response provides a strong quantitative framework for developing a high-precision measuring tool and supports the process of identifying learners based on objective psychometric parameters (Boone et al., 2020; DeMars, 2018).

This research is expected to make a significant contribution to the development of inclusive education, especially in terms of early identification and diagnostic assessment of children with numeracy barriers. The practical benefits of this study include the availability of empirically tested instruments to detect dyscalculia in elementary school-aged children. For teachers and educators, this instrument can be used as the basis for planning individualized learning interventions (IEPs). Meanwhile, from the theoretical side, this study strengthens the literature on the application of IRT in the field of special education and adds empirical references regarding numeracy assessment among students with special needs. Thus, the problem solving in this study departs from real field needs and is based on a strong methodological approach. The results of this study are expected to not only contribute to the academic world, but also have a direct impact on educational practices in inclusive schools, especially in dealing with students with specific learning difficulties such as dyscalculia.

METHOD

This study uses a quantitative approach with instrument development and validation methods, using the Item Response Theory (IRT) approach of the Rasch model. The main focus of the study is to develop and evaluate the psychometric quality of early numeracy assessment instruments to detect the risk of dyscalculia in grade 3 elementary school students. The Rasch model approach was chosen for its ability to independently generate estimates of item and individual parameters, as well as its ability to measure the unidimensionality and validity of constructs (Boone et al., 2020; Linacre, 2023). The research procedure consists of several stages, namely:

1. Literature study and identification of early indicators of dyscalculia based on basic numerical characteristics in early primary school-aged children (Butterworth et al., 2019).
2. The preparation of instrument grids based on four numerical essential domains.
3. Development of question items that are contextual and adapted to the abilities of grade 3 elementary school students.

4. Expert validation test involving three experts (two from special education and one from psychometrics).
5. Field trial on grade 3 students from several elementary schools in Tangerang.
6. Data analysis used jMetrics software to estimate the parameters of the Rasch model of one logistics parameter (1PL).

The research population included all fourth-grade elementary school students in Tangerang City. The sample was selected purposively by taking into account the variety of types of schools (public and private), the conditions of children's academic development, and the willingness of schools to be involved in research. The total participants of the study were 147 students, consisting of:

1. Public Schools: 116 students
2. Private School: 31 students
3. Male: 67 students
4. Female: 80 students

Grade 3 of elementary school was chosen because at this stage of development the child has obtained a numerical basis that is stable enough to be measured objectively (Dowker, 2019). In addition, the number of participants has met the minimum number of respondents for optimal Rasch analysis (Linacre, 2023). The instrument developed is an initial numeracy assessment tool in the form of an objective test with 10 multiple-choice items developed based on four essential domains of children's numerical abilities (Geary, 2019; Fuchs et al., 2021), namely:

1. Number Sense – the ability to understand the meaning of numbers, sequences, values, and numerical representations (2 items)
2. Number Fact – the ability to remember and use basic arithmetic facts such as addition and subtraction (2 items)
3. Calculation – the skill of performing basic calculation operations with strategies or algorithms (3 items)
4. Mathematical Reasoning – the ability to think logically in solving simple numerical problems (3 items)

The instrument grid can be seen in the following table:

ISSN 2579-6461 (Online) ISSN 2460-6324 (Print)
DOI: https://doi.org/10.29407/jpdn.v11i1.26446

Yes	Measured Domain	Key Indicators	Number of Items
1	Number Sense	Recognizing and comparing numbers; Place value	2
2	Number Fact	Considering the results of the basic addition and subtraction operations	2
3	Calculation	Solving one- and two-digit number operations	3
4	Mathematical Reasoning	Solve simple story problems and logical patterns	3

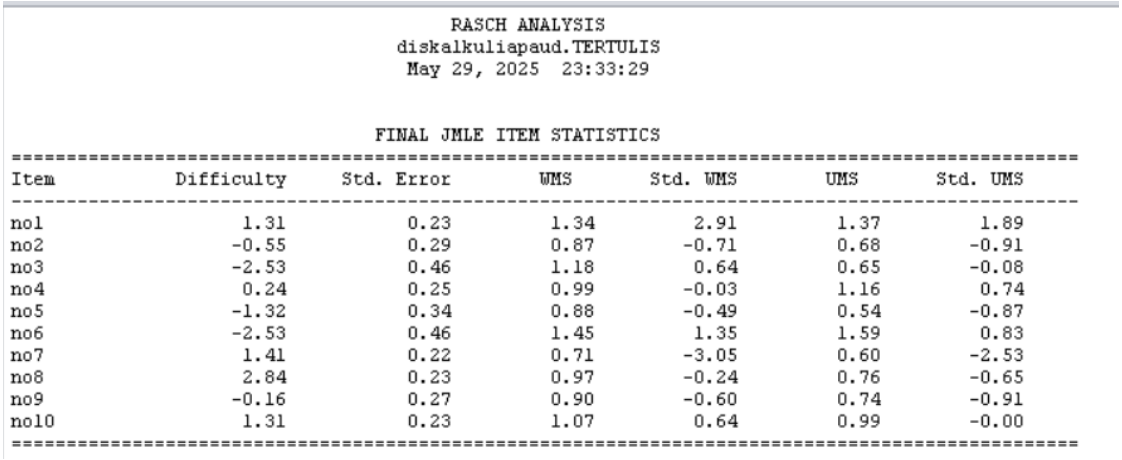
Each question is arranged with attractive visual illustrations and the context of daily life so that it is easy for students aged 8–9 years old to understand. The instruments are validated by a panel of experts and revised based on input to ensure content and language suitability. Data analysis was carried out using jMetrics software with the Rasch Model (1PL) approach. The analysis is carried out to evaluate the psychometric quality of the instrument through the following parameters:

1. Item difficulty parameter (difficulty, in logit units)
2. Standard error for each item
3. Item fit statistics, including infit and outfit mean square
4. Estimation of student ability (theta) and distribution of logit values
5. Reliability index (participant reliability and item reliability)
6. Item/person separation index and number of ability strata
7. Construct validity through correlation between raw scores and logite estimations

The Rasch model is used because it allows for fairer and more accurate measurements, especially in the context of the education of children with special needs (Boone et al., 2020). Items that indicate misfit will be evaluated for possible revision or removal in order to improve the quality of the instrument.

RESULTS

The data consisted of ten questions and were analyzed using jMetrics software. The following are the data obtained (1) item statistical data (Item Statistic), (2) participant statistical data (Score Table), (3) Scale Quality Statistical data and (4) Correlation and Covariance data of Sum and Theta Values.



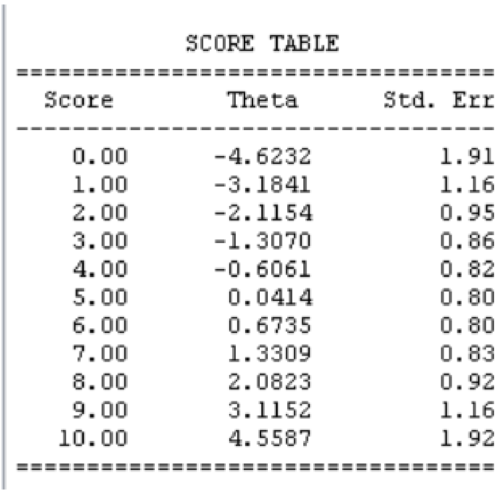
```
                              RASCH ANALYSIS
                           diskalkuliapaud.TERTULIS
                           May 29, 2025  23:33:29

                           FINAL JMLE ITEM STATISTICS
=============================================================================================
Item            Difficulty    Std. Error         WMS      Std. WMS         UMS      Std. UMS
---------------------------------------------------------------------------------------------
no1                   1.31          0.23        1.34          2.91        1.37          1.89
no2                  -0.55          0.29        0.87         -0.71        0.68         -0.91
no3                  -2.53          0.46        1.18          0.64        0.65         -0.08
no4                   0.24          0.25        0.99         -0.03        1.16          0.74
no5                  -1.32          0.34        0.88         -0.49        0.54         -0.87
no6                  -2.53          0.46        1.45          1.35        1.59          0.83
no7                   1.41          0.22        0.71         -3.05        0.60         -2.53
no8                   2.84          0.23        0.97         -0.24        0.76         -0.65
no9                  -0.16          0.27        0.90         -0.60        0.74         -0.91
no10                  1.31          0.23        1.07          0.64        0.99         -0.00
=============================================================================================
```

Figure 1 Item Statistical Test Results

Based on the results of the item parameter estimation with the Rasch model approach, the difficulty level of the item is in the range of -2.53 to 2.84 logits. Items no6 and no3 are the easiest items (logit = -2.53), while item no8 is the most difficult item (logit = 2.84). Most items have an evenly distributed level of difficulty, reflecting a good distribution of items in reaching different levels of ability of the participants.

In terms of item fit to model, the Weighted Mean Square (WMS) and Unweighted Mean Square (UMS) values are mostly in the acceptable range (0.5–1.5). However, item no6 shows an indication of overfit (WMS = 1.45; UMS = 1.59), while item no7 shows an indication of underfit (WMS = 0.71; UMS = 0.60), which can reflect a response that is inconsistent or not in line with the model's expectations. Nevertheless, in general the items in this instrument can be said to be quite valid and work according to the assumptions of the Rasch model.

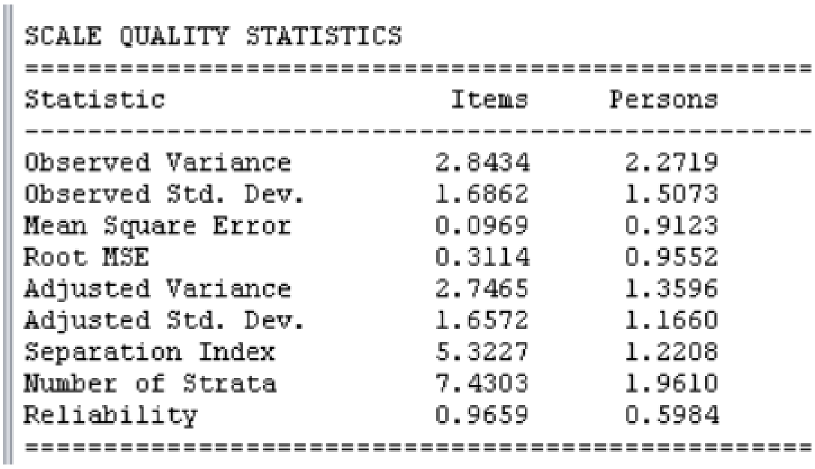


```
           SCORE TABLE
==================================
  Score          Theta     Std. Err
----------------------------------
   0.00        -4.6232         1.91
   1.00        -3.1841         1.16
   2.00        -2.1154         0.95
   3.00        -1.3070         0.86
   4.00        -0.6061         0.82
   5.00         0.0414         0.80
   6.00         0.6735         0.80
   7.00         1.3309         0.83
   8.00         2.0823         0.92
   9.00         3.1152         1.16
  10.00         4.5587         1.92
==================================
```

Figure 2 Results of the Participant Statistical Test (Score Table)

ISSN 2579-6461 (Online) ISSN 2460-6324 (Print)
DOI: https://doi.org/10.29407/jpdn.v11i1.26446

The theta score table shows the relationship between the participant's total raw score and the estimated ability (theta) based on the Rasch model. The participant's score range ranged from 0 to 10, with estimated theta values ranging from -4.6232 to 4.5587 logits. The higher the participant's score, the higher the estimate of their numerical ability. This reflects the consistency between the raw score and the latent ability measured by the instrument. The standard error for theta estimation is higher at extreme scores (both very low and very high), which is 1.91 at scores of 0 and 10. This suggests that estimation of ability in participants with extreme scores tends to be less precise, as is prevalent in Rasch modeling. In contrast, middle scores such as scores of 5 and 6 have a smaller standard error (0.80), which indicates that theta estimates in those score ranges are more accurate. Overall, these results support the assumption of unidimensionality and model consistency in mapping the relationship between scores and participant ability.



```
SCALE QUALITY STATISTICS
==================================================
Statistic                      Items     Persons
--------------------------------------------------
Observed Variance             2.8434      2.2719
Observed Std. Dev.            1.6862      1.5073
Mean Square Error             0.0969      0.9123
Root MSE                      0.3114      0.9552
Adjusted Variance             2.7465      1.3596
Adjusted Std. Dev.            1.6572      1.1660
Separation Index              5.3227      1.2208
Number of Strata              7.4303      1.9610
Reliability                   0.9659      0.5984
==================================================
```

Figure 3 Scale Quality Test Results (Scale Quality Statistics)

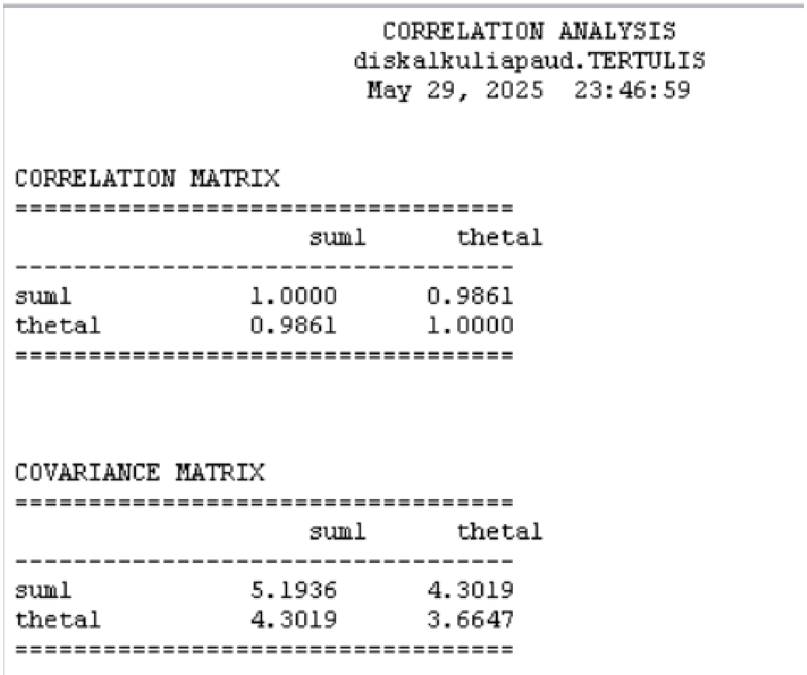
The Scale Quality Statistics table presents information about the quality of the measurement scale in terms of items and persons based on the Rasch model. In terms of grains, the observed variance is 2.8434 with a very high reliability of 0.9659, showing that this instrument has an excellent ability to distinguish the level of difficulty between grains. The separation index value of 5.3227 and the number of strata of 7.4303 also confirm that the instrument is able to group items into more than 7 strata of significantly different difficulty levels.

Meanwhile, in terms of participants, the reliability obtained was 0.5984, classified as moderate. This value indicates that the ability to distinguish between participants can still be improved, possibly due to an uneven distribution of participants' scores or because

ISSN 2579-6461 (Online) ISSN 2460-6324 (Print)
DOI: https://doi.org/10.29407/jpdn.v11i1.26446

the number of items is not optimal. The separation index of 1.2208 and the number of strata of 1.9610 indicate that this instrument is able to differentiate participants into approximately two significantly different skill groups.

In addition, the Mean Square Error and Root Mean Square Error (Root MSE) values on the participant side (0.9123 and 0.9552) were relatively high compared to the items (0.0969 and 0.3114), indicating that the estimated uncertainty was greater on the participant's ability than the item's difficulty. However, adjusted variance and adjusted standard deviation still show that the variations generated by the instrument are still reliable to measure students' numerical ability. Thus, in general, this instrument shows excellent quality in terms of items and sufficient quality on the participant side, so that it can be used to measure basic numerical ability validly and reliably.



```
                         CORRELATION ANALYSIS
                        diskalkuliapaud.TERTULIS
                         May 29, 2025   23:46:59


CORRELATION MATRIX
==================================
                    suml       thetal
----------------------------------
suml              1.0000      0.9861
thetal            0.9861      1.0000
==================================



COVARIANCE MATRIX
==================================
                    suml       thetal
----------------------------------
suml              5.1936      4.3019
thetal            4.3019      3.6647
==================================
```

Figure 4 Correlation and Covariance Data

This table presents the results of the correlation and covariance analysis between the student's total raw score (Suml) and the estimation of the logite ability of the Rasch model (Thetal) analysis. Based on the correlation matrix, a Pearson correlation coefficient of 0.9861 between Suml and Thetal was obtained, which shows a very strong and positive relationship between raw scores and students' estimated abilities. This correlation value close to 1.00 indicates that both measures have a very high consistency in representing students' numerical abilities.

This result confirms that the estimation of students' ability in logit units obtained through the Rasch model is in line with the total score obtained conventionally. However, Thetal provides more advantages because it takes into account the characteristics of the item (item difficulty), so that it is able to present a fairer and more precise estimate than a regular raw score. Meanwhile, the value in the covariance matrix shows a covariance of 4.3019 between Suml and Thetal. This suggests that changes in raw scores have a considerable linear relationship with changes in ability estimates. The covariance value of the Suml variable is 5.1936 and Thetal is 3.6647, which also reinforces the existence of a positive and stable linear relationship between the two variables. Thus, this table confirms that the use of the ability score (Thetal) in logit units as a result of the Rasch model analysis is a valid and highly consistent approach to the measurement of learners' numerical abilities based on their raw scores.

DISCUSSION

I. Item Statistical Analysis

a. Item Difficulty

The difficulty parameter in the Rasch model indicates the location of an item on the capability spectrum, expressed in logit. The higher the logit value, the more difficult it is for the item to be answered correctly by the participant (Boone, Staver, & Yale, 2014; Bond & Fox, 2015). Based on the results of the analysis, the difficulty level of the grains in this instrument ranges from –2.53 to 2.84 logits. The items with the lowest difficulty level are items number 3 and 6, which means most participants can answer correctly. In contrast, item number 8 has the highest level of difficulty, signifying that only participants with very high abilities can answer correctly.

The distribution of difficulty levels shows an imbalance, as there are some items that are classified as very easy and some that are very difficult. Items with moderate difficulty are relatively few, which can impact the sensitivity of the instrument in measuring abilities evenly (Boone et al., 2014). In the context of early childhood, items with high levels of difficulty may be less relevant due to the limitations of the child's cognitive development (Huang & Benson, 2019). Meanwhile, items that are too easy can lower the discriminating power of the instrument.

Table 1. Item Difficulty

Items	Difficulty	Interpretation of Difficulties
No3, No6	-2.53	**It's easy**
No5	-1.32	**Easy**
No2	-0.55	**Quite simple**
No9	-0.16	**Intermediate – lower**
No. 4	0.24	**Medium – upper**
No1, No10	1.31	**Difficult**
No7	1.41	**Difficult**
No8	2.84	**Very difficult**

b. Standard Error (SE)

Standard error indicates the level of uncertainty in the estimated difficulty of the item. Items with high SE are usually very easy or very difficult items due to low participant answer variation (Boone et al., 2014). Items numbers 3 and 6 have the highest SE value (0.46), as including items is very easy. In contrast, items 7 and 8 with lower SE values provided a more stable difficulty estimate because the participants' responses were more diverse (Arias, 2020).

Table 2 Item Analysis against Standard Error

Items	Std. Error
No3, No6	0.46 (**height**)
No5	0.34
No2	0.29
No9	0.27
No. 4	0.25
No1, No10	0.23 (**low**)
No7	0.22 (**low**)
No8	0.23

c. Fit Statistics: WMS (Infit) and UMS (Outfit)

The accuracy of the Rasch model to participants' responses was analyzed through infit and outfit statistics, with ideal values ranging from 0.6 to 1.4 (Linacre, 2022). Values outside this limit indicate a mismatch between the data and the model. Item number 6 has a WMS value of 1.45 and UMS 1.59, indicating potential misfit due to ambiguity or incompatibility with child development. On the other hand, items with very low fit values such as items 5 and 7 can indicate overfit, i.e. items that are too easy to predict (Karaman, 2016).

On the other hand, there are some items that are too model-appropriate (underfit), such as items numbers 5 and 7, which have WMS and UMS values below 0.8. While at

ISSN 2579-6461 (Online) ISSN 2460-6324 (Print)
DOI: https://doi.org/10.29407/jpdn.v11i1.26446

first glance it looks good, underfit can signify that an item is too "predictable" or too explicit, so it doesn't reflect the variability of the participant's actual abilities. Items such as number 1 show a WMS value of 1.34 and UMS of 1.37, are close to the upper limit that is still acceptable, but still require attention. Ideally, all items maintain a good fit with the model so that the instrument as a whole provides an accurate and bias-free estimate of capabilities.

Table 3 Item Items, WMS and UMS

Items	WMS	UMS	Interpretation
No6	1.45	1.59	**Overfit/misfit**: Response does not match the model
No1	1.34	1.37	**Height**, close to the upper limit
No3	1.18	0.65	**High WMS**, Low UMS
No10	1.07	0.99	Good fit
No. 4	0.99	1.16	Good fit
No2	0.87	0.68	**Slightly underfit**
No5	0.88	0.54	**Underfit (too deterministic)**
No7	0.71	0.60	**Extreme underfit**
No8	0.97	0.76	Good fit
No9	0.90	0.74	Good fit

d. The Relationship Between Difficulty and Fit

1. Item 6: Easy (difficulty = -2.53), misfit → too easy but instead has an inappropriate response pattern → possible ambiguity or misspelling.
2. Item 8: Difficult (difficulty = 2.84), good fit → shows that participants with high abilities respond to it consistently.
3. Item 3: It's easy, but high WMS → low consistency, needs formatting improvement.

Table 4 Summary of Key Findings Statistics Item

Aspects	Key Findings	Implication	Recommendations
Difficulty Distribution	Items are widely spread (–2.5 to 2.8) but uneven	Not all skill levels of participants are accommodated	Add intermediate items; Reduce items too easy/difficult
Fit Statistics	Some items are unfit or underfit	There are items that answer not according to the Rasch model pattern	Revision of items 6, 7, 5, 1
Standard Error	Extreme items have high SE	Unstable estimation at the end of the spectrum	Consider eliminating items with high SE and poor fit
Item Function	Some items are too easy or too difficult	Can create *a ceiling* or *floor effect*	Revise item content, use retry

The results of this analysis provide some important notes regarding the quality of the instrument:

1. The uneven distribution of difficulty levels can cause the instrument to be less sensitive in distinguishing students with moderate ability. The balance between easy, medium, and hard grains needs to be reconsidered.
2. Some items show misfits that indicate that the participants' answer patterns do not match those predicted by the Rasch model. This may reduce the validity of the construct of the instrument.
3. The presence of underfit and high errors in some items indicates that revisions to the content and format of the questions are necessary, especially to ensure that the item actually measures initial numerical abilities accurately, without bias or inconsistency with the characteristics of the child's age.
4. The high reliability of items (non-participants) (0.9659) indicates that items are generally capable of forming a consistent scale, even though some items individually exhibit anomalies.

II. Participant Statistical Analysis (*Score Table*)

a. Participant Capability Estimation (Theta)

The estimation of the participant's ability represented by the theta value shows the interval-scale logit ability. The higher the theta value, the higher the ability the participant demonstrates. Theta values represent the estimation of a participant's ability on a logit scale. The theta range from –4.6232 to 4.5587 indicates a wide range from very low to very high (Boone et al., 2014). The symmetrical distribution of theta values and their correlation to the raw score indicate that the calibration of the instrument is adequate. This indicates that the raw score reflects true ability linearly (Liu & Boone, 2021). This shows that a one-point change in the relative raw score provides a consistent difference in theta estimation, a very important characteristic in Rasch-based instruments.

b. Standard Error Estimating Capabilities

Standard error (SE) indicates a level of uncertainty in theta estimation. In general, the more extreme the score (either very low or very high), the higher the SE value. Example:

1. Score 0 (theta –4.6232) has an SE of 1.91

2. Score 10 (theta 4.5587) also has an SE of 1.92

In contrast, scores that are around the middle value (4–6) have a lower SE value, which is between 0.80 and 0.82, which reflects the most stable and most precise estimation of ability being in the middle score range. This distribution is typical in Rasch's analysis and shows that the instrument is most sensitive in distinguishing participants who are in the moderate ability range. This is also an indication that this instrument is best used for the general (non-extreme) population. For students with very high or very low ability, it is necessary to have additional special items so that the estimation of their ability is more precise.

c. Score Range and Logit Scale

The difference between theta values from a score of 0 to a score of 10 is about 9.18 logites (from –4.62 to 4.55), which indicates that the scale of the instrument is quite wide and capable of measuring significant differences in abilities between individuals. However, because standard error increases at extreme scores, the use of this instrument to identify children with very low or very high ability needs to be done with caution. The interpretation of the results needs to take into account the high uncertainty in the estimate.

Based on this score table, some of the implications that can be drawn are:

1. The instrument shows a linear and balanced measurement scale, indicated by a theta increase proportional to the raw score.
2. The most reliable capability estimation occurs at the middle score (score 4–6), which is the highest area of estimation precision (low SE).
3. Participants with extreme scores (0 or 10) had less precise estimates of ability, so caution was needed in drawing diagnostic conclusions or making important decisions (e.g. referrals for clinical assessments).
4. The instrument is effective for the purpose of screening or early classification in the general population, but it needs to be enriched if it is used to detect extreme cases (e.g. children with very severe indications of dyscalculia).

III. Scale Quality Statistical Analysis

a. Score Variability and Ability Spread

Observed Variance indicates the observed variation in score between items and participants. The values of 2.8434 (items) and 2.2719 (persons) indicate that both items and participants have a fairly good spread of scores. Variance at the item level is higher, which indicates that your question items vary quite a bit in terms of difficulty. The Observed Standard Deviation of 1.6862 (items) and 1.5073 (persons) respectively indicates a moderate to high spread of scores. This is a positive sign that the scale is not experiencing too high a homogeneity (which can lower the sensitivity of the measurement).

b. Measurement Error

Mean Square Error (MSE) and Root MSE describe the average measurement error. For items, the MSE value of 0.0969 and the Root MSE of 0.3114 are very low, indicating that the estimation of item parameters is done with great precision. On the other hand, at the persons level, MSE of 0.9123 and Root MSE of 0.9552 indicate that there is a relatively higher measurement uncertainty in the estimation of participants' abilities. This can be caused by:

1. Limited number of items (only 10 items),
2. The composition of the item has not fully reached the spectrum of the participants' abilities,
3. The response of participants who are not yet varied enough.

c. Adjusted Variance and Real Spread of Capabilities

Adjusted variance eliminates the influence of measurement errors. The Adjusted Variance values of 2.7465 (items) and 1.3596 (persons) show that this scale is better at distinguishing the characteristics of items than distinguishing participants. The adjusted standard deviation of items (1.6572) is still greater than that of persons (1.1660), which indicates that the variation in the difficulty level of the item is more distributed than the ability of the participants. This could mean that you have compiled questions with varying levels of difficulty, but the participant population tends to be homogeneous in ability, or the items do not sufficiently reach the overall variety of participants.

d. Separation Index and Number of Ability Strata

The Separation Index measures the ability of scale to separate items or participants based on the attributes being measured. Index values:

1. Items: 5.3227
2. Persons: 1.2208

A value of >2.0 for an item indicates that this scale is excellent at separating the difficulty level of the item. However, the score for participants (<1.5) is still relatively low, meaning that this instrument has not been able to clearly separate participants into different ability groups. Number of Strata indicates the number of different groups (statistically) that can be formed based on ability or difficulty level.

For items: 7.43 strata, this means that your items can be grouped into 7 difficulty levels — this is very good. For persons: 1.96 strata, which means that participants can only be differentiated into two levels of ability statistically. This is an indication that the instrument's discriminating ability against participants is still limited, and the number of items needs to be increased to remedy this.

e. Measurement Reliability

Reliability for Item: 0.9659, is very high and indicates that the estimated difficulty of the item is very stable if the instrument is reused in a group of similar participants. Reliability for Person: 0.5984, indicates moderate reliability. This value is in the lower limit of the general Rasch criterion (0.60–0.70 for screening instruments). Therefore, the reliability of participant ability measurements is not yet adequate if used for individual decision-making, but it is still acceptable for initial screening purposes.

IV. Correlation and Covariance Analysis

Based on the above data, a Pearson correlation coefficient value of 0.9861 was obtained, which shows a very strong and positive relationship between the number of raw scores and the estimation of the ability of the Rasch model. This value is close to the maximum number of 1, which indicates that the higher the number of raw scores obtained by the participants, the higher the estimated logit ability (theta) calculated through Rasch's analysis. These findings reflect that the developed instrument has high consistency between the observation results (raw score) and the theoretical measurement model (Rasch theta) (Bond & Fox, 2015). These results support the validity of the constructs of the instruments used, because both measure the same ability aspects, namely initial competencies related to numeracy or risk detection of dyscalculia in grade 3 elementary school students.

In the context of psychometrics, the high correlation coefficient between raw scores and logit estimates suggests that raw scores have practical utility as a rough indicator of ability, although they do not take into account item characteristics such as difficulty and statistical fit. Nevertheless, logit scores are still recommended to be used in more precise decision-making because they are able to provide a fairer estimation of abilities and are free from the influence of the number of items answered correctly (Wright & Masters, 1982; Bond & Fox, 2015).

In addition to correlations, the results of the covariance matrix showed a value of 4.3019 between sum1 and theta1, which is also quite high compared to the variance of each variable (5.1936 for sum1 and 3.6647 for theta1). This reinforces the proportional linear relationship between the two forms of scoring and shows that changes in raw scores are directly associated with changes in participants' estimated abilities. Overall, these results indicate that the Rasch model applied works effectively in estimating abilities based on participants' response patterns. In addition, the existence of a very high correlation is also evidence that the data obtained from the instrument has a good internal structure and is in accordance with the basic assumptions of Rasch modeling, namely unidimensionality and monotonicity (Wilson, 2005; Linacre, 2022).

CONCLUSION

The early detection of dyscalculia risk instruments developed in this study show several important strengths. First, the instrument has demonstrated high item reliability and adequate differentiation power based on the Rasch model, which means that the question items can effectively differentiate learners based on their numerical ability level. Second, the results of the analysis also indicate that the estimation of student ability (person ability) is highly correlated with the raw score, which strengthens the convergent validity of the instrument. Third, the dimensions of the numerical ability measured—including number sense, number fact, calculation, and mathematical reasoning—have

ISSN 2579-6461 (Online) ISSN 2460-6324 (Print)
DOI: https://doi.org/10.29407/jpdn.v11i1.26446

been designed based on a strong theoretical foundation, so that the construct of the instrument is considered relevant for detecting the risk of dyscalculia early.

Nevertheless, some shortcomings were also identified. The distribution of the difficulty level of the grains is still uneven, with most grains being classified as too easy or too difficult. This causes ability estimation to be less precise, especially in students with intermediate ability levels. In addition, the reliability value at the person level which is classified as indicating that the variation in students' abilities has not been fully covered by this instrument. The variability of the score also shows that the number of questions currently available is still limited to cover the full spectrum of dyscalculia risk that may arise in early elementary school classes.

Based on these findings, some of the improvements that can be made include: increasing the number of items with a moderate difficulty level to increase the distribution of logit evenly, developing new items that are more representative of the variation of learners' profiles, and considering the preparation of special sub-scales for each numerical domain so that the ability analysis can be carried out more deeply. In addition, at a later stage, it is necessary to conduct an external validity test by comparing the results of this instrument with other assessments or professional diagnostic results to strengthen the evidence of its validity.

The practical implications of this study suggest that primary school teachers, particularly in lower grades, can use this instrument as an early screening tool to identify students who have the potential to have difficulty learning mathematics, especially in the form of dyscalculia. The use of this instrument needs to be combined with classroom observation, portfolio analysis, and other qualitative assessments to form a comprehensive understanding of the student's learning profile. The application of a tiered assessment approach is highly recommended so that interventions can be provided in a timely manner and as needed.

In terms of policy implications, the results of this study underscore the importance of developing an early detection system for learning difficulties that is quantitative, data-based, and psychometrically valid in the basic education system. Local governments, inclusive school administrators, and teacher education institutions need to start integrating assessments like this into the learning needs identification system of students, as well as provide adequate training to educators in reading assessment results and designing appropriate learning follow-ups.

Theoretically, this research contributes to the development of special education assessments in Indonesia, especially in the application of item response theory in compiling early detection instruments for specific learning needs. The application of the Rasch model has been proven to provide an objective and adaptive approach, as well as in accordance with modern assessment principles that demand high accuracy and relevance to the context of students. This study opens up opportunities for further research on the development of similar instruments in other cognitive domains, as well as the exploration of differential item functioning to ensure measurement fairness across demographic groups.

REFERENCES

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Arias, V. B. (2020). Rasch Measurement and Item Response Theory. In The Palgrave Handbook of Psychometrics. Springer. https://doi.org/10.1007/978-3-030-44648-3_8

Berch, D. B., & Mazzocco, M. M. M. (Eds.). (2020). Development of mathematical cognition: Neural substrates and genetic influences. Academic Press. https://doi.org/10.1016/C2016-0-04382-3

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. (2020). Rasch analysis in the human sciences (2nd ed.). Springer. https://doi.org/10.1007/978-3-030-46444-5

Butterworth, B., & Laurillard, D. (2016). Low numeracy and dyscalculia: Identification and intervention. ZDM–Mathematics Education, 48(4), 479–490. https://doi.org/10.1007/s11858-015-0704-7

Butterworth, B., Varma, S., & Laurillard, D. (2019). Dyscalculia: From brain to education. Science, 332(6033), 1049–1053. https://doi.org/10.1126/science.1201536

DeMars, C. (2018). Item response theory. Oxford University Press.

Dowker, A. (2019). Individual differences in arithmetic: Implications for psychology, neuroscience and education (2nd ed.). Routledge. https://doi.org/10.4324/9780429283605

ISSN 2579-6461 (Online) ISSN 2460-6324 (Print)
DOI: https://doi.org/10.29407/jpdn.v11i1.26446

Fisher, W. P., & Wright, B. D. (1994). Applications of Rasch measurement. Journal of Outcome Measurement.

Fuchs, L. S., Schumacher, R. F., & Powell, S. R. (2021). Children's acquisition and generalization of arithmetic facts: The role of working memory and conceptual understanding. Journal of Educational Psychology, 113(3), 447–464. https://doi.org/10.1037/edu0000590

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. Developmental Psychology, 47(6), 1539–1552. https://doi.org/10.1037/a0025510

Geary, D. C. (2019). Development of mathematical understanding. In D. Geary, D. Berch, & K. M. Koepke (Eds.), Cognitive foundations for improving mathematical learning (pp. 13–44). Academic Press. https://doi.org/10.1016/B978-0-12-815952-1.00002-7

Huang, C.-W., & Benson, J. (2019). The Influence of Cognitive Development on Item Functioning in Young Children. Educational and Psychological Measurement, 79(6), 1135–1152. https://doi.org/10.1177/0013164419832065

Hutchison, L. A., & Thomas, M. S. C. (2020). Dyscalculia: From brain to education. Nature Reviews Psychology, 1(6), 304–316. https://doi.org/10.1038/s44159-021-00035-5

Karaman, Ö. (2016). Evaluating Test Items with Rasch Analysis: Fit Statistics and Item Invariance. International Journal of Assessment Tools in Education, 3(2), 122–134.

Linacre, J. M. (2023). A User's Guide to WINSTEPS: Rasch-Model Computer Programs. Winsteps.com.

Liu, X., & Boone, W. J. (2021). Applications of Rasch modeling in science education. In Handbook of Research on Science Education (Vol. III, pp. 419–443). Routledge.

Nelson, G., & Powell, S. R. (2018). A systematic review of instruction for students with mathematics learning disability: Implications for tiered systems of support. Exceptional Children, 84(3), 219–240. https://doi.org/10.1177/0014402917748576

Peterson, R. L., Boada, R., Saygin, Z. M., Pennington, B. F., & Fiez, J. A. (2017). Development of ventral temporal cortex is linked to reading skill in children with and without dyslexia. Developmental Cognitive Neuroscience, 25, 139–148. https://doi.org/10.1016/j.dcn.2016.12.002

Rao, S., Hofer, M., Al Otaiba, S., & Puranik, C. (2021). A systematic review of early mathematics screening tools for young children. Early Childhood Research Quarterly, 56, 1–16. https://doi.org/10.1016/j.ecresq.2021.02.003

Skagerlund, K., Träff, U., & Östergren, R. (2019). How symbolic number processing and working memory predict arithmetic development: A longitudinal study from kindergarten to Grade 3. Frontiers in Psychology, 10, 1832. https://doi.org/10.3389/fpsyg.2019.01832

Tang, Y., Zhen, Z., Sun, H., Chen, Q., Luo, Y., & Liu, J. (2020). Neural correlates of arithmetic and reading skills in children with math difficulties. NeuroImage: Clinical, 25, 102171. https://doi.org/10.1016/j.nicl.2020.102171.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (2013). Handbook of modern item response theory. Springer Science & Business Media.

Wilson, M. (2005). Constructing Measures: An Item Response Modeling Approach. Lawrence Erlbaum Associates.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. MESA Press.