Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 13, No. 1, March 2023
ISSN 2476-5880

Using Item Analysis to Evaluate a Teacher-Made PISA-Like Reading Test Model

Nia Kurniasih[1]*, Emi Emilia[2], Eva Tuckyta Sari Sujatna[3]

Received: 31 July 2022 Accepted: 7 January 2023

Abstract
This study aimed at evaluating a PISA-like reading test developed by teachers participating in the teacher training for teaching PISA-like reading. To serve this purpose, an experimental test was administered to 107 students aged 15-16 using a set of text and questions constructed according to the criteria of the PISA Reading test Level 1. Item analysis was performed following the sampling using Rasch Measurement, deemed essential for determining the ideal index of test items relative to students' ability in making the correct response. The component of the calculation comprises reliability, separation, dan standard error. The Rasch model was constructed manually using Microsoft Excel to obtain the result of the calculation, and a Wright Map was also made manually to illustrate the result of the calculation. The results of the item analysis indicate that the test and the items the teachers constructed have met good criteria. The results revealed an even distribution of test item difficulty at the targeted level. The samples' ability to make correct answers, however, was decentralized towards the test items of moderate level of difficulty. Only a limited number of students showed good ability in their response to test items of higher difficulty. These findings have the practical implication of advancing PISA-like reading test teaching and writing models by providing more information on teachers' ability to write PISA-like reading test items and the levels of difficulty of the items written by the teachers, as indicated by the students' responses to the test items.

Keywords: ability; difficulty; item analysis; PISA reading test; Rasch Measurement

1. Introduction

An important indicator to monitor a country's progress in achieving the target of sustainable development goals 4 is the proportion of students aged 15 years who achieve the minimum level of competence, i.e., achieving level 2 of the six levels of the Program for International Student Assessment (PISA) test in science reading, and math for long life learning of education (Education International Toolkit, 2017). PISA is not aimed at measuring 'school' knowledge; instead, it seeks to evaluate students' performance outside of the school curriculum. This 'outside of school' curriculum' performance test, conducted in over 90 countries, assesses the students' thinking processes required for problem-solving and decision-making in their everyday life. The basis for the approach to assessing higher-order thinking abilities is a dynamic concept of lifelong learning in which new information and skills required for effective

[1] Institut Teknologi Bandung, Email: nia.kurniasih@itb.ac.id
[2] Universitas Pendidikan Indonesia, Email: emi.emilia@upi.edu
[3] Universitas Padjadjaran, Email: eva.tuckyta@unpad.ac.id

 ISSN 2476-5880

adaptation to an environment that is continuously changing are learned throughout life (Koruts et al., 2020).

Reading literacy, mathematics literacy, and scientific literacy are the three domains that are evaluated by the PISA. These domains are highly related to the subjects that students learn at school. Yet, PISA test focuses more on the values of the acquired skills through their application in real-life situations. The questions in each PISA test demand students' problem-solving skills and the ability to reason about new conditions they encounter or will encounter. Indonesia's PISA scores in the last seven rounds, according to PISA national report 2018, have been less encouraging, especially in mathematics. In PISA 2018, however, it is the reading ability that is the weakest (OECD, 2017).

The reading literacy of Indonesian students is lower than that of the average OECD ASEAN and of a number of other countries with characteristics similar to Indonesia, such as Peru and Brazil. In countries participating in OECD, an average of 14% of students can solve level 1a questions in the field of reading. In Indonesia, the largest proportion of students' test results, about 37%, are at competency level 1a. Around 27% of Indonesian students have a competency level of 1b, implying students only can solve the easiest text comprehension problems, such as locating explicitly stated information (OECD, 2017). About 6% of the students have reading competence level 1c, implying their mastery of basic reading skills, such as understanding sentences with literal meaning, but they are unable to unify and apply these skills to longer texts or to draw simple conclusions (Hayati, 2016).

One of the recommended measures to respond to the problem to alleviate students' low reading competence as proposed in the PISA National Report 2018 is that education in Indonesia should aim at, among others, improving teacher training and professional development programs in the school where they teach (Kemendikbud, 2019). Moreover, surveys among teaching staff should be conducted to ensure the kind of training needed that specifically directs student learning to be more active, interactive, and at a high level (higher-order thinking skills or HOTs) (Badan Bahasa, 2018).

Improving the skills of elementary school teachers in teaching reading is deemed the most crucial as students' reading skills develop in the early days of elementary school. PISA 2018 results, for instance, show that SMP/MTs (junior high school) students in rural areas tend to score lower in reading literacy than their peers in other groups (Kemendikbud, 2019).

PISA defines reading literacy as "the ability to understand, use and reflect on written texts in order to achieve one's goals, to develop one's knowledge and potential, and to participate effectively in society" (Park, 2008). This literacy challenges students to complete tasks with varied types of text, reflecting the range of written forms they will encounter in their later stages of life.

The literacy program hitherto developed in Indonesia, especially in improving the reading ability of Indonesian students, such as in reading on the PISA test, has not yet shown a significant impact. Results of a previous study involving class X students conducted at the national scale in Indonesia have indicated that the students' reading ability still requires significant improvement (Badan Bahasa, 2018; Marta & Salman Alqo, 2022). Only 6.24% of the total 6.539 students were able to achieve the ability to evaluate and reflect, i.e., the highest reading ability tested in PISA.

The problem in the literacy program is that the improvement of teacher skills, which is the key to the success of the learning program, especially the PISA-equivalent reading strategy learning, has not been carried out continuously and systematically. To respond to this problem, joint-research involving three state universities in Indonesia has been conducted aiming at improving Indonesian students reading literacy through a teacher training model for teaching PISA-like reading.

This current study is part of this research whose training model and duration have followed the implementation of teacher training carried out by Emilia and Gustine in 2015, in which teacher training is carried out 6 times a year using text-based teaching (Aunurrahman et al., 2017).

As part of the teacher training model, the focus of this study was to develop a PISA-like reading test model, which will, in a later stage of research, serve as the basis for the design of an application that will contain several sets of PISA-like reading test questions with various topics, text types, and text forms, both single and multiple. Different text types are deemed crucial in the design as they are, in some measure, related to the differences in students' reading comprehension achievement (Eason et al., 2012; Şahin, 2013).

In the PISA-like reading test model this research designed, the types of text follow those of the PISA tests, namely description, recount, explanation, exposition, discussion, and procedure. Similarly, the topics of the texts used in the PISA-like reading test model are in accordance with the PISA criteria, namely topics in everyday life, such as education and profession. The other important aspect is the focus of the reading ability in the test model, which comprises three main levels as specified in the PISA standard, i.e., locating, understanding, and evaluating detailed information in the text (OECD, 2017; Park, 2008).

2. Literature Review

2.1 PISA-like Reading Test Model

In the case of PISA, students' performance is not evaluated on the most fundamental reading abilities. The 15-year-olds are required to show their proficiency in retrieving information, developing a broad general understanding of the text, interpreting it, reflecting on the content and form of the text in relation to their own knowledge of the world, and debating their own point of view.

Text content, text type, text organization, and text readability serve as the text variables for the reading comprehension of the PISA-like reading test model development. Specifically, the model for PISA-like assessments includes a framework for evaluating various types of text (mainly narrative and expository) and several comprehension levels (Santos et al., 2016). Reading tasks vary in their difficulty level depending on the text, the test-tasks (e.g., multiple-choice items, true or false questions), the readers, and the reader's engagement with the material (Alderson, 2000).

The other important aspect required in designing a test model is the assessment that will validate the quality and validity of the test items hence the test model itself. For that purpose, item analysis using Rasch Measurement Model was used.

 ISSN 2476-5880

2.2 Rasch Measurement Model

Rasch Measurement Model is used to predict the probability of a person of a given capability to provide a correct response to items of a certain level of difficulty (Hamdu et al., 2020). In other words, Rasch Measurement can reveal the relationship between a person and an item on the basis of mutual latent trait. The probability of success of conducting the test relies on the difference between the difficulty of the item and the ability of the person conducting the test (Bond, 2015).

Rasch model analysis uses two parameters in its calculation: a difficulty parameter for each item (bi) and an ability parameter for each person (θ) (Bond, 2015; Chang & Ying, 2009). These two parameters are positioned on a single logit scale, and a continuum is constructed based on the corresponding parameter values of the objects and persons. If the distance between the person's value x and the item's value y is higher, the chance of responding correctly to the item likewise increases. If the distances that favor the item are bigger – that is, if the value for the item's difficulty is greater than the person's ability - the likelihood of a successful answer decreases.

The two assumptions underpinning the Rasch measurement model are therefore that a person with higher capability has a greater likelihood of making correct answers to all items presented, and easier items are more likely to be answered correctly by all persons (Wright, 1977).

2.3 Item Analysis

Item analysis, the process of receiving, summarizing, and evaluating information from students' responses, is used to evaluate the quality of the test items, which ultimately leads to an improvement in the overall quality of the test (Shete et al., 2015). In constructing a reading test model, item analysis is particularly beneficial in that it ensures if the items fit the criteria for inclusion on a test or need modification for improvement.

The findings of the item analysis may reveal crucial information on the test-takers' responses to each question or item and how those responses are connected to the entire performance of the test (Nunnally & Bernstein, 1994). Item analysis is considered beneficial for selecting test items during their construction and revision, as well as for reviewing how effective the items are with a certain set of test takers This is because item analysis is used to evaluate the effectiveness of individual items in a PISA reading test model that was developed by teachers participating in a training.

Item revision, the stage closely related to item analysis, is used as a means for identifying and classifying items that are too difficult or too easy or questions with implausible distractors (Shin et al., 2019). In day-to-day classes, item analysis and revision can be used to determine whether an item can differentiate between students who have or have not learned materials given by their teachers. Therefore, the items can change, and instructions can be modified to fix any misunderstanding about the content (Quaigrain & Arhin, 2017).

Various studies regarding the use of the Rasch measurement model have been conducted in different fields. In one study, Rasch measurement model is preferred over other approaches due to its capability in identifying and assessing concepts that students find easy, moderate, and difficult to understand (Mahmud et al., 2013). Another study using Rasch measurement method focused on item and test quality and explored the relationship between difficulty index

(p-value) and discrimination index (DI) with distractor efficiency. Item analysis is deemed essential in improving items which will be used again in later tests of this research. It was also used to eliminate misleading items in a test (Quaigrain & Arhin, 2017). Crossley et al., (2011) used Rasch Wright Map to assess students' conceptual understanding of electricity, while Soeharto & Csapó, (2022) conducted a similar study among Indonesian students, focusing on science education.

Models that are founded on item response theory have seen extensive usage in the context of measurement applications in social domains including educational assessment and language testing (Boone et al., 2011; Boone & Scantlebury, 2006; Wilson & Moore, 2011). In their work, Santos et al., (2016) demonstrate that the Rasch model contributes significantly to the measuring of reading comprehension, particularly in multiple-choice tests. Their study aimed to investigate the psychometric properties of the items, assess the dimensionality of the test forms (including local independence and reliability), and perform vertical scaling of the test forms of the TRC-n and the TRC-e.

Other researchers utilized the Rasch Measurement model to evaluate the English Paper 1 (EP1) items of the UPSR trial examination for sixth graders in terms of their reliability, validity, item characteristics, and difficulty levels, which included vocabulary, language and social expression, grammar, cloze comprehension, and reading comprehension (Mokshein et al., 2019). The reliability indices suggest that the EP1 test items are repeatable across all samples with comparable skills. The distribution pattern of item difficulty and student ability demonstrates a decent fit between the two despite some that are beyond the range of student capability.

Very few studies have utilized the Rasch Wright Map to identify students' reading ability and item difficulty related to the PISA reading test. This current research discusses the outcomes of an experimental PISA-like reading test prepared by 24 teachers participating in a teacher training for teaching PISA reading skills.

When developing test items, test developers need to execute one of the most important statistical procedures, i.e., item analysis. How difficult a particular item might be and how valid the test is are mostly based on speculation. Test developers need the data resulting from a field trial to be used as corroboration or refutation to their earlier expectations (Aryadoust & Raquel, 2019). The objective of the experimental test conducted in this study was, therefore, to assess the items and the test quality as well as to explore the relationship between item difficulty and students' ability hence the teacher training was deemed to meet its objectives.

3. Method

Following the reviewing process of the PISA-like reading test items written by the teachers participating in the training, only a set of questions were evaluated for the purpose of testing the level of difficulty of the test items due to the limited time and complexity of conducting the experimental test to students during the pandemic.

Sampling during the Covid-19 pandemic was carried out online, utilizing a network of English teachers at state high schools. The respondents conducted the test after class, hence there was no interference with their learning process. The consequence was that students were not in their best condition after participating in the learning process.

 ISSN 2476-5880

Table 1 shows the classification of respondents taken, students aged 15 to 16 years old, as a representation of the upper secondary class (grade X) students in the male and female gender. No further classification by grade level was made because in certain cases there were students younger than their peers in school but belonged to a higher grade, or vice versa. The educational regulations in Indonesia allow students with good academic ability to accelerate their studies; on the contrary, there are also students who enter the next stage of education late. Therefore, the classification was carried out in less detail. Similarly, gender differences are not used as a reference in the sampling method on the grounds that the experimental test was carried out in state high schools, not male-only or female-only schools hence male and female students were assumed to receive the same type of learning.

Table 1
Sampling Detail

Parameters	Group
Age	15-16 (Years)
Teaching Level	X
Teaching Area	English
Gender	Male and Female

The criteria of the parameters determined for sampling (see Table 1) were adjusted to the needs of the study and the sampling method commonly used in similar studies. Then, the data obtained were processed using the Rasch method. The results were then analyzed using statistical and linguistic approaches. This process is important to determine the ideal scale of the question item on the student's ability in making correct answers. Important components in the calculation of the Rasch model include reliability, separation, and standard error. The Rasch model was built manually using Microsoft Excel to get the result of the calculation. Furthermore, a Wright Map was visually made with the aim of making it easier for readers, in general, to understand the results of the research conducted.

The objective of this experimental test was to assess items and test quality and to explore the relationship between item difficulty and students' ability. To suit the purpose, a set of items of the PISA-like reading test composed by the teachers involved in the training were selected consisting of the following.

1. A text with the characteristics (PISA-like) is detailed in Table 2.

Table 2
Text Characteristics

Situation	: Education
Text type	: Biography Recount
Taken from	: https://www.hellomagazine.com/profiles/queen-elizabeth-ii/
Total word count	: 406
Hard Words	: 49 (12,07%)
Lexical Density	: 57,88

2. A set of questions consisting of 9 items of PISA-like reading test items with the following characteristics. The original set was designed with ten (10) questions, but only nine (9) were considered valid for the test and the analysis.

Table 3
Summary of Framework Characteristics Items

Item No.	Aspect	Question Intent	Item Format
1	Access and retrieve information from text	locate information explicitly stated	Closed Constructed Response
2	Access and retrieve information from text	locate information explicitly stated	Multiple choice
3	Reflect and evaluate content	identify whether each statement is relevant to the text	Complex multiple choice
4	Reflect and evaluate content	identify whether each statement is relevant to the text	Complex multiple choice
5	Reflect and evaluate content	identify whether each statement is relevant to the text	Complex multiple choice
6	Reflect and evaluate content	identify whether each statement is relevant to the text	Complex multiple choice
7	Reflect and evaluate content	identify whether each statement is relevant to the text	Complex multiple choice
8	Access and retrieve information from text	locate information explicitly stated	Multiple choice
9	Access and retrieve information from text	locate information explicitly stated	Multiple choice

Unlike the other items in the test, question 3 is broken down into five questions (items 3 to 7), asking about the relevance of the information in the text.

3. The time allocated for the samples (students) to do the test was 10 minutes.
4. For the purpose of Rasch measurement, a correct answer for each item was given a value of 1 (one) and an incorrect answer was given a value of 0 (zero).

4. Results

The results of the experimental test show 97 out of 107 samples were deemed valid for analysis. For a measurement to be accurate, the item's difficulty must correspond to the person's ability (Hayat et al., 2020), and misfit may result from ambiguous item wording, random answers or

 ISSN 2476-5880

distraction, and lack of cooperation and motivation (Bond, 2015). Based on the results of the experimental PISA-like reading test, the following result (Figure 1) was obtained.

Figure 1
PISA-like Reading Test Score

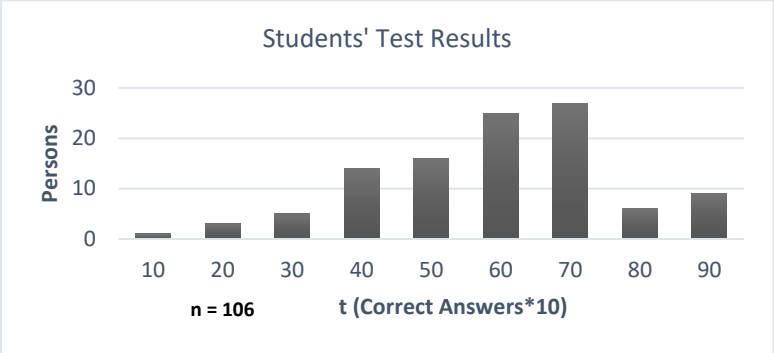


Figure 1 shows the majority of students involved in the experimental test have done the test relatively well with 68 students out of a total of 107 or 64% scoring between 50 – 70 (out of a maximum of 90), showing an average ability in performing the test; 23 students or 21% scored between 10 – 40, implying lower ability, and only 15 students or 14% scored between 80 – 90 (perfect score).

Table 4
Summary of Test Results Based on Rasch Measurement

	Reliability	Separation
Person	0,75	1,01
Item	0,91	2,18

Table 4 reveals that the average person reliability value from the sampling results is 0.75 and the item reliability is 0.91, indicating that the reliability index value is > 0.70 and hence categorized as adequate. It means that there is an interaction between item difficulty and students of good ability. Person reliability of 0.75 implies that the instrument used is relatively sensitive to respondents who are categorized as having good (high), moderate, and low abilities. The high-reliability score of 0.91 informs a very good question differentiation, i.e., the level of difficulty is evenly distributed. The unsatisfactory result is shown by the person separation of 1.01. This situation indicates that the students' cognitive abilities are not good, only dominated by one or a few abilities. The separation threshold is 2> hence item separation is moderately good.

 ISSN 2476-5880

Table 5
Item Fit to Statistic Difficulty and Standard Error

Item	Logit (difficulty)	SE (Standard Error)	Level of difficulty	Type
1	0,24	0,22	Slightly difficult	CCR
2	0,73	0,22	Slightly difficult	MC
3	-0,5	0,23	Easy	YES/NO
4	-2,27	0,34	The easiest	YES/NO
5	2,03	0,24	Difficult	YES/NO
6	-0,43	0,23	Easy	YES/NO
7	-0,76	0,24	Slightly Easy	YES/NO
8	-0,89	0,24	Slightly Easy	MC
9	0,94	0,22	Slightly difficult	MC

Table 5 shows the level of difficulty of the items of the sample test. One question was found to be very easy and successfully and correctly answered by all samples with a 100% answer success rate hence elimination from the calculation. The value of the logit item as an index of problem difficulty is in the range of -2.27 to 2.03 indicating that the distribution of the questions is fairly good. It was also revealed that the question with the highest level of difficulty is question number 5.

This research differs from previously conducted and published studies, as can be seen in Table 6. Some limitations in each of these studies have been found. For instance, there are still several methodologies and instruments to investigate similar cases. In addition, sampling is restricted to a certain location, region, and age or gender group. The void created by these prior studies, therefore, must be filled with research employing different methodologies, hence the data acquired enhance educational literacy, particularly within the context of PISA.

Table 6
Previous Related Studies

Author	Purpose	Method	Finding
(Sharfeldin, 2021)	To identify the effect of integrating critical thinking skills in the Arabic language reading curriculum for seventh graders to improve the performance of PISA Qatari Students	Descriptive-analytical method	Shows a clear improvement in the performance of Qatari students in the PISA test in general, as they made remarkable progress between 2006 and 2015. This indicates the positive effect of integrating critical thinking

 ISSN 2476-5880

			skills into the reading curriculum.
(Durrant & Brenchley, 2019)	To improve public's understanding of children's use of vocabulary in changes in their writing as they progress through their school careers.	Examining the extent to which a model of lexical sophistication as use of low-frequency, register-appropriate words, adequately captures development in vocabulary use (distribution of texts across year groups, genres, and disciplines).	As they mature, children tend to make greater use of academic vocabulary in both their literary and non-literary writing, though this increase is greatest in their non-literary writing.
(Marta & Salman Alqo, 2022)	To describe class X students' ability in reading exposition texts in SMA Negeri 1 Bungaraya (Siak) during the odd semester 2021/2022. .	The analysis uses inferential statistical procedures, i.e., one sample t-test and unidirectional ANOVA test using SPSS facility, in which the exposition text test instrument is made in two languages, namely Indonesian and English.	Through the two-way ANOVA test approach, it was shown that there were no differences in sample group-based reading comprehension of exposition texts. In other words, neither science class nor science and social studies class is superior to the other. This is a normal occurrence because they are all taught by the same instructor.

5. Discussion

5.1 Student performance in PISA-like Reading Test Model according to the Wright Map

The results of the calculation show the mean obtained is μ*Person* = 0.5598, implying a positive value. It is therefore evident that the 9 items in the test are not deemed challenging by these respondents as the mean value for the item, μ*item,* is -0.6013, meaning greater than the μ*Person* = 0.5598. The items can also be classified as easy or difficult depending on their position on the map. Items located on the right side of the map and those located above the mean value (line 0) are ones with a higher level of difficulty.

ISSN 2476-5880

5.1.1 Easy items

The easiest is item number 4 with -2,2673, meaning most students did not find any substantial difficulty in identifying explicitly stated information in the text. Item number 3 and 6 with their respective -0,4978 and -0,4357 logits are questions of the same aspects. Items 3 and 6 deal with the aspect of reflecting and evaluating the content, while the question intends to identify whether each statement is relevant to the text. Both are of complex multiple choice type format with two options yes or no. The values of the difficulty of items 3 and 6 are located below the mean value, indicating that the students did not experience any significant problems in reflecting and evaluating the content of the text. Item number 7 at -0,7583 logit and item 8 at -0,8977 are considered easy. Item 7 deals with the aspect of reflecting and evaluating content while the question intends to identify whether each statement is relevant to the text. It is of a complex multiple-choice type format with two options yes or no. Item 8 deals with the aspect of accessing and retrieving information from text, while the question intent is to locate information explicitly stated in the text. Item 8 is of closed constructed response, a multiple choice with three (3) options. The difficulty values of items 7 and 8 that are located below the mean value show that the students had only a slight problem in locating the specific information.

5.1.2 Difficult items

Item 1 is considered slightly difficult as it is located at +0.2490 logit above the mean value, meaning students found only a slight difficulty in locating information explicitly stated. The aspect of the question is to access and retrieve information from the text, and the type of format closed structure response, a multiple choice of three options. Items 2 located at +0,7266 logit and item 9 at 0,9376 logits are both above the mean value in the map. This indicates that most of the students found no ease yet no significant difficulty in accessing and retrieving the specific information from the text (item 2) and in identifying whether each statement is relevant to the text (item 9). Item 2 is of closed structure response, a complex multiple choice with two options (yes/no), while item 9 is a multiple choice with three options. This shows that the response type format does not significantly determine their difficulty in choosing the answer. Item 5 is located relatively the highest at +2,0270 logit above the mean value, indicating the students found significant difficulty in accessing and retrieving the specific information tested from the text. Item 2 is of complex multiple choice with two options (yes/no).

The range persons' measure distribution is indicated in the model of the Wright Map of the PISA-like reading test as follows. Items with the most difficulty are located at the top left of the map, while the easiest is at the bottom left of the map. This indicates that students with a higher ability (higher critical thinking) in comprehending information in the text were located in the upper left quadrant of the map.

The bottom left of the map shows students with lower ability in understanding items of higher-order thinking types that require students' HOTS. Students are deemed to have a 50:50 probability of making correct answers when their logit measure is located at the same difficulty level as an item. Students whose logit measure in ability is higher than that of item difficulty are shown to have a bigger probability, higher than 50 percent, to respond correctly to that particular item. Likewise, students have less probability, lower than 50 percent, to make correct

 ISSN 2476-5880

answers for a particular item if their logit measure is located at a level lower than that of the difficulty level.

Figure 2 shows a Wright Map of the items being tested and the level of the samples' ability. The test results reveal that the difficulty of the questions is evenly distributed. However, the level of the sample's ability to make correct answers tends to be decentralized to items with moderate difficulty. In other words, only a few samples have the ability to work with a high level of difficulty.

Figure 2
Wright Map of Students' Ability and Item Difficulty

The students, as depicted in Figure 2, have a relatively good comprehension of the text. Reading aspects as presented in the items are situated above the +0.0-logit mark, only two are respectively situated significantly above and below the mean value. This means that these items, with the mean of students' ability (μ*person* = 0,56) and mean of item difficulty (μ*item* = 0,60) are deemed of an average level of difficulty and hence not very challenging for the students.

On average, the students experienced no major difficulties in solving these problems. Item 5 has the highest logit measure of the nine (9) items, indicating it is the most difficult, confirmed by the student responses to item 5 which showed 24 out of 97 students answered correctly.

5.2 Proposed model for PISA-like Test Item Writing

Based on the results of the experimental test as well as evaluation of the overall process of teachers' writing the items that may qualify the standards of the PISA reading test. To achieve the expected result of increasing students' ability to respond to items of a higher level of difficulty, teachers need to have their classroom instruction synchronized with the test items. This calls for teacher's ability to develop good test items and of analyzing the items leading to teachers' understanding and use of statistical analysis of the test materials to improve the test construction and the teachers' teaching strategies (Quaigrain & Arhin, 2017; Saka, 2016). A modified RASCH measurement, i.e., item analysis, is among the tools teachers may find useful for such purposes.

Composing a reasonably good set of test items begins with selecting the right kind of texts that greatly represent the real occurrences of the day-to-day life of test takers, in this case, students of age 15 to 16. This refers to the choice of words, the number of words used, and the diversity of the vocabulary used in the text (Crossley et al., 2011; Durrant & Brenchley, 2019;

Olinghouse & Wilson, 2013). This can be measured based on lexical density, reading levels, or the FOG Index, of the text, which is an indication of how easily the work can be read or index of readability (Nurhayati & Kurniasih, 2016). Altogether the texts should also offer chances for teachers to compose questions that meet the characteristics required for PISA reading test. The findings of this current research also confirm a study by Yu, Xiaoli (2021), which suggested that authentic materials and extensive reading offer promising values for English learners. The authentic texts teachers used in preparing the PISA-like reading comprehension tests of this current research has significantly helped teachers in writing the test items during and after the training.

As for composing items with an increased level of difficulty, teachers' mastery of high order thinking skills (HOTS), i.e., the top three levels in Bloom's Taxonomy: analysis, synthesis, and evaluation, is deemed considerably beneficial. Referring to the items written by the teachers, prior to the reviewing process, many still belong to the lower order thinking skills (LOTS) that involve memorization. To write items of higher-order thinking (HOTS), understanding and application of that knowledge are prerequisite, even on a test that consists primarily of multiple-choice questions (Semsar & Casagrand, 2017).

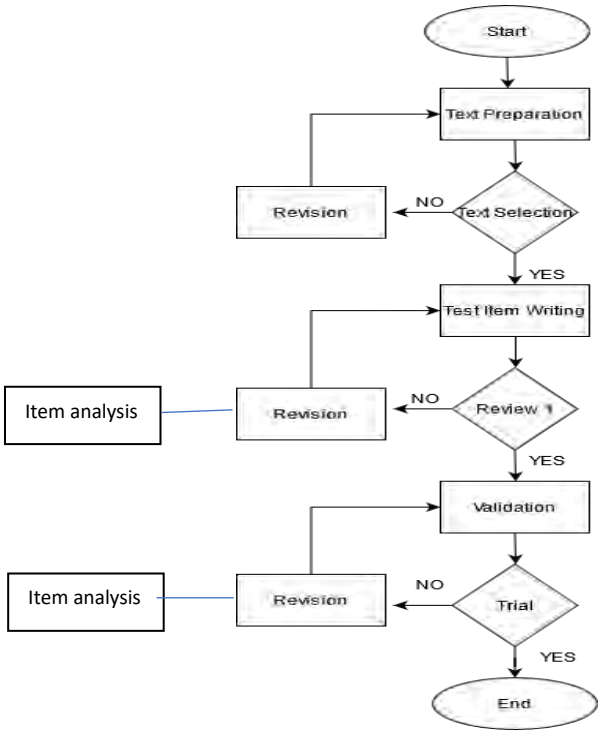
Dufresne et al. (2002) state that in a test that uses multiple-choice items, students' success in performing a test is determined by how distractors are developed. Students' consistent failure in choosing the correct option in a multiple-choice item type may have nothing to do with the item itself but rather with implausible options. Composing distractors is, therefore, equally important in determining the relative usefulness of each item of the test. As distractors indicate a relationship between the distractor chosen by the student and the total test score, teachers are required to enhance their mastery of designing plausible distractors.

Based on the results of the experimental PISA-like reading test detailed above and according to the theoretical discussion regarding the needs to enhance teachers' ability in composing test items (Lahay, 2022; Mohammadkhah et al., 2022; Quaigrain & Arhin, 2017). This study proposes the writing model of a PISA-like reading test item as depicted in Figure 3 as part of the bigger research scheme whose primary objective is to develop a training model for class X teachers.

 ISSN 2476-5880

Figure 3

Proposed Model of Using Item Analysis for Evaluating Teacher-developed PISA-like Reading Test Items



5.3 Novelty of this proposed model for PISA-like Test Item Writing

Existing studies on PISA employing the Rasch method are still scarce and tend to focus on a small number of specific case regions in Indonesia. The aim of this study is therefore to fill in the voids left by previous studies, addressing issues that have not yet been investigated by either Indonesian researchers or those overseas. The variables in Rasch's inputs were subject to improvisation, beginning with gender, age range, and levels of education. The research conducted by Durrant and Brenchley (2019) uses the sampling principle to collect data from kindergarten students and tracks the development of students' writing skills throughout their elementary schooling. These conditions are highly intriguing because to assess the procedure used, it is important to carry out sampling in various situations. Sharfeldin (2021) took a different approach, conducting research on reading ability from a sample population. Marta and Alqo (2022) have conducted a study along these lines in Indonesia, utilizing a representative sample of senior high school students from class X. The researchers used an ANOVA-based analytic method in their research.

This study has utilized grouping method previously used by Durrant and Brenchley (2019) with a sample of students aged 16-17 years (grades X/XI junior and high schools) in a reading test study following PISA reading test rules. Previous studies only used samples in the same school and in one area/region, while the samples used in this study were taken from two different provinces so that the sample variations were not polarized. This study also indicates its specific objectives, i.e., to determine the level of difficulty of the questions on the PISA-like reading test written by teachers participating in the training, as indicated by the students'

ability to answer the PISA-like prototype test items. Furthermore, the current research proposes a model for the development of the PISA-like reading test in Indonesia, making this model country specific.

6. Conclusion

Item analysis was deemed essential to validate and assess the quality of the texts prepared and items written by the teachers. The items were written following the criteria for PISA 2018 comprising PISA reading skills of locating explicitly stated information and identifying whether each statement is relevant to the text. A sampling of 107 students aged 15-16 was performed through teacher network and the test was conducted online in 10 minutes for 10 questions. The results, through the Rasch item analysis, are in conjunction with the analysis of the items' content. It was found that the test items developed by the teachers met the criteria for a good PISA-like Level 1.

The whole process of developing the test, starting from selecting texts up to performing the experimental analyzing the test items have revealed that the PISA-like Reading test model has fulfilled the target set in the teacher training, i.e., to qualify for standards of PISA reading test Level 1. To achieve the expected result of increasing students' ability to respond to items of a higher level of difficulty, teacher trainings on this topic that use the training model proposed in this research need to be continued with wider coverage and more detailed aspects to be measured. It can also be concluded that the use of the Rasch Measurement model to analyze the items is deemed helpful for improving teachers' strategies in teaching and in constructing tests. This study has implemented sampling with various spatial data, where samples were taken from two different provinces. Thus, in future research, spatial analysis can be applied to obtain better information regarding the PISA test in Indonesia, which has a wide area coverage.

Acknowledgment

The authors received funding for this research from the Program Penelitian Kolaborasi Indonesia (PPKI) scheme. We'd also like to express our gratitude to LPPM Institut Teknologi Bandung, LPPM Universitas Pendidikan Indonesia, and LPPM Universitas Padjadjaran as parts of the PPKI scheme.

Declaration of Conflicting Interests

The authors of this work disclosed no potential conflicts of interest with its research, authorship, or publishing.

References

Alderson, J. (2000). *Assessing reading*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511732935

Aryadoust, V., & Raquel, M. (Eds.). (2019). *Quantitative data analysis for language assessment volume I: Fundamental techniques (1st ed.)*. London: Routledge. https://doi.org/10.4324/9781315187815

Aunurrahman, A., Hamied, F. A., & Emilia, E. (2017). A joint construction practice in an

academic writing course in an Indonesian university context. *Celt: A Journal of Culture, English Language Teaching & Literature*, *17*(1), 27-44. https://doi.org/10.24167/celt.v17i1.1137

Badan Bahasa. (2018). *Laporan kajian kemampuan membaca siswa kelas X di Indonesia*. Kemdikbud. Jakarta: Badan Bahasa Kemdikbud.

Bond, T. (2015). *Applying the Rasch model*. New York: Routledge. https://doi.org/10.4324/9781315814698

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*(2), 253–269. https://doi.org/10.1002/sce.20106

Boone, W. J., Townsend, J. S., & Staver, J. (2011). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education*, *95*(2), 258–280. https://doi.org/10.1002/sce.20413

Chang, H.-H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, *37*(3), 1466-1488. https://doi.org/10.1214/08-AOS614

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*(3), 282–311. https://doi.org/10.1177/0741088311410188

Dufresne, R. J., Leonard, W. J., & Gerace, W. J. (2002). Marking sense of students' answers to multiple-choice questions. *The Physics Teacher*, *40*(3), 174–180. https://doi.org/10.1119/1.1466554

Durrant, P., & Brenchley, M. (2019). Development of vocabulary sophistication across genres in English children's writing. *Reading and Writing*, *32*(8), 1927–1953. https://doi.org/10.1007/s11145-018-9932-8

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader–text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology*, *104*(3), 515–528. https://doi.org/10.1037/a0027182

Education International Toolkit. (2017). *Guide to indicators for SDG 4 quality education*. Education International. https://download.ei-ie.org/Docs/WebDepot/2017_SDGs_Toolkit_eng_v1.1.pdf

Hamdu, G., Fuadi, F. N., Yulianto, A., & Akhirani, Y. S. (2020). Items quality analysis using Rasch model to measure elementary school students' critical thinking skill on stem learning. *JPI (Jurnal Pendidikan Indonesia)*, *9*(1), 61-74. https://doi.org/10.23887/jpi-undiksha.v9i1.20884

Hayat, B., Putra, M. D. K., & Suryadi, B. (2020). Comparing item parameter estimates and fit statistics of the Rasch model from three different traditions. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(1), 39-50. https://doi.org/10.21831/pep.v24i1.29871

Hayati, A. (2016). The Correlation between Indonesian students' vocabulary mastery and their reading comprehension. *Al-Ta Lim Journal*, *23*(2), 95–99. https://doi.org/10.15548/jt.v23i2.217

Kemendikbud. (2019). *Pendidikan di Indonesia: Belajar dari hasil PISA 2018*. https://repositori.kemdikbud.go.id/16742/1/Laporan Nasional PISA 2018 Indonesia.pdf

Koruts, U. Z., Petkov, V. P., Nazymko, E. S., Denysova, T. A., & Oliinyk, U. M. (2020). Formation of lifelong learning competences in the process of professional training of future lawyers. *International Journal of Learning, Teaching and Educational Research*, *19*(4), 130–149. https://doi.org/10.26803/ijlter.19.4.9

Lahay, M. Y. (2022). Improving teacher ability in preparing daily test assessment questions through workshops. *Pinisi Journal of Education and Management*, *1*(1), 67–78. https://ojs.unm.ac.id/PJoEM/article/view/33031/pdf#

Mahmud, Z., Ghani, N. A. M., & Rahim, R. A. (2013). Assessing students' learning ability in a postgraduate statistical course: A Rasch analysis. *Procedia - Social and Behavioral Sciences*, *89*, 890–894. https://doi.org/10.1016/j.sbspro.2013.08.951

Marta, F., & Salman Alqo, D. (2022). Kemampuan membaca teks eksposisi siswa kelas X SMA. *Jurnal Pembelajaran Bahasa Dan Sastra*, *1*(1), 53–64. https://doi.org/10.55909/jpbs.v1i1.13

Mohammadkhah, E., Kiany, G. R., Tajedin, Z., & Shayestefar, P. (2022). Teachers' conceptions of language assessment: Affective and theoretical knowledge dimensions of language assessment literacy model. *International Journal of Language Testing*, *12*(1), 82–102. https://www.ijlt.ir/article_146986_2da1bde1e143f415de88d27ab136571d.pdf

Mokshein, S. E., Ishak, H., & Ahmad, H. (2019). The use of Rasch measurement model in English testing. *Jurnal Cakrawala Pendidikan*, *38*(1), 16–32. https://doi.org/10.21831/cp.v38i1.22750

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill, Inc.

Nurhayati, I. K., & Kurniasih, N. (2016). Improving academic writing standard: A challenge for universities. *Pertanika Journals Social Sciences & HUmanities*, *24(5)*, 187–204.

OECD. (2017). *PISA for development assessment and analytical framework: Reading, mathematics and science (Preliminary)*. Paris: OECD Publishing.

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing*, *26*(1), 45–65. https://doi.org/10.1007/s11145-012-9392-5

Park, H. (2008). Public policy and the effect of sibship size on educational achievement: A comparative study of 20 countries. *Social Science Research*, *37*(3), 874–887. https://doi.org/10.1016/j.ssresearch.2008.03.002

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013. https://doi.org/10.1080/2331186X.2017.1301013

Şahin, A. (2013). The effect of text types on reading comprehension. *Mevlana International Journal of Education, 3(2)*, 57–67. https://doi.org/10.13054/mije.13.27.3.2

Saka, F. ÿzlem. (2016). What do teachers think about testing procedure at schools? *Procedia - Social and Behavioral Sciences*, *232*, 575–582. https://doi.org/10.1016/j.sbspro.2016.10.079

Santos, S., Cadime, I., Viana, F. L., Prieto, G., Chaves-Sousa, S., Spinillo, A. G., & Ribeiro,

I. (2016). An application of the Rasch model to reading comprehension measurement. *Psicologia: Reflexão e Crítica*, *29*(1), 1-19. https://doi.org/10.1186/s41155-016-0044-6

Semsar, K., & Casagrand, J. (2017). Bloom's dichotomous key: a new tool for evaluating the cognitive difficulty of assessments. *Advances in Physiology Education*, *41*(1), 170–177. https://doi.org/10.1152/advan.00101.2016

Sharfeldin, S. (2021). The impact of integrating critical thinking skills in reading curriculum in developing Qatari students performance in PISA exams. *Global Journal Al-Thaqafah*, *11*(1), 102–122.

Shete, A., Kausar, A., Lakhkar, K., & Khan, S. (2015). Item analysis: An evaluation of multiple choice questions in physiology examination. *Journal of Contemporary Medical Education*, *3*(3), 106-109. https://doi.org/10.5455/jcme.20151011041414

Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in Psychology*, *10, 285*. https://doi.org/10.3389/fpsyg.2019.00825

Soeharto, S., & Csapó, B. (2022). Exploring Indonesian student misconceptions in science concepts. *Heliyon*, *8*(9), 1-10. https://doi.org/10.1016/j.heliyon.2022.e10720

Wilson, M., & Moore, S. (2011). Building out a measurement model to incorporate complexities of testing in the language domain. *Language Testing*, *28*(4), 441–462. https://doi.org/10.1177/0265532210394142

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116. https://doi.org/10.1111/j.1745-3984.1977.tb00031.x

Yu, X. (2021). Text complexity of reading comprehension passages in the national matriculation English test in China: The development from 1996 to 2020. *International Journal of Language Testing*, 11(2), 142-167.