Kasetsart Journal of Social Sciences

journal homepage: http://kjss.kasetsart.org

Description of students' abilities using estimated difficulty level of test items: An Indonesian case study

Rukli Rukli*

Mathematics Education Study Program, Faculty of Teacher Training and Education, Muhammadiyah University Makassar, Makassar 90221, Indonesia

Article Info

Article history:
Received 4September 2023
Revised 27 January 2024
Accepted 6 March 2024
Available online 28 March 2025

Keywords:
assessment for learning,
case study,
description of students' abilities,
difficulty level of test items,
mathematics

Abstract

In the concept of classical test theory, the difficulty level is one of the characteristics of certain test items that is directly related to the level of student ability, but the level of difficulty of the items is not considered in detail in examining student work. This study aims to describe the pattern of students' ability to solve math test items using the estimated of the difficulty level of the test items. This study used an exploratory-descriptive cross-sectional study with a case study approach. The research subjects were 21 elementary school students. The time allocation for working on one item is three minutes while the estimation of the characteristics of the item is one minute. The results of the research show the following things. Grade IV students experienced errors in working on test items in all content domains, especially the geometry and data domains. Then, fifth-grade students experienced more errors in the cognitive domain, especially the cognitive domain of knowing. In general, students are only able to answer low-category test items with the cognitive domain of knowing. Teachers can use this approach to describe student work patterns on math test items according to assessments for learning.

Introduction

The design and manufacture of the assessment have shifted in three ways. First, the school is the evaluation maker, implementer, and user. This approach gives principals and teachers the role to improving their ability to create, implement, and use assessments. The activity is intended so that teachers can improve the preparation, process, and learning outcomes in class. Second, teachers follow and apply communication and technology in the school ecosystem. Communication and technology are not merely secondary but are already a primary need for school residents. Teachers and students as school residents who do not have the readiness for communication and technology will be crushed by the times. The era of dynamism is also changing rapidly

* Corresponding author.
E-mail address: rukli@unismuh.ac.id.

https://doi.org/10.34044/j.kjss.2025.46.1.18
2452–3151/© 2025 Kasetsart University.

due to the influence of the expansion of the digital era (Halili, 2019; Shahroom & Hussin, 2018). Third, students get the impact of the assessment conducted by the teacher as much as possible. The teacher and students directly involved will give energy to both. The combination of both will have a greater impact when working collaboratively rather than alone. Otherwise, the teacher is placed on the subject side, and the other side of the student is an object where the learning system uses one direction (Johan & Harlan, 2017). It can result in students being less creative, and their ability to compete is low.

International research data show that Indonesian students struggle to compete globally. Indonesian participation in the TIMSS race for 20 years shows that Indonesia's position has always been below the rank of other participants (Luschei, 2017; Fenanlampir et al., 2019). Likewise, the TIMSS race results show Indonesian students can only work on low-category items while other participating countries can reach high categories or even advanced ones (Luschei, 2017). However, at the same time, during this period, Indonesia made several changes to its education system. For example, the curriculum has changed four times, namely, in 1997, 2004, 2006, and 2013. Even the 2013 curriculum has undergone several revisions (Muth'im, 2014). Other changes were made, namely, teacher training, procurement of free books, improvement of school infrastructure facilities, increase in teacher salaries, and changes in teacher competency and learning mechanisms (Schleicher, 2015). That is, some need to be fundamentally revised, not only aspects outside the school and physical aspects but also aspects of touching the classroom, namely, the interaction of teachers and students.

Students and teachers collaborate in learning by utilizing test item characteristics. The measurement features of the test item are most related to the examinees' ability, namely, the difficulty level of the test item (Magno, 2009; Chae et al., 2019). The characteristics have never been applied to identify student difficulties working on test items in class. Some studies identify the errors students make while solving problems. For example, using a partial credit model polytomous scoring for identifying benchmarks for the polytomous rating scale (Dogan, 2018), and Newman interview procedure for determining error analysis of students working (Reid O'Connor & Norton, 2020).

Determining the Difficulty Level of Test Items (DLTI) can be done using a computer program, for example, the Iteman program, the Bilog-MG program, and other programs. On the other hand, DLTI can be determined using an adjustment approach. The results are not different from the computer program (Stone et al., 2020). Further, research results show that the estimation of the teachers and students to the DLTI is not significantly different from the outcome of a computer program (Rukli et al., 2021). The advantage of the method is that students predict DLTI after working on the item, so the DLTI is more factual and accurate to track students mistakes in working on the test item. Objective and precise are according to the abilities and needs of students when working on these items. This means it can be easier for teachers and schools to conduct assessments in the classroom. Furthermore, the evaluation involves students directly predicting the DLTI after solving the items according to the adaptive learning (Dolenc & Aberšek, 2015; Griff & Matter, 2013; Klenin et al., 2020). TIMSS and PISA are very comprehensive evaluation materials for monitoring and providing information about the state of education in the form of mathematics and science in a country (Fenanlampir et al., 2019).

The results of studies in several countries show that school context and student background have an impact on students TIMSS results in Sweden (Wiberg, 2019). Students low performance on the TIMSS assessment test is related to reasons such as a lack of interest in the TIMSS test and unfamiliarity with the TIMSS test item in Kuwait (Al-Mutawa et al., 2021), and there is an imbalance in the learning achievement of lower and upper-grade students where science subjects in Russian schools are more focused on acquiring and demonstrating knowledge, but to a lesser extent on implementing and implementing practice development scientifically (Pentin et al., 2018). Indonesia is ranked even lower compared to the three countries (Fenanlampir et al., 2019; Haerani et al., 2021). The three studies revealed that students' weaknesses were only limited to cognitive descriptions without providing a detailed presentation of student work. The study aims to describe the involvement of students in getting to know TIMSS test items through students working on the TIMSS test item and then DLTI. In this way, the teacher can carry out further studies on the item map of test items so that students can work on TIMSS test items so that it can be easier to carry out AfL abilities on TIMSS test items.

This paper reports a case study conducted to investigate students' abilities using test takers assessment of the DLTI in grades IV, V, and IV Indonesian elementary schools. The findings of this study identify whether the estimated student DLTI that have been worked on can be used to track in detail the ability of each examinee? If the teacher can track it in detail, they can use it to diagnose students who are having difficulties so that treatment can be improved. That finding informs assessment practice, especially AfL for mathematics teachers. Although this research is limited specifically to the Indonesian context, these findings apply to all teachers, especially in countries that experience limitations for students to understand international standard mathematics items as the development of starting student academic literacy seems to be an ongoing challenge in all countries.

Literature Review

Difficulty Level of Test Items

The level of difficulty is a characteristic of test items apart from other characteristics, namely, discrimination and guessing opportunities. The level of difficulty represents student ability, while differential power and guessing opportunities result from this representation. In theory, the level of difficulty refers to two theories, namely, Classic Test Theory (CTT) and Item Response Theory (IRT) (Subali et al., 2020).

CTT is characterized by a soft level of difficulty. The softness is caused by the underlying assumptions (Quesque & Rossetti, 2020; Yunida & Riyan, 2023). There are many underlying reasons, one of which is dependence on sample characteristics. However, this level of difficulty characteristic has several advantages, for example it is easier to apply in the field.

IRT is characterized by a solid level of difficulty. Contrary to the assumptions of classical test theory, the assumptions of item response theory are independent of the sample (Ackerman et al., 2022) . Once the estimated characteristics of the level of difficulty are invariant the difficulty level remains constant. You can estimate the level of difficulty using a computer program or an adjustment approach. DLTI uses an adjustment approach according to CTT. This approach makes it easier to apply difficulty levels in schools (Rukli et al., 2021). The adjustment approach or using a computer program, for example the Iteman program, is similar.

Students Abilities

The ability of the examinee is a characteristic of the examinee (Halama & Biescad, 2011; Mancheño et al., 2018). These characteristics are latent so there needs to be external stimulation so that they can be measured. These measurements usually use tests. The test results can be in the form of a score according to the CTT concept or in the form of a trait with theta symbol according to the IRT concept.

Scores according to the CTT concept refer to examinees test results after responding to questions in an exam. These measurements estimate the true score with some measurement error. The difference between the real score and the true score results in measurement error. In order for the examinee ability to estimate accurately, the measurement error is as small as possible because it is difficult to find measurements with zero estimation error. The score in the form of the results of the responses to each question until they have been answered in whole or in part is the composite score of the examinee. In simple terms, these results are called the examinee abilities.

In terms of the concept of IRT theory, the examinee ability has the same scale as the level of difficulty (Shaw, 1991; Zanon et al., 2016). Having one scale has several advantages. For example, examinees' abilities and difficulty levels can be compared directly. There are several models expressing this, namely, the logistic model and the Rasch measurement model. The two models are mathematically similar even though their philosophical basis is different.

The ability of examinees in the logistic model consists of three, namely, the one-parameter, two-parameter and three-parameter logistic models (Hambleton & Swaminathan, 1985). This size depends on the characteristics of the questions that are related to ability. On the other hand, the Rasch measurement model only has one thing, namely, linking the level of ability with the level of difficulty of the questions.

Methodology

This study uses a cross-sectional exploratory research type with a descriptive case study approach that describes students' ability to solve math test questions after predicting DLTI. The study of students' ability to answer TIMSS test questions, particularly in mathematics for Grade IV, was conducted through surveys and in-depth observations. The initial survey used a questionnaire

to gather preliminary information on students' perception of TIMSS test items. Subsequent studies include direct trials of TIMSS test items without adaptation. The results of the two studies indicated that students in grades IV, V, and VI experienced difficulties in working on the test questions. Therefore, further research is in the form of a case study to explore this matter by involving students in assessing the level of difficulty of the questions after working on the test questions from seven schools.

Test Item

The test items use the adapted TIMSS test. Adaptation involves measurement, language, and material evaluation experts. There are one hundred questions adapted both in terms of measurement, language and material. Apart from that, adaptations are made to the condition of the question stem, sentence structure or numbers. The number of test item is 40 items. Test item specifications and test blueprint are in Table 1.

Where N1 is whole number, N2 is fractions and decimals, N3 is number sentences with whole numbers, N4 is patterns and relationships, G1 is points, lines, and angles, G2 is two- and three-dimensional shapes, D1 is reading and interpreting, D2 is organizing and representing, and GSM is geometric shapes and measures.

Subjects

The subjects of this study were elementary school students in Indonesia. The number of students involved is 21 people for the case study. Subjects are taken as follows. Research subjects involved grades IV, V, and VI from seven schools. Participants were taken using a purposive sample from seven students in class IV, seven in class V, and seven in class VI. Each school takes one student in grades IV, V, and VI. Selection of students is left to each school. School names and research subjects use pseudonyms. The subjects of this study were elementary school students in Indonesia.

Data Collection

Implementation of DLTI tests and assessments use an adjustment approach according to the CTT concept in the classroom. This activity involved students from the mathematics education study program as well as other parties including teachers and school principals. Each student who has responded to the test questions for three minutes or less immediately estimates the DLTI for 1 minute. DLTI assessment uses a semantic differential scale. The student assessment of the level of difficulty of the questions is on a scale of [0.7]. Student answer data and DLTI estimation results for 40 multiple choice questions are stored on the student desk after they have finished estimating the DLTI for each question. Then the committee collected the data for further analysis.

Data Analysis

Data analysis used a descriptive statistical approach in the form of percentages. Student work in each class is grouped into three sections: high DLTI if students answering true are above 70 percent, moderate DLTI if students answering true are 30 percent – 70 percent, and easy DLTI if students answering true are below 30 percent. Students who are in the middle group of DLTI are not considered. So, the descriptions of the test questions were carried out by students of grades IV, V, and VI taking into account the level of DLTI. Analysis of student job descriptions is to find patterns in working on test questions by paying attention to the upper and lower groups.

Table 1 Distribution of Test Item

Cognitive Domain	Content Domain								∑
	Number				GSM		Data Display		
	N1	N2	N3	N4	G1	G2	D1	D2	
Knowing	1, 3, 4, 6, 7, 9, 10, 11, 12, 15, 33	13;24	8	19, 26, 28	36	18, 30, 39			21
Applying	5	34	14, 22, 23	2, 16, 17, 20, 21, 38	27	40		35	14
Reasoning			25, 32	29	37		31		5
∑	12	3	6	10	3	4	1	1	40

DLTI predictions use the Semantic Differential Scale (SDS) with a scale of 0–7, where zero is the easiest item and seven is the most difficult item. Completion of the SDS pays attention to two guidelines, namely, the procedure for filling in the SDS and the SDS rating scale qualitatively. The SDS filling process consists of three points, namely, reviewing the test questions that have been done, the range of assessment scores, and the DLTI assessment using three decimal places. The qualitative SDS filling scale uses a reference that the movement of scores starting from zero means the easiest questions, while a score of seven means the most difficult questions.

Results and Discussion

Description for High Category

The Venn diagram contains the high category of DLTI which consists of three sets of students in grades IV, V and VI. The intersection of the three class sets consists of eight test items. Figure 1 shows the number and slice of high DLTI for each class. All difficult items in grade IV are also difficult in grade V except for test item 29. Test item 29 has the content domain of numbers, topic areas of patterns and relationships, and the cognitive domain of reasoning.

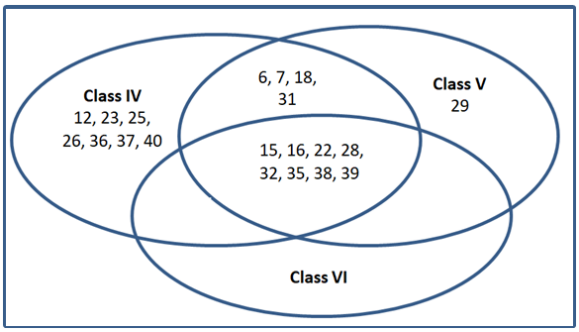

Figure 1 High category of DLTI

The answers of DLTI students in the high category have the following characteristics. The characteristics of test item 28 measure the domain of number content, topic areas of patterns and relationships, and cognitive domain of knowing. The item is written, Ati made 57 bags of cakes in March, 62 bags in April, and 59 bags in May. The best assessment of the number of bags Ati needs for three months is: The item options are (1) 50 + 50 + 50; (2) 55 + 55 + 55; (3) 60 + 60 + 60; (4) 65 + 65 + 65. The answer key is option 3. 29 percent of Grade IV students answered correctly. The prediction DLTI is 5.303. 14 percent of Grade V students answered correctly. The prediction DLTI is 4.07. 0 percent of Grade VI students answered right. The prediction DLTI is 2.591. Grade IV students chose options 1 and 2. Grade V chose option 1, while grade VI chose options 1 and 2. None chose the answer key. The data show that the higher the grade level, the lower the student ability, but students predict DLTI decreases. Students say the test item is easy but cannot answer correctly.

Students experienced deficiencies in the topic areas of patterns and relationships based on these two test items. By paying attention to the DLTI of the test item, the test items require the stage of knowing while students have a weak understanding of the concept. So, the fifth-grade and even sixth-grade students have forgotten the concept of the test item, so the answer is wrong, but the assesses of the DLTI of the test item is very low. However, in contrast to grade IV, they still remember the stages of the test item. Another cause is that students in grades V and VI lack understanding, so they are limited to memory when in grade IV. The failure of students to do additional operations is due to an inadequate understanding of basic operational steps (Confrey, 2011). It means that students have difficulty understanding the purpose of the test item, understanding the concept of place value, translating test items into mathematical sentences, having difficulty doing addition calculations and having low self-confidence (Sidik et al., 2021). Therefore, fourth-grade teachers need to deepen the procedural concept of the material so that students not only remember but also understand the items by paying attention to the DLTI of test items.

Characteristics of test item 32 measures students' ability in the content domain of numbers, topic areas of number sentences with whole numbers, and cognitive domain of reasoning. The test item is written, Ambo Dawi goes to Ampalang garden and returns home at 9 in the morning. The trip to the garden takes 1 hour 30 minutes. What time does Ambo Dawi go to the garden? The key test item is 6 a.m. Grade IV students answered 28 percent correctly with the prediction of the DLTI of 4.097. Grade V students answered 14 percent correctly with the prediction of the DLTI by 4.173. Grade students VI answered 0 percent correctly with the prediction DLTI equal to 2.591. The data are similar to the previous two items: the higher the grade level, the more students get the test items wrong, but on the contrary, the DLTI assessment results are lower. The test item measures

reasoning to require students to reason, where most students answer at 6 30 minutes in the morning. Students experience deficiencies in number sentences with integers in the topic domain for reasoning. Students have not been able to think logically to make connections between empirical facts and the problems at hand, so they are not able to conclude, students have not been able to carry out the thought process to make arguments so that new statements are based on facts, it is necessary to develop a learning model to improve the mathematical reasoning of elementary school students, especially Indonesian students (Sulianto et al., 2020; Syafitri et al., 2020). Teachers need innovative learning models so that students reasoning develops optimally. Likewise, they gave feedback by improving their understanding of the reasoning steps while still paying attention to DLTI predictions because upper-grade students show high DLTI predictions.

The data show that grade IV students predict higher DLTI than other classes. Students experience the limitations of rectangles when the concept is broad in the context of everyday life. It can happen because students in grades IV are less thorough in reading the subject matter about the number of rectangles. Therefore, teachers need to provide feedback when learning in class, especially related to the topic of two- and three-dimensional shapes. The weakness of students working on geometry problems is not only students of grade VI and below. However, grade VII students experienced misconceptions, lack of background knowledge, reasoning and basic operating errors on the topic (Özerem, 2012). The same is true for adults who do not always have direct access to fractional quantities on a number line (DeWolf & Vosniadou, 2015).

Therefore, teachers need to emphasize identifying, classifying, and comparing common geometric shapes (e.g., classifying or comparing based on shape, size, or property). In addition, the fourth-grade teacher provided feedback on remembering, describing, and using the basic properties of plane figures, including line symmetry and rotation.

Description for Low Category

The Venn diagram in Figure 2 contains the low of the DLTI category which consists of three sets of students in grades IV, V and VI. The intersection of the three class sets consists of ten test items. The ten test items measure the same, namely, the content of the domain of the number and the cognitive of knowing domain but only differ in topic areas. The most common topic areas are whole numbers. It shows that the easiest test item for students is the domain content number test items in the topic areas of whole numbers in the cognitive domain of knowing. The ten test items do not measure the reasoning domain.

The several DLTI low categories have as follows. The characteristics of test item 8 measure students' ability in the content domain of numbers, topic areas of number sentences with whole numbers, and cognitive domain of knowing. The test item is written, Ulhaq will use a calculator to calculate 213 + 13. He entered 203 + 13, but it is wrong. What should Ulhaq do to fix it? The key test item is added 10. Grade IV students answered 86 percent correctly, with the predicted DLTI being 1.584. Grade V students answered 86 percent correctly, but the prediction DLTI was higher, i.e., 2.416. Grade VI students answered 100 percent correctly with the prediction of DLTI of easier items at 1.303. Test item 8 is similar to the case in test item 4, where the second position of the item requires a low-thinking stage which is knowing the content domain of the number.

The characteristics of test item 9 measure students' ability in the content domain of numbers, topic areas of the whole number, and cognitive domain of knowing. The test item is written, Which number is the smallest of the following numbers? The choices are 2753, 2573, 2735, and 2537. The key test items are 2537. The result shows that: (1) grade IV students answer 100 percent correctly, with the prediction of the DLTI being 1.892;(2) grade V students answered 86 percent correctly, with a prediction of the DLTI being 1.086, and (3) grade VI students have the same answer as for Grade IV, but predictions of DLTI are smaller at .766. The data show the ability to work on items with a prediction of the DLTI in the same direction even though grade V misses the prediction. However, these items are known by students in all grades.

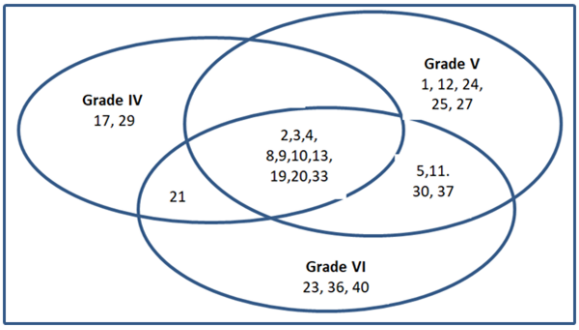

Figure 2 Easy category of DLTI

The student ability to work on test items and predictions of the DLTI by students on test items is more consistent with the theory that if the DLTI is high, then students' ability is low. Otherwise, if the DLTI is low, students' ability increases. Students experience many mistakes in the initial stages, so students do not understand the material, especially in grade IV, because new students get exam material.

The description of the student's ability in the three grades with the DLTI in the high, and low categories provides the following information. First, the information relates the characteristics of test items to the order of items in tests, the content domain, the topic areas, and the cognitive domain with the description of students' abilities to be complete and faster. Second, the teacher can use the information to conduct a detailed AfL related to the teaching material. Third, student involvement in assessing the DLTI increases their knowledge and understanding of the material. Fourth, schools get faster, more complete, and more accurate information about the map of students' abilities. Fifth, a description of the topic areas to the cognitive domain helps students and teachers understand higher-order thinking skills test items.

The results of the five-point study, tracking students' ability to solve test items with DLTI, can show test items that are too difficult and too easy for students. If the DLTI increases, the student abilities decrease, and so does their commitment to the goals of the performance approach, while their support to avoid work increases (Kumar & Jagacinski, 2011). Some of the information provides benefits and support in the learning process of mathematics in the classroom. However, the description of students' ability to use the assessed DLTI is still limited where the study is based on CTT. Therefore, these descriptions require deeper exploration before they are applied in other schools for a more comprehensive application using the item response theory approach. In addition, the description of students' abilities is limited to a description using a quantitative approach, so a qualitative study needs to be carried out further. Likewise, the test items only use the multiple-choice type of test. The study of the essay test is a suggestion for further research. However, multiple-choice test items also require language accuracies, such as the number of propositions and syntactic structure. Fundamentally, the presence of difficult words contributes to the prediction of DLTI (Brizuela & Montero-Rojas, 2014). However, language difficulties have been assessed by manipulating the size and distribution of gaps based on predictions of absolute and relative disparity difficulties (Lee et al., 2020).

Description of the assessed DLTI and Test Item Answers

Students assess that the DLTI varies according to the SDS. The DLTI scores of the four students are in Figure 3 from test item number 11 to 20. There are four examples of student work where the first column contains the Item Number (IN) while the second column is the result of the DLTI estimate for each item.

The results of the DLTI assessment of the four students were not the same from test items 11 to 20. Of the 10 test items described, only test items 13 and 14 were. Test items 13 and 14 in Figure 4 result from MN answer. Test item 13 measures the content of the number domain with topic areas of fractions and decimals and the cognitive domain of knowing. Test item 14 measures the content domain of the number with topic areas number sentences with whole numbers and the cognitive domain of applying.

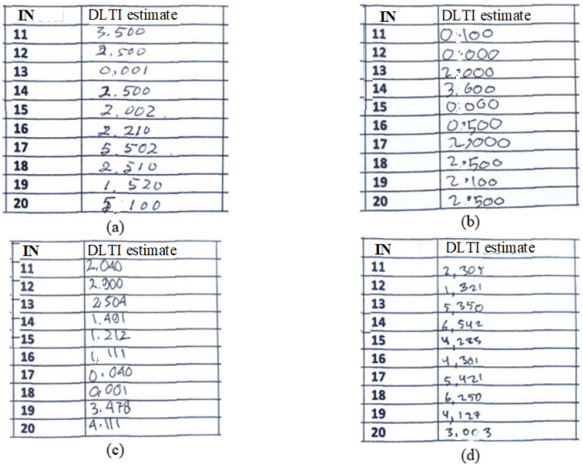

Figure 3 The DLTI assessment from (a) MN, (b) NN, (c) RK, and (d) SK

13. **0.8 means**
    a. 8/10
    b. 8
    c. 10
    d. 1/4

14. **Habibi uses 4 tomatoes to make half a liter of tomato sauce. How much tomato sauce can be made from 16 tomatoes?**
    a. **One liter**
    b. **One and a half liters of sauce**
    c. **Two and a half liters of sauce**
    d. **Three liters of sauce**
    e. **Two liters of sauce**

Figure 4 MN answer choice on test items 13 and 14

There are two test items, namely, 13 and 14 test items. Test item 13 is related to changing decimal numbers to fractions while test item 14 is related to fractions in the form of word problems. Test item 13 in Figure 4 is written, 0.8 means, and the key to item 13 is 8/10. Test item 14 is written, Habibi uses four tomatoes to make half a litre of tomato sauce. How many litres of sauce can he make from 16 tomatoes? The key to item 14 is two litres of milk. Table 2 shows the two test items DLTI assessment results and the answer choices of the four students.

Items 13 and 14 measure the same content, namely, numbers, but are different in areas and cognitive domain topics. There are two fourth-grade students, namely, RK and NN and two fifth-grade students, namely, SK and MN. SK of grade V experienced errors in both test items by saying the two test items were difficult, while NN grade IV answered both test items correctly by saying that both test items had a moderate level of difficulty. The two students were consistent in their answers with the DLTI, but the abnormality was that the fifth-grade students considered both test items difficult while the fourth-grade students considered them moderate.

Furthermore, RK of grade IV correctly answered test item 13 with a high DLTI but incorrectly answered test item 14 with a low DLTI. On the other hand, MN did not answer test item 13 with a low DLTI but correctly answered test item 14 with a high DLTI. The difference was that the two did not assess the same difficult test items. MN said that test item 14 was difficult in the cognitive domain of application. In comparison, RK said that test item 13 was difficult in the cognitive domain of knowing. The four students need to be considered by the teacher to discuss further the pattern of students' ability to work on the test item. The results showed that the items with high and low DLTI showed that the class IV, V, and VI students had the answers to the test items with the DLTI that did not match. In test theory, the DLTI is related to students' abilities. If the DLTI is high, the student ability to answer truly is low, and vice versa (Foster, 2020; Kohli et al., 2015; Martins et al., 2020). Therefore, the discussion is focused on high DLTI and low DLTI by taking certain test items according to content accommodation and cognitive domains.

The four students worked on test items 13 and 14 with variations of answers with the DLTI not adaptive. SK, who is in grade V, has the highest estimation of DLTI, and the answers to both test items are wrong. It means that she is working on a complicated matter. SK considered test items 13 and 14 difficult, both about the cognitive domain of knowing and applying. SK is a forgetful type of memory error, forgetting how to do the fourth-grade test item one year ago. Forgetfulness can occur due to conceptual errors because the teacher does not focus on learning retention, identifying errors, and increasing student scores (Ancheta & Subia, 2020). SK needs additional practice, so they do not forget. Teachers avoid blaming students (Hong & Hobbs, 2021), or when working on test items keeping away from interference from friends can cause forgetting (Kreitewolf et al., 2019). So, the teachers need to emphasize or provide a link to previous material information and emphasize certain concepts so as not to forget and avoid students from being disturbed, especially the presence of their friends during exams.

RK assessed DLTI 13 and 14 in the medium category, where the DLTI scores were close to the average on the SDS scale. RK encountered an error in topic areas fractions and decimals and number sentences with whole numbers. However, the DLTI's estimation results are not reasonable because the test items considered easy are wrong, while the test items considered difficult are correct. RK thought the test item was difficult but answered correctly, while the test item was easy but answered incorrectly. The risk of memory is high but unable to apply concepts or rules. RK is the type of lack of concentration and carelessness. This pattern of working on test items requires practice to learn from mistakes (Pan et al., 2020) and avoid interference from close people (Kreitewolf et al., 2019). Teachers need to give other test items to take lessons from previous mistakes.

Table 2 Assesses the two DLTI from four students

Name /Grade	DLTI/Cognitive Domain		Answer	
	Item 13/ Knowing	Item 14/Applying	Item 13	Item14
SK/ V	5.350	6.542	False	False
RK/ IV	2.504	1.491	True	False
MN/V	0.001	2.500	False	True
NN/IV	2.000	3.600	True	True

MN, in class V, actually worked on test 13 but forgot to choose the answer, so he ended up wrong. MN considered the test item very easy with a DLTI assessment of only .001. MN did the 14th test item, thinking that the test item was moderate, but the answer was correct. MN's ability to work on the two test items was inconsistent, where the DLTI assessment was in a low category, but he was unable to work on the test item. MN is similar to RK, but MN is wrong in doing the knowing test item, which is considered easy but can correctly answer the applying test item he feels is difficult. MN is the type of student who is less careful or careless in doing the test item. It can happen due to limited ability, language, and confidence in the content of the test item (Clements, 1982) or overconfidence (San Pedro et al., 2014). The teacher needs to provide information to re-check the answer or provide information that there is still time left. There is no need to rush.

On the other hand, the DLTI assessment was moderate, but he could do the problem correctly. However, after tracing the answer sheets to the test items, MN did not choose one of the available options. There are two possibilities: MN did not know how to do test item 13, or he could do it but forgot to choose the answer. Therefore, the teacher needs to warn or pay attention to MN not to be careless when doing the test items. MN's behavior is categorized as mathematics anxiety, and it can be due to his nature or new things, for example, someone else or online exams. MN's stress about doing the test items could affect his exam results (Jiang et al., 2021; Vanbinst et al., 2020). Carelessness in doing the test items can cause inaccurate ability documentation.

NN considered the two test items to be in the medium category, where the answers to both test items were correct by assuming that test item 14 was more difficult than test item 13. The DLTI estimation results and accurate answers to both test items showed that NN had no problems with both test items. NN does not need to take part in remedial but can participate in enrichment activities or help friends with these test items. Teachers need to provide a different approach between RK and NN. RK answered inconsistent test items, the low DLTI but answered wrong test items. Conversely, high DLTI answered the test item correctly. It can happen when RK answers by guessing the answer key. Multiple-choice test items are weak because the examinee can guess the answer (Koediger & Marsh, 2005; Slepkov et al., 2021). The teacher needs to provide unique guidance in remedial test items for these two test items.

On the other hand, NN consistently does the test items, and everything is correct. Therefore, teachers need to provide NN while RK is doing remedial learning. NN can be a learning model for teachers to motivate other students to learn. Modelling is essential for other students to get up, namely, consistently answering and correcting answers. Based on the four data on the student DLTI assessment results, the teacher can provide feedback by paying attention to each test item characteristics. Both test items measure the same content domain, namely, numbers, but test item 13 measures students' abilities in topic areas of fractions and decimals. On the other hand, test item 14 measures students' abilities in topic areas of number sentences with whole numbers. It means that teachers need to deepen the material in the form of remedial for the three students in the content domain, namely, the number on topic areas of fractions and decimals and number sentences with whole numbers.

Item 13 measures the cognitive domain of knowing, while test item 14 measures the cognitive domain of applying. MN had an error in test item 13, where the test item demanded the cognitive domain of knowledge, but he got the correct answer to test item 14. RK answered correctly on test item 13 but experienced an error when test item 14 required the cognitive domain application stage. SK experiences an error in a test item demanding the cognitive domain thinking stage of knowledge and application. NN answered correctly all the test items that required the cognitive domain thinking stage of knowledge and application.

Judging from the characteristics of test item 13, only MN considered the test item easy, namely, a DLTI of .001, even though MN did not answer the test item. Only RK considered the 14th test item easy, namely, the DLTI of 1.491 but had an error doing it. So, MN and RK considered that they could work on the test items by paying attention to the DLTI assessments in the low category, but in reality, they failed. Although MN's approach was wrong, it was different from RK. MN had an error on the knowledge test item, while RK had an error on applying. It provides information that the description shows that the teacher offers further emphasis assistance in both psychology and material to the two students. Psychologically, the teacher gave a normal/relaxed atmosphere, especially to MN that exams were a common thing in learning so that MN could avoid being careless and overworked in working on the remembering type test items. Judging from the learning material, the teacher must repeat the material. Several concepts, procedures, or principles have

been forgotten to understand or apply to work on the applying test items, especially for RK students, where it is also necessary to deepen the material. SK rated DLTI the highest on both test items and answered both incorrectly. Deepening the wrong material can be an experience for teachers to correct learning methods and for students to prepare themselves to work on these types of test items. Based on this, DLTI can help teachers to carry out learning assessments in class, especially for students who have difficulty working on TIMSS items. Thus, Indonesian students are expected to increase their ranking in international-level mathematics competitions.

Conclusion and Recommendation

The description of students' abilities with DLTI estimates can indicate the ability of each student in grades IV, V, and IV. The high category test questions indicate that the DLTI predictions do not match students' abilities in working on test items because students have difficulty doing the questions correctly. But for easy category test items, the results are consistent. The involvement of students in predicting DLTI after completing test items has provided a pattern of students' ability to complete test items. The study can be a source for teachers to provide feedback according to AfL.

Conflict of Interest

The author declares that there is no conflict of interest.

Acknowledgments

This research was funded by an Internal Fiscal grant from the Muhammadiyah University of Makassar through the Institute for Research Development and Community Service in 2019. Therefore, the author would like to thank the Chancellor and Chair of the Institute for Research, Development and Community Service, University of Muhammadiyah Makassar for the financial support.

References

Ackerman, T., Ma, Y., Ma, M., Pacico, J. C., Wang, Y., Xu, G., Ye, T., Zhang, J., & Zheng, M. (2022). Item response theory. In International Encyclopedia of Education: Fourth Edition. https://doi.org/10.1016/B978-0-12-818630-5.10010-7

Al-Mutawa, F., Al-Rasheedi, G., & Al-Maie, D. (2021). Kuwaiti students' achievements in Mathematics: Findings from the TIMSS assessments: Reality and reasons. SAGE Open, 11(3), 21582440211031903. https://doi.org/10.1177/21582440211031903

Ancheta, C. M. D., & Subia, G. S. (2020). Error analysis of engineering students' misconceptions in Algebra. International Journal of Engineering Trends and Technology, 68(12), 66-71. https://doi.org/10.14445/22315381/IJETTV68I12P212

Chae, Y. M., Park, S. G., & Park, I. (2019). The relationship between classical item characteristics and item response time on computer-based testing. Korean Journal of Medical Education, 31(1), 1. https://doi.org/10.3946/kjme.2019.113

Clements, M. A. (1982). Careless errors made by sixth-grade children on written mathematical tasks. Journal for Research in Mathematics Education, 13(2), 136. https://doi.org/10.2307/748360

Confrey, J. (2011). Better measurement of higher cognitive processes through learning trajectories and diagnostic assessments in mathematics: The challenge in adolescence. In The adolescent brain: Learning, reasoning, and decision making, 155–182. https://doi.org/10.1037/13493-006

DeWolf, M., & Vosniadou, S. (2015). The representation of fraction magnitudes and the whole number bias reconsidered. Learning and Instruction, 37, 39–49. https://doi.org/10.1016/j.learninstruc.2014.07.002

Dogan, E. (2018). An application of the partial credit IRT model in identifying benchmarks for polytomous rating scale instruments. Practical Assessment, Research and Evaluation, 23(7), 1–10. https://doi.org/https://doi.org/10.7275/1cf3-aq56

Dolenc, K., & Aberšek, B. (2015). TECH8 intelligent and adaptive e-learning system: Integration into Technology and Science classrooms in lower secondary schools. Computers and Education, 82, 354–365. https://doi.org/10.1016/j.compedu.2014.12.010

Fenanlampir, A., Batlolona, J. R., & Imelda, I. (2019). The struggle of Indonesian students in the context of TIMSS and Pisa has not ended. International Journal of Civil Engineering and Technology, 10(2), 393–406.

Foster, R. C. (2020). A generalized framework for classical test theory. Journal of Mathematical Psychology, 96, 102330. https://doi.org/10.1016/j.jmp.2020.102330

Griff, E. R., & Matter, S. F. (2013). Evaluation of an adaptive online learning system. British Journal of Educational Technology, 44(1), 170–176. https://doi.org/10.1111/j.1467-8535.2012.01300.x

Haerani, A., Novianingsih, K., & Turmudi, T. (2021). Analysis of students' errors in solving word problems viewed from mathematical resilience. JTAM (Jurnal Teori Dan Aplikasi Matematika), 5(1), 246–253. https://doi.org/10.31764/jtam.v5i1.3928

Halama, P., & Biescad, M. (2011). Measurement of psychotherapy change: Comparison of classical test score and IRT based score. Ceskoslovenska Psychologie: Casopis Pro Psychologickou Teorii a Praxi, 55(5), 400–411.

Halili, S. H. (2019). Technological advancements in education 4.0. The Online Journal of Distance Education and E-Learning, 7(1), 63–69. https://www.tojsat.net/journals/ tojdel/articles/v07i01/v07i01-08.pdf

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Springer. https://doi.org/10.1007/978-94-017-1988-9

Hong, M. R., & Hobbs, S. D. (2021). Simulated memory error and blame attribution in cases of child sexual abuse. Psychology, Crime and Law, 27(5), 494–516. https://doi.org/10.1080/1068316X.2020.1837128

Jiang, R., Liu, R. de, Star, J., Zhen, R., Wang, J., Hong, W., Jiang, S., Sun, Y., & Fu, X. (2021). How mathematics anxiety affects students' inflexible perseverance in mathematics problem-solving: Examining the mediating role of cognitive reflection. British Journal of Educational Psychology, 91(1), 237–260. https://doi.org/ 10.1111/bjep.12364

Klenin, A. I., Donskov, A. V., Spasskaya, D. D., & Khussein, A. M. A. (2020). Digital technologies in teacher training: New experience. ITM Web of Conferences, 35, 06002. https://doi.org/10.1051/itmconf/20203506002

Koediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. In Journal of Experimental Psychology: Learning Memory and Cognition, 31(5), 1555.

Kohli, N., Koran, J., & Henn, L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models, Educational and Psychological Measurement, 75(3), 389–405. https://doi.org/10.1177/ 0013164414559071

Kreitewolf, J., Wöstmann, M., Tune, S., Plöchl, M., & Obleser, J. (2019). Working-memory disruption by task-irrelevant talkers depends on degree of talker familiarity. Attention, Perception, and Psychophysics, 81(4), 1108–1118. https://doi.org/10.3758/s13414-019-01727-2

Kumar, S., & Jagacinski, C. M. (2011). Confronting task difficulty in ego involvement: Change in performance goals. Journal of Educational Psychology, 103(3), 664. https://doi.org/10.1037/a0023336

Luschei, T. F. (2017). 20 Years of TIMSS: Lessons for Indonesia. Indonesian Research Journal in Education (IRJE), 1(1), 6–17. https://doi.org/10.22437/irje.v1i1.4333

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory ssing derived test data. The International Journal of Educational and Psychological Assessment, 1(1), 1–11.

Mancheño, J. J., Cupani, M., Gutiérrez-López, M., Delgado, E., Moraleda, E., Cáceres-Pachón, P., Fernández-Calderón, F., & Lozano Rojas, Ó. M. (2018). Classical test theory and item response theory produced differences on estimation of reliable clinical index in World Health Organization Disability Assessment Schedule 2.0. Journal of Clinical Epidemiology, 103, 51–59. https://doi.org/10.1016/j.jclinepi.2018.07.002

Martins, P. S. R., Barbosa-Pereira, D., Valgas-Costa, M., & Mansur-Alves, M. (2020). Item analysis of the Child Neuropsychological Assessment Test (TENI): Classical test theory and item response theory. Applied Neuropsychology: Child, 1–11. https://doi.org/10.1080/21622965.2020.1846128

Muth'im, A. (2014). Understanding and responding to the change of curriculum in the context of indonesian education. American Journal of Educational Research, 2(11), 1094–1099. https://doi.org/10.12691/education-2-11-15

Özerem, A. (2012). Misconceptions in geometry and suggested solutions for seventh grade students. Procedia - Social and Behavioral Sciences, 55, 720–729. https://doi.org/10.1016/j.sbspro.2012.09.557Pan, S. C., Sana, F., Samani, J., Cooke, J., & Kim, J. A. (2020). Learning from errors: students' and instructors' practices, attitudes, and beliefs. Memory, 28(9), 1105–1122. https://doi.org/ 10.1080/09658211.2020.1815790

Pan, S. C., Sana, F., Samani, J., Cooke, J., & Kim, J. A. (2020). Learning from errors: Students' and instructors' practices, attitudes, and beliefs. Memory, 28(9), 1105–1122. https://doi.org/10.1080/09658211.2020.1815790

Pentin, A., Kovaleva, G., Davidova, E., & Smirnova, E. (2018). Science education in Russia according to the results of the TIMSS and PISA international studies. Voprosy Obrazovaniya/Educational Studies Moscow, (1), 79–109. https://doi.org/10.17323/1814-9545-2018-1-79-109

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. Perspectives on Psychological Science, 15(2), 384–396. https://doi.org/10.1177/1745691619896607

Reid O'Connor, B., & Norton, S. (2020). Supporting indigenous primary students' success in problem-solving: Learning from Newman interviews. Mathematics Education Research Journal, 34(2), 293–316. https://doi.org/10.1007/s13394-020-00345-8

Johan, R. & Harlan, J. (2017). Education nowadays. International Journal of Educational Science and Research, 4(5), 51–56.

Rukli, R., Ma'rup, M., Bahar, E. E., & Ramdani, R. (2021). The estimation of test item difficulty using focus group discussion approach on the semantic differential scale. Kasetsart Journal of Social Sciences, 42(3), 599–606. https://doi.org/10.34044/j.kjss.2021.42.3.22

San Pedro, M. O. Z., Baker, R. S. J. D., & Rodrigo, M. M. T. (2014). Carelessness and affect in an intelligent tutoring system for mathematics. International Journal of Artificial Intelligence in Education, 24(2), 189–210. https://doi.org/10.1007/s40593-014-0015-y

Schleicher, A. (2015). Educational in Indonesia: Rising to the challenge. https://repositori.kemdikbud.go.id/8414/1/ACDP017%20-%20Final-Report-Education-Indonesia-Rising-Challenge.pdf

Shahroom, A. A., & Hussin, N. (2018). Industrial revolution 4.0 and education. International Journal of Academic Research in Business and Social Sciences, 8(9), 314–319. https://doi.org/10.6007/IJARBSS/v8-i9/4593

Shaw, F. (1991). Descriptive IRT vs. Prescriptive Rasch. Rasch Measurement Transactions. https://www.rasch.org/rmt/rmt51f.htm

Sidik, G. S., Suryadi, Di., & Turmudi. (2021). Learning obstacle of addition operation whole number in elementary schools. Journal of Physics: Conference Series, 1842(1), 012070. https://doi.org/10.1088/1742-6596/1842/ 1/012070

Slepkov, A. D., Van Bussel, M. L., Fitze, K. M., & Burr, W. S. (2021). A baseline for multiple-choice testing in the university classroom. SAGE Open, 11(2), 21582440211016838. https://doi.org/10.1177/21582440211016838

Stone, J. C., Glass, K., Munn, Z., Tugwell, P., & Doi, S. A. R. (2020). Comparison of bias adjustment methods in meta-analysis suggests that quality effects modeling may have less limitations than other approaches. Journal of Clinical Epidemiology, 117, 36–45. https://doi.org/10.1016/j.jclinepi.2019.09.010

Subali, B., Kumaidi, & Aminah, N. S. (2020). The comparison of item test characteristics viewed from classic and modern test theory. International Journal of Instruction, 14(1), 647–660. https://doi.org/10.29333/IJI.2021.14139A

Sulianto, J., Sunardi, S., Anitah, S., & Gunarhadi, G. (2020). Classification of student reasoning skills in solving mathematics problems in Elementary School. Jurnal Pendidikan Indonesia (JPI), 9(1), 95–105. https://doi.org/10.23887/jpi-undiksha.v9i1.23103

Syafitri, R., Putra, Z. H., & Noviana, E. (2020). Fifth grade students' logical thinking in mathematics. Journal of Teaching and Learning in Elementary Education (JTLEE), 3(2), 157–167. https://doi.org/10.33578/jtlee.v3i2.7840

Vanbinst, K., Bellon, E., & Dowker, A. (2020). Mathematics anxiety: An intergenerational approach. Frontiers in Psychology, 11, 1648. https://doi.org/10.3389/fpsyg.2020. 01648

Wiberg, M. (2019). The relationship between TIMSS mathematics achievements, grades, and national test scores. Education Inquiry, 10(4), 328–343. https://doi.org/10.1080/ 20004508.2019.1579626

Yunida, H., & Riyan, A. (2023). Bloom's taxonomy approach to cognitive space using classic test theory and modern theory. East Asian Journal of Multidisciplinary Research, 2(1), 95–108. https://doi.org/10.55927/eajmr.v2i1.2331

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. Psicologia: Reflexao e Critica, 29, 18. https://doi.org/10.1186/s41155-016-0040-x