Jurnal Pijar MIPA
https://jurnalfkip.unram.ac.id/index.php/JPM

ISSN 1907-1744 (Print)
ISSN 2460-1500 (Online)

Item Analysis of the Minimum Competency Assessment in Science Subjects at the Junior High School Level

Durriyatus Salsabila

Department of Science Education, Universitas Negeri Surabaya, Surabaya, Indonesia
e-mail: durriyatus.22057@mhs.unesa.ac.id

Received: December 25, 2025. Accepted: February 6, 2026. Published: February 12, 2026

Abstract: Minimum Competency Assessment (MCA), known in Indonesia as Asesmen Kompetensi Minimum (AKM), is a national assessment of basic mastery that students must complete. Teachers must be able to compile MCA questions that accurately measure student knowledge. Well-written questions are crucial for creating high-quality assessments. This study analyzes the quality of MCA items on the topic of linear motion in science education at a junior high school in Surabaya. The study aims to evaluate multiple-choice questions using key psychometric parameters, such as validity, reliability, difficulty index, discrimination index, and distractor efficiency. The questions were created by the author and administered to 15 ninth-grade students on 8 May 2024. The analysis used descriptive quantitative methods, and data processing was performed in Microsoft Excel. The results showed that only 4 of 10 items were valid, and the overall reliability coefficient was low (KR-20 = 0.213). Based on the difficulty index, six items were categorized as moderate, three were difficult, and one was easy. The discrimination index analysis revealed that five items had good to excellent discrimination power, whereas others showed poor discrimination and required revision. Distractor efficiency analysis indicated that several distractors were non-functional, particularly in items with low discrimination indices. These findings imply that although some MCA items are suitable for use, many require improvement to ensure accurate, reliable, and diagnostic measurement of student competencies. The study highlights the importance of item analysis as a basis for improving the design of MCA-based assessments in science education.

Keywords: Item Analysis; Linear Motion; Minimum Competency Assessment (MCA); Science.

Introduction

Freedom of learning in the 21st century is often associated with the development of technology and varied self-skills [1]. This is one of the bases for the independent learning policy initiated by the Minister of Education, Culture, Research, and Technology, Nadiem Makarim. The policy contains 4 points, one of which is to replace the National Examination with the Minimum Competency Assessment (MCA) and character surveys [2]. Minimum Competency Assessment (MCA), known in Indonesia as Asesmen Kompetensi Minimum (AKM), is a national assessment of basic mastery that students must follow to improve their abilities and be able to participate in society actively [3]. MCA is one of the national assessment instruments for students based on two basic abilities, namely reading literacy and numeracy [1]. Therefore, the preparation of MCA items should not be arbitrary.

Teachers must be able to develop MCA questions that can accurately measure student knowledge. The items must be well-written to produce high-quality questions for students. Therefore, it is necessary to analyze the question items, which is the stage of evaluating the quality of the questions to determine whether the questions are valid, reliable, and suitable for use in the assessment process [4]. Item analysis will make it easier for teachers to analyze the quality of questions and revise questions if necessary, before the questions are used as an assessment tool [5]. Item analysis includes testing the validity, reliability, difficulty index, discrimination index, and distractor efficiency.

The item validity is a test used to determine the extent to which a test can measure what it is intended to measure [6]. A test can have high validity when its measuring function is precise and in accordance with the purpose of the measurement being carried out [7]. A questionnaire is said to be valid if the calculated r value is greater than the r-table (r-calculated > r-table) [8]. After a validity test, a question is considered consistent if, when retested, the results remain the same as in the first trial. The word reliability means the extent to which the results of a measurement can be trusted. A measurement can be trusted when the measurement is carried out repeatedly on the same subject, and the results obtained are relatively the same or unchanged [9]. The reliability of an assessment tool is the fixity or consistency of a measurement in assessing what you want to assess [10]

The difficulty index of a question can be interpreted as the probability that students will answer it correctly at the given skill level. This can be used as a benchmark to prove whether the question is classified as easy or difficult [11]. A good question is a question that has a medium difficulty index, because the question requires high intelligence and effective pseudo-choice [12]. Therefore, if students participating in the trial cannot answer a question at all, it can be said that the question is bad. Likewise, vice versa for the easy question category [5].

Item quality parameters are also determined by the discrimination index. The discrimination index measures a

How to Cite:
D. Salsabila, "Item Analysis of the Minimum Competency Assessment in Science Subjects at the Junior High School Level", *J. Pijar.MIPA*, vol. 21, no. 1, pp. 117–123, Feb. 2026. https://doi.org/10.29303/jpm.v21i1.11243

question's ability to distinguish between highly and less intelligent test takers [13]. Multiple-choice questions include answer choices that serve as distractors in addition to the answer key. A distractor functions well when the distractor is chosen by more than 5% of test takers [10]. Multiple-choice questions often include many distractor answers [14]. The effectiveness of the distractor is evident from how well it tricks selection participants by shaping the pattern of answer choices [5].

Most previous studies have primarily focused on teachers' perceptions, readiness, and students' achievement profiles in literacy and numeracy [3][1]. Other studies discuss MCA within the broader framework of educational reform and independent learning policies, without examining the quality of assessment items themselves [2]. Several studies have emphasized the importance of item analysis in improving assessment quality, including validity, reliability, difficulty index, discrimination index, and distractor efficiency [10][15][5]. However, these studies analyze MCA instruments and do not specifically address teacher-developed MCA science items at the junior high school level. In addition, research focusing on item-level psychometric evaluation in science subjects remains limited compared to studies in literacy and numeracy domains [11].

Therefore, this study addresses the existing research gap by conducting a comprehensive item analysis of MCA science questions on linear motion using established psychometric parameters. The novelty of this research lies in its focus on item-level evaluation of teacher-developed MCA science assessments, providing empirical evidence on their validity, reliability, difficulty index, discrimination index, and distractor efficiency. This analysis is necessary to support teachers in improving assessment literacy and ensuring that MCA instruments function as valid and reliable tools for measuring student competencies and informing instructional decision-making.

Research Methods

The research design used was a quantitative descriptive method (related to its statistical characteristics). The analyzed questions are MCA questions for science subjects on the linear motion topic at the junior high school level. The questions were created by the author, and the sampling technique was purposive, involving 15 ninth-grade students from a junior high school in Surabaya who had completed instruction on linear motion. Data were collected through a test administered on 8 May 2024. Ethical considerations were addressed by obtaining permission from the school and ensuring that student identities remained anonymous. Data processing was performed using Microsoft Excel. The questions consisted of 7 multiple-choice questions (numbers 1, 3, 4, 6, 7, 8, and 9), 1 matching question (number 2), and 2 true/false questions (numbers 5 and 10). The application used to analyze the questions is Microsoft Excel 2007. This application will determine the quality of multiple-choice items, including item validity, item reliability, difficulty index, discrimination index, and distractor efficiency.

Item Validity

Validity testing is one of the steps taken to evaluate the accuracy of an instrument used in research. The higher the instrument's validity, the more accurately it measures data [16]. The correlation value between each item score and the total score can be calculated using the Pearson Product-Moment formula as follows [17].

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\{n\sum x^2 - (\sum x)^2\}\{n\sum y^2 - (\sum y)^2\}}} \quad (1)$$

r = Pearson correlation coefficient
n = Number of respondents
x = Score for the item
y = Total score (or external variable)
xy = Product of paired scores

Table 1. Categories of validity [18]

Value	Category
$0.80 < r \leq 1.00$	Very high validity
$0.60 < r \leq 0.80$	High validity
$0.40 < r \leq 0.60$	Medium validity
$0.20 < r \leq 0.40$	Low validity
$0.00 < r \leq 0.20$	Very low validity
$r \leq 0.00$	Not valid

Item Reliability

Item reliability can be influenced by several factors, including the characteristics of test takers, test conditions, variations in test administration and scoring, the length of the test, the homogeneity of student abilities, and the difficulty index of test items [19]. The reliability coefficient in this test was calculated using the Kuder-Richardson formula (20), also known as KR-20, as follows [20].

$$\mathrm{KR} - 20 = \frac{\mathrm{k}}{\mathrm{k}-1}\left(1 - \frac{\sum \rho \mathrm{q}}{\sigma^2}\right) \quad (2)$$

KR-20 = Kuder-Richardson formula
k = The number of items
ρ = Number of students who answered correctly
q = Number of students who answered incorrectly
σ^2 = The variance of the total score

KR-20 is a good choice for measuring the reliability of questions with several question formats, such as multiple choice, true/false, and fill-in-the-blank [21].

Table 2. Categories of reliability values [18]

Value	Category
$0.80 < r \leq 1.00$	Very high reliability
$0.60 < r \leq 0.80$	High reliability
$0.40 < r \leq 0.60$	Medium reliability
$0.20 < r \leq 0.40$	Low reliability
$0.00 < r \leq 0.20$	Very low reliability
$r \leq 0.00$	Unreliable

Difficulty Index

The difficulty index is obtained by dividing the number of students who answer correctly on each item by the number of students who are respondents [22]. The higher the

percentage of students who answer a question correctly, the easier the question is, and vice versa [23]. The difficulty index of the questions can be expressed as a 3:4:3 ratio, meaning 30% in the easy category, 40% in the medium category, and 30% in the difficult category [10].

Table 3. Categories of difficulty index [15]

Value	Category
0.00 - 0.30	Difficult
0.30 - 0.70	Medium
0.7 - 1.00	Easy

Discrimination Index

The discrimination index is used to measure the ability of a question to distinguish between highly talented and less talented test takers [13]. The discrimination index can be calculated by ranking students from the highest to the lowest score. After that, the top 50% of scores are taken as the upper group and the bottom 50% of scores as the lower group [24]. The discrimination index (DI) is estimated by the formula [25],

$$\mathrm{DI} = 2\frac{(\mathrm{H}-\mathrm{L})}{\mathrm{N}} \quad (3)$$

DI = Discrimination index
H = The number of students in the upper group
L = The number of students in the lower group
N = Number of students in the upper and lower group

Table 4. Categories of discrimination index [26]

DI	Category	Recommendation
Negative	Defective items	Discard /correct the key
DI < 0.19	Poor	Discard
0.20 ≤ DI ≤ 0.29	Marginal	Reviewed
0.30 ≤ DI ≤ 0.39	Good	Need improvement
0.40 < DI	Excellent	Accepted

Distractor Efficiency

A distractor is an incorrect option in a multiple-choice question. Its purpose is to determine whether the person being tested can identify differences in a test [27]. A distractor that works is called a Functional Distractor (FD) and is chosen by more than 5% of students. Options picked by less than 5% of students are called Non-Functional Distractors (NFD) [28]. Questions that have NFDs should be revised with more plausible answer choices or removed from the test [29].

Item Validity of Minimum Competency Assessment (MCA)

Based on the MCA results at a junior high school in Surabaya, 15 students participated. Each multiple-choice question was scored as follows: 1 point for each correct answer and 0 points for each incorrect answer. This scoring system aligns with the MCA scoring framework published by the Ministry of Education, Culture, Research, and Technology. The following is a graphic of the results of total student scores.

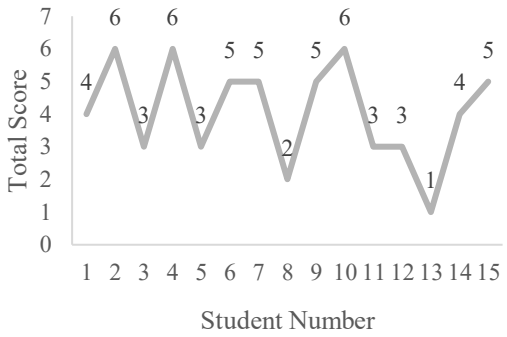

Figure 1. Graphic total student scores

Table 5. Validity test results of multiple-choice questions

No.	r-calculated	r-table	Criteria	Category
1	-0.228	0.514	Not valid	Not valid
2	-0.018	0.514	Not valid	Not valid
3	0.771	0.514	Valid	High
4	0.536	0.514	Valid	Medium
5	0.349	0.514	Not valid	Low
6	0.499	0.514	Not valid	Medium
7	0.770	0.514	Valid	High
8	0.588	0.514	Valid	Medium
9	-0.220	0.514	Not valid	Not valid
10	0.349	0.514	Not valid	Low

Based on Table 5, question numbers 1, 2, 5, 6, 9, and 10 were declared invalid, while question numbers 3, 4, 7, and 8 were declared valid. These valid and invalid criteria are generated from the comparison between the r-calculated and r-table. A questionnaire is said to be valid when the value of r-calculated is greater than the r-table (r-calculated > r-table) [8]. The calculated r value (0.514) is obtained from the correlation between each item's score and the total score. The r-table value is obtained from the Pearson r correlation table at the 5% significance level with n = 15.

The category column in Table 1 classifies the calculated r value. Based on this table, questions 1, 2, and 9 are categorised as invalid; questions 3 and 7 as having high validity; questions 4, 6, and 8 as having medium validity; and questions 5 and 10 as having low validity. This category is aligned with the contents of Table 1, as outlined by Guilford (1956). A test can have high validity when its measuring function is precise and the purpose of the measurement being carried out [7]. In this MCA test, there are still invalid and low-validity questions. Therefore, it is necessary to improve the question so that, when used as an assessment instrument, it can accurately assess students' knowledge.

Item Reliability of Minimum Competency Assessment (MCA)

Table 6 shows that the reliability of the multiple-choice questions is low. The reliability category for multiple-choice questions is consistent with the one in Table 2, as described by Guilford (1956). The reliability of the multiple-choice MCA questions was 0.213, which is low. These results indicate that the MCA questions are unreliable when retested, as the results will differ from the previous ones, even when tested on the same person.

Table 6. Reliability test for multiple-choice questions

No.	ρ	q	$\sum \rho q$	σ^2	KR-20	Category
1	0.53	0.47				
2	0.13	0.87				
3	0.60	0.40				
4	0.73	0.27				
5	0.07	0.93				
6	0.47	0.53	1.902	2.352	0.213	Low
7	0.47	0.53				
8	0.60	0.40				
9	0.40	0.60				
10	0.07	0.93				

Difficulty Index of Minimum Competency Assessment (MCA)

Table 7 presents the results of data processing for the difficulty index of multiple-choice items. Based on this table, it can be seen that question 4 is categorized as an easy question, questions 1, 3, 6, 7, 8, and 9 are categorized as medium questions, and questions 2, 5, and 10 are categorized as difficult questions. The category is adjusted to the difficulty index value (p) in Table 3. The results showed that 10% of the questions were categorized as easy questions, 60% of the questions were categorized as medium questions, and 30% of the questions were categorized as difficult questions. Thus, the MCA questions still lack a good balance. The difficulty index of 6 questions categorized as medium is a good question, because these questions have a high discrimination index and effective distractors [12].

Table 7. Difficulty index of multiple-choice questions

No.	Number of correct	Difficulty Index	Category
1	8	0.53	Medium
2	2	0.13	Difficult
3	9	0.60	Medium
4	11	0.73	Easy
5	1	0.07	Difficult
6	7	0.47	Medium
7	7	0.47	Medium
8	9	0.60	Medium
9	6	0.40	Medium
10	1	0.07	Difficult

Discrimination Index of Minimum Competency Assessment (MCA)

Based on Tables 8 and 9, the upper and lower groups are determined by the number of questions students answered correctly. Student number 2, who belongs to the upper group, answered 6 questions correctly, while student number 13, who answered only 1 correctly, was placed in the lower group. There is also a student who is not in the upper group or lower group, namely, student number 14. Although student number 14 and student number 1 have the same score, they belong to different groups because the difficulty index of the questions answered by student number 1 is lower than that of student number 14. Therefore, student 14 was placed in the middle group. The grouping of students is made to facilitate the determination of the question's discrimination index.

Table 8. Students grouping in upper group

Question No.	Student Number						
	2	4	10	6	7	9	15
1	0	0	1	0	0	0	1
2	0	0	0	0	0	0	1
3	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1
5	0	1	0	0	0	0	0
6	1	1	1	1	1	1	0
7	1	1	1	1	1	1	0
8	1	1	1	1	1	1	0
9	0	0	0	0	0	0	1
10	1	0	0	0	0	0	0
Total	6	6	6	5	5	5	5

Table 9. Students grouping in lower group

Question No.	Student Number						
	1	5	3	11	12	8	13
1	1	1	1	1	0	1	0
2	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0
4	1	1	0	0	1	1	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0
8	1	0	0	1	1	0	0
9	1	0	1	1	1	0	0
10	0	0	0	0	0	0	0
Total	4	3	3	3	3	2	1

Table 10. Discrimination index of multiple-choice questions

No.	DI	Category	Recommendation
1	-0.571	Defective item	Discard
2	0.000	Poor	Revision for incorrect key/Discard
3	0.714	Excellent	Accepted
4	0.429	Excellent	Accepted
5	0.143	Poor	Revision for incorrect key/Discard
6	0.714	Excellent	Accepted
7	0.714	Excellent	Accepted
8	0.429	Excellent	Accepted
9	-0.571	Defective item	Discard
10	0.143	Poor	Revision for incorrect key/Discard

Table 10 shows the discrimination index (DI), obtained by subtracting the number of lower groups who answered correctly from the number of upper groups who answered correctly, and dividing by the number of students in the lower and upper groups. The index is given a category according to Table 4. After that, it is decided whether the question can be used for assessment. Based on the table, questions 1 and 9 were categorized as defective items and should be discarded, questions 2, 5, and 10 were categorized as poor and should be revised for the incorrect key, and questions 3, 4, 6, 7, and 8 were categorized as excellent and

accepted. Rejected questions must be replaced with other questions because they cannot differentiate students' abilities, while questions with revised keys can be revised before being used as an assessment. There are 2 negative discrimination index results, likely due to damaged items/incorrect keys, and to low-ability students answering more correctly than high-ability students [30].

Distractor Efficiency of Minimum Competency Assessment (MCA)

Questions 2, 5, and 10 did not include a distractor analysis because they were presented as complex multiple-choice questions. Based on Table 11, in question number one (Q1), 33% of upper grade students chose answer A, 13% chose answer B, 0% chose answers C and D. As shown in Table 11 answer choices A and B are categorized as FD and do not need to be revised, while answer choices D and C are categorized as NFD and need to be revised because no students chose those answers. A good distractor will be chosen by lower group participants, while upper group participants tend to choose the correct answer [31]. According to the table, students in the upper grades tend to use the answer key. Consequently, answer options that are selected by fewer than 5% of students need to be revised.

Table 11. Distractor analysis of multiple-choice questions in upper-group students

Answer options	Q1	Q3	Q4	Q6	Q7	Q8	Q9
A	33%	0%	0%	0%	0%	7%	40%
B	13%	0%	0%	0%	40%	0%	7%
C	0%	47%	47%	40%	7%	40%	0%
D	0%	0%	0%	7%	0%	0%	0%
Answer key	B	C	C	C	B	C	B
Answer options	Q1	Q3	Q4	Q6	Q7	Q8	Q9
A	FD	NFD	NFD	NFD	NFD	FD	FD
B	FD	NFD	NFD	NFD	FD	NFD	FD
C	NFD	FD	FD	FD	FD	FD	NFD
D	NFD	NFD	NFD	FD	NFD	NFD	NFD

Table 12. Distractor analysis of multiple-choice questions in the lower group of students

Answer choice	Q1	Q3	Q4	Q6	Q7	Q8	Q9
A	7%	27%	7%	7%	7%	0%	13%
B	33%	13%	7%	0%	0%	27%	20%
C	7%	7%	27%	7%	27%	20%	7%
D	0%	0%	7%	33%	13%	0%	7%
Answer key	B	C	C	C	B	C	B
Answer options	Q1	Q3	Q4	Q6	Q7	Q8	Q9
A	FD	FD	FD	FD	FD	NFD	FD
B	FD	FD	FD	NFD	NFD	FD	FD
C	FD	FD	FD	FD	FD	FD	FD
D	NFD	NFD	FD	FD	FD	NFD	FD

Based on Table 12, the distribution of student answers is more even across all answer choices. Therefore, fewer answer choices need to be revised than in the upper grades. Compared with Table 11, upper-grade students tend to choose the answer key more often than lower-grade students. Questions 3, 4, 6, 7, and 8 have good distractors. However, in questions 1 and 9, the situation is reversed: lower-grade students choose the answer key more often than upper-grade students, so these questions can be considered to have weak distractors. This is because the questions' discrimination index is insufficient to distinguish between upper- and lower-grade students.

Based on the instrument analysis, several MCA items were selected for further discussion. The selection of these items was based on their psychometric characteristics, particularly those with negative or low discrimination indices and those that demonstrated good quality. This item sampling was conducted to provide a clearer evaluation of the assessment instrument's strengths and weaknesses.

Items 1 and 9 showed negative discrimination indices, indicating that lower-group students answered these items correctly more frequently than upper-group students. This result suggests that the items did not function properly in distinguishing students' ability levels. Similar findings were reported by [27] and [32], who stated that negative discrimination indices often indicate defective items caused by unclear wording, inappropriate distractors, or incorrect answer keys. Therefore, items with these characteristics should be discarded or thoroughly revised before being reused in assessments.

In contrast, items 3 and 7 demonstrated high validity and excellent discrimination indices. Most upper-group students selected the correct answers, while lower-group students tended to choose the distractors. This pattern indicates that the items effectively differentiated student ability levels. Similar results were also reported by [21], who found that items with moderate difficulty indices and effective distractors tend to show better discrimination power. Items with these characteristics can be retained as examples of well-constructed MCA questions.

Furthermore, items 2, 5, and 10 showed low discrimination indices and ineffective distractors, as several answer options were rarely chosen by students. This finding is consistent with studies by [28] which emphasized that non-functional distractors reduce the diagnostic value of multiple-choice items and negatively affect discrimination indices. These results indicate that future MCA item development should focus on improving distractor plausibility and alignment between indicators, cognitive levels, and answer options to ensure accurate and reliable competency measurement.

Conclusion

The analysis of the MCA multiple-choice questions revealed several key findings. Out of ten items, only four (questions 3, 4, 7, and 8) were found to be valid, and the other six are invalid. Invalid questions need to be revised to ensure they accurately assess student understanding. The overall reliability score was low (0.213), which indicates that the test results might not be consistent if given again. Most questions were moderately difficult, which is an ideal level for assessments. However, a few questions were either too easy

or too difficult and need to be adjusted. The discrimination index showed that questions 1 and 9 didn't effectively differentiate between upper and lower group students and should be replaced, while questions 2, 5, and 10 should be revised. Additionally, the distractor analysis showed that, in some cases, especially in questions 1 and 9, lower-group students chose the correct answer more often than upper-group students. This study highlights the importance of teachers developing and evaluating MCA questions. Reliable and valid assessment tools are essential to support meaningful learning evaluation and informed instructional decision-making. Further research is recommended to involve larger samples, different schools, and diverse science topics to strengthen the generalisability of findings and improve MCA question development practices.

Author's Contribution

Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing original draft preparation, writing review and editing, visualization, supervision, project administration, funding acquisition.

Acknowledgements

The author would like to express his gratitude to (1) the lecturer in charge of the Assessment of Learning and Outcomes course, (2) the teachers and administrators at the school, and (3) the students who participated in this research.

References

[1] Nurhikmah, I. Hidayah, and S. Kadarwati, "Persepsi dan Kesiapan Guru dalam Menghadapi Asesmen Kompetensi Minimum," *CJPE: Cokroaminoto Juornal of Primary Education*, vol. 4, no. 1, pp. 78–83, 2021, [Online]. Available: https://e-journal.my.id/cjpe

[2] Ansori and F. Putridiyanti, "Merdeka Belajar dalam Pendidikan Indonesia," *Jurnal Manajemen Pendidikan*, vol. 3, no. 1, pp. 1–13, 2022, [Online]. Available: https://ejournal.unuja.ac.id/index.php/jumpa

[3] Nofita, N. Kurniasih, and R. Yudi Purwoko, "Analisis Aspek Numerasi Siswa Kelas IX SMP pada Implementasi Asesmen Kompetensi Minimum," vol. 4, no. 2, pp. 13–22, 2022, [Online]. Available: http://jurnal.umpwr.ac.id/index.php/jipm

[4] F. Alfarisa, I. Ni, mah Chudari, and F. Robiansyah, "Analisis Butir Soal IPS Kelas V Sekolah Dasar Menggunakan Software ITEMAN Article Info," 2019. [Online]. Available: https://ejournal.upi.edu/index.php/edubasic

[5] D. P. Andini and M. Mukhlis, "Analisis Butir Soal pada Instrumen Asesmen Kompetensi Minimum Literasi Membaca di SMP IT Insan Utama Pekanbaru," *Diglosia: Jurnal Kajian Bahasa, Sastra, dan Pengajarannya*, vol. 6, no. 2, pp. 401–412, Apr. 2023, doi: 10.30872/diglosia.v6i2.658.

[6] H. Taherdoost, "Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research," *SSRN Electronic Journal*, Aug. 2018, doi: 10.2139/ssrn.3205040.

[7] S. Azwar, *Reliabilitas dan Validitas Aitem*, vol. 3, no. 1. Yogyakarta: Pustaka Belajar, 2019.

[8] F. D. P. Anggraini, A. Aprianti, V. A. V. Setyawati, and A. A. Hartanto, "Pembelajaran Statistika Menggunakan Software SPSS untuk Uji Validitas dan Reliabilitas," *Jurnal Basicedu*, vol. 6, no. 4, pp. 6491–6504, May 2022, doi: 10.31004/basicedu.v6i4.3206.

[9] A. Musyarofah, S. Abdurrahman Siddik Bangka Belitung, and I. Syaikh Abdurrahman Siddik Bangka Belitung, "Al-Mu'arrib: Jurnal Pendidikan Bahasa Arab Validitas dan Reliabilitas dalam Analisis Butir Soal," 2021. [Online]. Available: https://jurnal.lp2msasbabel.ac.id/index.php/AL-MUARRIB

[10] N. Sudjana, *Penilaian Hasil Proses Belajar Mengajar*. Bandung: PT Remaja Rosdakarya, 2017.

[11] S. Santosa and J. Badawi, "Analisis Butir Soal Pilihan Ganda Tema Pertumbuhan dan Perkembangan Makhluk Hidup Kelas III Madrasah Ibtidaiyah," *Jurnal Basicedu*, vol. 6, pp. 1678–1686, Feb. 2022, doi: 10.31004/basicedu.v6i2.2206.

[12] Purwanto, *Evaluasi hasil belajar*. Yogyakarta: Pustaka Pelajar, 2011. [Online]. Available: https://books.google.co.id/books?id=C6i_ZwEACAAJ

[13] O. Saleha and J. Marudut, "Analisis Butir Soal Ujian Akhir Sekolah Mata Pelajaran Bahasa Indonesia Kelas VIII SMP Negeri 3 Kutacane Tahun Pelajaran 2020/2021," 2022.

[14] Z. Arifin, *Evaluasi Pembelajaran*. Bandung: Remaja Rosdakarya, 2017. [Online]. Available: www.diktis.kemenag.go.id

[15] S. Erita and N. Amelia, "Eksplorasi Validitas dan Reliabilitas Soal Pemahaman Konsep dalam Asesmen Pembelajaran," *Jurnal Bima : Pusat Publikasi Ilmu Pendidikan bahasa dan Sastra*, vol. 2, no. 1, pp. 222–22131, Jan. 2024, doi: 10.61132/bima.v2i1.578.

[16] Y. Utami, P. M. Rasmanna, and Khairunnisa, "Uji Validitas dan Uji Reliabilitas Instrument Penilaian Kinerja Dosen," *Jurnal Sains dan Teknologi*, vol. 4, pp. 21–24, Feb. 2023, doi: 10.55338/saintek.v4i2.730.

[17] J. P. Guilford, *Fundamental statistics in psychology and education*, 3rd ed., vol. 3191. McGraw-Hill, 1956.

[18] H. S. Setyaedhi, "Comparative Test of Cronbach's Alpha Reliability Coefficient, Kr-20, Kr-21, And Split-Half Method," *Journal of Education Research and Evaluation*, vol. 8, no. 1, pp. 47–57, Feb. 2024, doi: 10.23887/jere.v8i1.68164.

[19] R. Sarwiningsih, "The Comparison Accuracy Estimation of Test Reliability Coefficients for National Chemistry Examination in Jambi Province on Academic Year 2014/2015," *JKPK (Jurnal Kimia dan Pendidikan Kimia)*, vol. 2, no. 1, p. 34, May 2017, doi: 10.20961/jkpk.v2i1.8740.

[20] S. Ntumi, S. Agbenyo, and T. Bulala, "Estimating the Psychometric Properties (Item Difficulty, Discrimination and Reliability Indices) of Test Items using Kuder-Richardson Approach (KR-20)," *Shanlax International Journal of Education*, vol. 11,

no. 3, pp. 18–28, Jun. 2023, doi: 10.34293/education.v11i3.6081.

[21] M. Fauzie, A. U. T. Pada, and S. Supriatno, "Analysis of the difficulty index of item bank according to cognitive aspects during the Covid-19 pandemic," *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 25, no. 2, Dec. 2021, doi: 10.21831/pep.v25i2.42603.

[22] D. Kumar, R. Jaipurkar, A. Shekhar, G. Sikri, and V. Srinivas, "Item analysis of multiple choice questions: A quality assurance test for an assessment tool," *Med J Armed Forces India*, vol. 77, pp. S85–S89, Feb. 2021, doi: 10.1016/j.mjafi.2020.11.007.

[23] S. Arikunto, *Dasar-Dasar Evaluasi Pendidikan Edisi 3*. Bumi Aksara, 2021. [Online]. Available: https://books.google.co.id/books?id=j5EmEAAAQBAJ

[24] I. Hendrika and C. Baka, "Pengaruh Pendekatan Pembelajaran Kontekstual Terhadap Hasil Belajar IPA Siswa Kelas IV UPT SDN 2 Tallunglipu," *Journal Elementary*, 2020.

[25] S. K. Bhat and K. H. L. Prasad, "Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology: A cross-sectional study," *Indian J Ophthalmol*, vol. 69, no. 2, pp. 343–346, Feb. 2021, doi: 10.4103/ijo.IJO_1610_20.

[26] R. M. Hadhoud, "Evaluating the effectiveness of the response-after MCQ exam item analysis on students' performance in a Communication Skills exam ," *Educational Research and Innovation Journal ERIJ*, vol. 5, no. 5, 2022.

[27] L. R. Sharma, "Analysis of Difficulty Index, Discrimination Index and Distractor Efficiency of Multiple Choice Questions of Speech Sounds of English," *International Research Journal of MMC*, vol. 2, no. 1, pp. 15–28, Feb. 2021, doi: 10.3126/irjmmc.v2i1.35126.

[28] A. A. Rezigalla *et al.*, "Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items," *BMC Med Educ*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12909-024-05433-y.

[29] M. Tarrant, J. Ware, and A. M. Mohammed, "An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis," *BMC Med Educ*, vol. 9, no. 1, 2009, doi: 10.1186/1472-6920-9-40.

[30] S. Bhattacherjee, A. Mukherjee, K. Bhandari, and A. Rout, "Evaluation of multiple-choice questions by item analysis, from an online internal assessment of 6 th semester medical students in a rural medical college, West Bengal," *Indian Journal of Community Medicine*, vol. 47, no. 1, pp. 92–95, Jan. 2022, doi: 10.4103/ijcm.ijcm_1156_21.

[31] Warju, R. S. Ariyanto, Soeryanto, and A. R. Trisna, "Analisis Kualitas Butir Soal Tipe HOTS Pada Kompetensi Sistem Rem Siswa di Sekolah Menengah Kejuruan," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 17, no. 1, 2020.