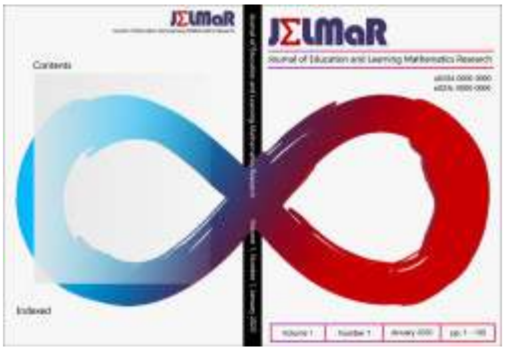


Journal of Education and Learning Mathematics Research (JELMaR)

Online ISSN : 2715-9787
Print ISSN : 2715-8535
Journal Homepage : http://jelmar.wisnuwardhana.ac.id/index.php/jelmar/index

Development Instrument Evaluation Mathematics HOTS - Based for Prospective Junior High School Teachers

Umy Zahroh, Dewi Asmarani

To cite this article Asmarani, D., & Zahroh, U. (2025). Development Instrument Evaluation Mathematics HOTS -Based for Prospective Junior High School Teachers. *Journal of Education and Learning Mathematics Research (JELMaR)*, *6*(2), 124-129. https://doi.org/10.37303/jelmar.v6i2.3840

To link this article: https://doi.org/10.37303/jelmar.v6i2.3840

Publisher
Department of Mathematics Education,
Faculty of Teacher Training and Education,
Universitas Wisnuwardhana Malang

Journal of Education and Learning Mathematics Research (JELMaR)
Volume: 6, Number: 2, November 2025, pp. 124-129
p-ISSN: 2715-8535, e-ISSN: 2715-9787

Received: 28 July 2025 *Revised: 30 August 2025* *Accepted: 30 September 2025* *Published: 3 November 2025*

Development Instrument Evaluation Mathematics HOTS - Based for Prospective Junior High School Teachers

[1]Umy Zahroh, [2*]Dewi Asmarani
State Islamic University of Sayyid Ali Rahmatullah Tulungagung
[*]Email: dewi_asmarani@uinsatu.ac.id ,

***Abstract**: Higher-order thinking skills (HOTS) are essential for meeting the challenges of the 21st century. Therefore, these skills must be possessed by every individual engaged in the learning process. Students, as learners, need frequent practice to develop these skills. To facilitate this, prospective teachers should be capable of designing assessment instruments that foster higher-order thinking. However, in practice, teachers often struggle to create such instruments effectively, as it is a complex task. This study was initiated to address that challenge. It employed a research and development (R&D) approach using the 4D model: define, design, develop, and disseminate. The research subjects were 15 prospective mathematics teachers who developed HOTS-based test items, which were then tested on 79 seventh-grade students at Darul Akhwan International Junior High School. The developed items were evaluated for their validity and practicality through construct and empirical validity tests. The practicality was further assessed using reliability tests, item difficulty levels, and discrimination indices. The results showed that the instrument achieved 100% validity based on expert judgment. The reliability test yielded an rcount=0.659r_{count} = 0.659rcount=0.659, which exceeds the rtable=0.2213r_{table} = 0.2213rtable=0.2213; thus, the instrument is considered reliable. The discrimination index analysis revealed that 60% of the items were in the "fair" category, while the remaining 40% were categorized as "good." Meanwhile, the difficulty level analysis showed that 15% of the items were "very difficult," 65% were "difficult," and 20% were "easy." These results indicate that the instrument has an appropriate level of difficulty. In conclusion, the assessment instrument developed in this study is both valid and practical for evaluating students' higher-order thinking skills.*

***Keywords**: HOTS, evaluation, mathematics*

INTRODUCTION

Problem-solving ability is one of the essential steps for students to successfully solve mathematical problems. In the 21st century, this ability has become a tool for selecting individuals to be admitted into certain institutions or organizations. Such ability relies heavily on a student's level of intellectual intelligence or IQ. This level can be measured through the student's performance in solving Higher-Order Thinking Skills (HOTS)-based questions.

In the 21st century, the ability to solve HOTS-based problems has become a key indicator in assessing the quality of human resources. According to a survey conducted by the OECD, Indonesian students' ability to solve PISA-related problems remains low. In 2015, Indonesia ranked 64th out of 72 participating countries. The average score in mathematics was 386, which marked an increase of only 11 points from the 2012 assessment. (Guria, 2015, p. 5) · The curriculum revision was carried out in 2016 (Regulation of the Minister of Education and Culture of the Republic of Indonesia Number 24 Core Competencies and Basic Competencies of Subjects in the 2013 Curriculum, 2016) and the 2018 PISA survey did not show a positive impact from the reforms. In that year, Indonesia was ranked 74th out of 79 countries with an average mathematics score of 379. (Summaries, 2019, p. 17) That matter show that ability student in finish question HOTS -based also still low.

This issue is influenced by two factors: internal and external. Internally, students' abilities play a key role in determining the extent to which they are capable of solving problems. However, external factors also contribute to students' problem-solving abilities, such as the availability of learning facilities, the learning environment, parental motivation, and teacher quality (Safiany & Maryatmi, 2018, p. 89).

Teachers are mediators for the students (Gunawan et al., 2023, p. 788) . Students' ability to solve HOTS questions is closely related to how well a teacher masters the process of solving such questions. A deep understanding of mathematical concepts and strong problem-solving strategies are essential before attempting to answer HOTS-based questions. Strong mathematical reasoning skills also significantly influence teachers, including Madrasah Tsanawiyah teachers, when designing steps to solve HOTS problems.

Teachers in the Madrasah Tsanawiyah setting are expected to be able to solve HOTS questions effectively. In this educational context, the ability to solve HOTS problems serves as an important indicator of a teacher's capacity to explain mathematical concepts in an integrative manner. The Madrasah Tsanawiyah environment is seen as a crucial moment for instilling fundamental mathematical concepts in an inclusive way. Therefore, students' understanding in Madrasah Tsanawiyah needs to be structured systematically so that mathematical concepts can be fully and coherently understood.

In order to measure the extent to which mathematical concepts have been acquired, an assessment tool commonly referred to as a mathematics test instrument is needed. While many instruments have been developed and studied with a focus on students, it is equally important to equip teachers with the skills to construct HOTS-based questions. This will enable them to better prepare students to face future challenges.

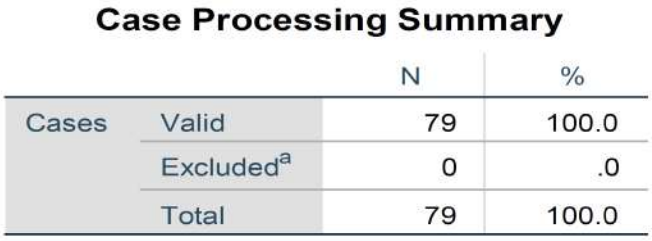
Teachers' ability to construct HOTS (Higher-Order Thinking Skills) questions is crucial in the evaluation process. Through evaluation, the effectiveness and success of the learning activities can be assessed. Moreover, evaluation serves as a tool to measure the success of previously planned programs. Therefore, to determine the achievement of learning objectives, a well-developed assessment instrument is essential. (Devi, 2011) Based on the results of observations conducted on 150 prospective mathematics teachers, only 50% were able to successfully construct HOTS questions appropriately. This means that the questions they developed met the criteria for levels C4, C5, and C6 of Bloom's Taxonomy. Meanwhile, the remaining 50% were still focused on lower-order thinking levels—C1, C2, and C3. This indicates that the ability of prospective Madrasah mathematics teachers at UIN SATU Tulungagung to construct HOTS-based questions still needs improvement. Therefore, as an effort to evaluate and enhance the abilities of future junior high school teachers, it is necessary to develop a HOTS-based assessment instrument designed by the prospective Madrasah Tsanawiyah teachers themselves, with the goal of producing instruments that are valid, practical, and effective.

METHOD

This study employs a research and development approach using the 4D model: define, design, develop, and disseminate. The subjects of the study were 15 prospective mathematics teacher students who developed HOTS-based questions, which were then tested on 79 seventh-grade students at Darul Akhwan International Junior High School. The HOTS questions developed were assessed for validity and practicality through both construct and empirical validity tests. To evaluate practicality, the instruments underwent reliability testing, item difficulty analysis, and discrimination index testing.

RESULTS AND DISCUSSION

The construct validity data of the developed HOTS questions indicates that the instrument is considered fairly valid based on expert judgment, with an average score of 2.7. However, experts recommended that 10 questions be eliminated, as they do not meet the criteria for HOTS-level questions.



Case Processing Summary

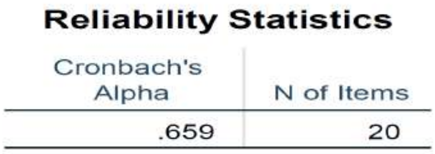
		N	%
Cases	Valid	79	100.0
	Excluded[a]	0	.0
	Total	79	100.0

The construct validity data of the developed HOTS questions indicates that the instrument is considered fairly valid based on expert judgment, with an average score of 2.7. However, experts recommended that 10 questions be eliminated, as they do not meet the criteria for HOTS-level questions.

Based on results, researchers get conclusion that question including in 100% valid category due to $r_{hitung}$is greater than r_{tabel}. This is in accordance with Holil et al. who stated that if r_{hitung}has a positive value and $r_{hitung} > r_{tabel}$with $\alpha = 0{,}05$, then the item is declared valid (Akbaryanto et al., 2023, p. 6568) . Statement Lismawari also expressed the same thing et al. who stated that If $r_{hitung} > r_{tabel}$then it can be concluded that the question can be declared valid (Sudiah et al., 2022, p. 4908) . In line with statement, Ni Nyoman stated that grains question is said to be valid if $r_{hitung} > r_{tabel}$ (Suariantini et al., 2023, p. 7).

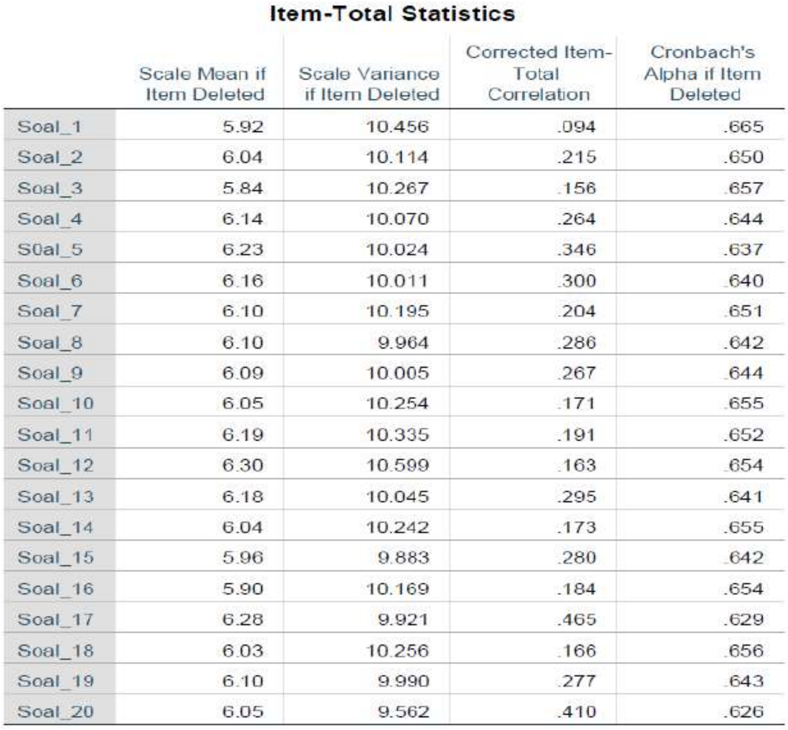
a. Reliability Test

The results of the reliability test, based on a trial involving all 79 seventh-grade students of Darul Akhwan International Junior High School, showed a Cronbach's Alpha value of 0.659. This indicates that the instrument is reliable and does not require revision. The following is a screenshot of the reliability test results along with the summary.



Reliability Statistics

Cronbach's Alpha	N of Items
.659	20



Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Soal_1	5.92	10.456	.094	.665
Soal_2	6.04	10.114	.215	.650
Soal_3	5.84	10.267	.156	.657
Soal_4	6.14	10.070	.264	.644
S0al_5	6.23	10.024	.346	.637
Soal_6	6.16	10.011	.300	.640
Soal_7	6.10	10.195	.204	.651
Soal_8	6.10	9.964	.286	.642
Soal_9	6.09	10.005	.267	.644
Soal_10	6.05	10.254	.171	.655
Soal_11	6.19	10.335	.191	.652
Soal_12	6.30	10.599	.163	.654
Soal_13	6.18	10.045	.295	.641
Soal_14	6.04	10.242	.173	.655
Soal_15	5.96	9.883	.280	.642
Soal_16	5.90	10.169	.184	.654
Soal_17	6.28	9.921	.465	.629
Soal_18	6.03	10.256	.166	.656
Soal_19	6.10	9.990	.277	.643
Soal_20	6.05	9.562	.410	.626

Based on criteria reliability, if $r_{hitung} \geq r_{tabel}$, then the item question is reliable and if $r_{hitung} < r_{tabel}$, so grains question it is not reliable. With $\alpha = 5\%$and $n = 79$, obtained $r_{tabel} = 0{,}2213$. Because the table above shows the value $r_{hitung} \geq r_{tabel}$, the items tested are included in the reliable category.

This reliability is in accordance with what was conveyed by Nining et al. who stated that question is reliable (consistent) if $r_{hitung} \geq r_{tabel}$ (NP Dewi et al., 2020, p. 145). In the study conducted by Aloisius, the Cronbach's Alpha test result was 0.699, indicating that the items fall into the category of questions with **moderate** reliability. (Son, 2019, p. 45) . In contrary Aloisius, Sabina Ndiung and Mariana Jediut state that If The *Cronbach Alpha* test value of 0.699 indicates that question the including in category grains questions that have reliability at the level tall (Ndiung & Jediut, 2020, p. 103) . However, if a common thread is drawn from the three opinions mentioned, it can be concluded that the items developed are reliable (consistent) test items.

b. Differential Test

The following is a classification of the discriminatory power test for HOTS questions that has been compiled.

Table 1. Classification of differential power question

$0{,}00 \leq DB < 0{,}20$	Bad
$0{,}20 \leq DB < 0{,}40$	Enough
$0{,}40 \leq TK < 0{,}70$	Good
$0{,}70 \leq TK < 1{,}00$	Very well

Furthermore For power test results different in HOTS questions can see as:

Table 2. Results of differential power test

Question Items	Pearson Correlation	Sig. (2-Tailed)	Information
1	.243	.031	Enough
2	.352	.001	Enough
3	.301	.007	Enough
4	.387	.000	Enough
5	.446	.000	Good
6	.416	.000	Good
7	.336	.002	Enough
8	.411	.000	Good
9	.395	.000	Enough
10	.310	.005	Enough
11	.309	.006	Enough
12	.251	.025	Enough
13	.409	.000	Good
14	.313	.005	Enough
15	.416	.000	Good
16	.329	.003	Enough
17	.542	.000	Good
18	.307	.006	Enough
19	.403	.000	Good
20	.527	.000	Good

Based on the results of the discrimination index test, 60% of the data indicate that the distributed instrument is fairly good, while the remaining 40% show that the tested instrument falls into the good category.

This finding is in line with the study conducted by Rahmawati et al., which defines four ranges of discrimination power. Two of these ranges are: 0.20 ≤ DI < 0.40, which falls into the category of sufficient discrimination power, and 0.30 ≤ DI < 0.70, which is categorized as having good discrimination power**.** (Rahmawati et al., 2022, p. 865). Based on the data obtained, it can be concluded that the discrimination power of the questions is fairly good**.**

c. Difficulty Level

Level of difficulty can classified as Table 4.9 as follows.

Table 3. Classification of difficulty levels

Mean	Information
$0{,}00 \leq TK < 0{,}20$	Show grains the test is very difficult
$0{,}20 \leq TK < 0{,}40$	Show grains test difficult
$0{,}040 \leq TK < 0{,}60$	Show grains test currently
$0{,}60 \leq TK < 0{,}80$	Show grains test easy
$0{,}80 \leq TK < 1{,}00$	Show grains the test is very easy

Furthermore, the results of the difficulty level test for the HOTS questions can be seen in Table 4 below.

Table 4. Results of HOTS question difficulty level

Question Items	Mean
Question 1	.49
Question 2	.37
Question 3	.57
Question 4	.27
Question 5	.18
Question 6	.24
Question 7	.30
Question 8	.30
Question 9	.32
Question 10	.35
Question 11	.22
Question 12	.10
Question 13	.23
Question 14	.37
Question 15	.44
Question 16	.51
Question 17	.13
Question 18	.38
Question 19	.30
Question 20	.35

Based on the data obtained, 15% of the questions fall into the very difficult category, 65% into the difficult category, and 20% into the easy category. These findings indicate that the instrument possesses a good level of difficulty, as it includes a range of question complexities that can effectively differentiate students' abilities. This is in accordance with the research conducted by Rahmawati et al., which outlines six difficulty level ranges. Among them are: 0.00 ≤ DL < 0.20, which classifies items as very difficult; 0.20 ≤ DL < 0.40, which indicates difficult items; and 0.60 ≤ DL < 0.80, which falls under the easy category. These ranges provide a clear framework for interpreting the appropriateness of item difficulty within an assessment instrument. (Rahmawati et al., 2022, p. 865).

Since 13 of the questions fall into the difficult category, it can be concluded that the overall set of questions possesses a moderate level of difficulty. This indicates that the questions are neither too difficult nor too easy, making them appropriate for assessing students' higher-order thinking skills in a balanced manner.

CONCLUSION

Based on the results of this research and development, it can be concluded that the developed instrument is considered fairly valid, as indicated by the construct validity test conducted by experts with an average score of 2.7. The empirical validity test of the items showed that the correlation coefficient $r_{x(y-1)}$ was greater than or equal to the critical value r_{tabel}, meaning that 100% of the tested items were valid and required no revision. The reliability test results showed a Cronbach's Alpha of 0.659, which also indicates that the instrument is reliable and does not require any revision. The discrimination index test results revealed that 60% of the items were categorized as fairly good, while the remaining 40% were in the good category. Furthermore, the difficulty level test indicated that 15% of the questions were very difficult, 65% were difficult, and 20% were easy, which suggests that the questions are not too difficult nor too easy. Therefore, the instrument's level of difficulty can be classified as appropriate. Based on these findings, the developed instrument can be concluded to be practical for use in assessing higher-order thinking skills.

REFERENCES

Devi, P.K. (2011). Pengembangan Soal "Higher Order Thinking Skill" dalam Pembelajaran IPA SMP/MTs. Diakses dari.https://www.academia.edu/8337926

Dewi, N. P., Rahmi, Y. L., Alberida, H., & Darussyamsu, R. (2020). Validitas dan Reliabilitas Instrumen Penilaian Kemampuan Berpikir Tingkat Tinggi ten-tang Materi Hereditas untuk Peserta Didik SMA/MA. *Jurnal Eksakta Pendidikan (Jep)*, *4*(2), 138–146.

Gunawan, A., Riyadi, A. A., & Musthofa, A. H. (2023). Kompetensi Guru Mata Pelajaran Agama Islam Dalam Meningkatkan Mutu Lulusan Peserta Didik di MTSN 1 Kota Kediri. *Jurnal Ilmu Multidisplin*, *1*(4), 788–798.

Guria, A. (2015). Pisa Result in Focus. In *PISA 2015*.

Ndiung, S., & Jediut, M. (2020). Pengembangan instrumen tes hasil belajar matematika peserta didik sekolah dasar berorientasi pada berpikir tingkat tinggi. *Premiere Educandum : Jurnal Pendidikan Dasar Dan Pembelajaran*, *10*(1), 94. https://doi.org/10.25273/pe.v10i1.6274

Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 24 Kompetensi Inti dan Kompetensi Dasar Pelajaran pada Kurikulum 2013, Pub. L. No. 24, 1 (2016).

Rahmawati, N. D., Komarudin, & Suherman. (2022). Pengembangan Instrumen Penilaian Matematika Berbasis HOTS Pada Calon Guru Sekolah Dasar. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *11*(2), 860–871.

Safiany, A., & Maryatmi, A. S. (2018). Hubungan self efficacy dan dukungan sosial teman sebaya dengan stres akademik pada siswa-siswi kelas XI di SMA Negeri 4 Jakarta Pusat. *IKRA-ITH HUMANIORA: Jurnal Sosial Dan Humaniora*, *2*(3), 87–95.

Son, A. L. (2019). Instrumentasi kemampuan pemecahan masalah matematis: analisis reliabilitas, validitas, tingkat kesukaran dan daya beda butir soal. *Gema Wiralodra*, *10*(1), 41–52.

Suariantini, N. N. G., Werang, B. R., & ... (2023). Instrumen Asesmen Numerasi Online Menggunakan Aplikasi Kahoot Pada Mata Pelajaran Matematika Kelas IV Sekolah Dasar. *Innovative: Journal Of …*, *3*(2). http://j-innovative.org/index.php/Innovative/article/view/930%0Ahttps://j-innovative.org/index.php/Innovative/article/download/930/767

Summaries, C. E. (2019). What Students Know and Can Do. *PISA 2009 at a Glance*, *I*. https://doi.org/10.1787/g222d18af-en