**Alifmatika: Jurnal Pendidikan dan Pembelajaran Matematika**
e-ISSN: 2715-6109 | p-ISSN: 2715-6095 Vol. 3, No. 2, December 2021
https://journal.ibrahimy.ac.id/index.php/Alifmatika **DOI:** 10.35316/alifmatika.2021.v3i2.159-169

# DEVELOPMENT OF MATHEMATICAL CRITICAL THINKING TEST INSTRUMENTS

**Dzulfiqar Satria Waliyuddin[1*], Dwi Sulisworo[2]**
[1,2]Ahmad Dahlan University, Yogyakarta, Indonesia
[1*]dzulfiqar2008050027@webmail.uad.ac.id, [2]dwi.sulisworo@uad.ac.id

| *Received:* | *Revised:* | *Accepted:* | *Corrigendum:* |
|---|---|---|---|
| *October 30, 2021* | *December 29, 2021* | *December 30, 2021* | *14, January 2022* |

**Abstract:**
Improving the quality of education to encourage students to have 21st century abilities needs to be balanced with a test instrument that is able to measure students' 21st century abilities. The ability to think critically is one of the 21st century abilities that every student must have. Based on the indications, this study was carried out to create a test instrument capable of testing students' critical thinking skills. This research is classified as development research since it is based on the following stages of development: information collection, planning, initial product development, limited trial, initial product revision, field trial, and final product revision. Scale and comparison are the topics. In Sleman, Indonesia, 11th-grade students participated in the product testing. To establish a good category in general, product validity is based on expert opinion, item validity, discriminatory power, level of difficulty, and dependability.

**Keywords:** Development, Mathematics, Critical Thinking, Instrument Test.

***How to Cite***: Waliyuddin, D. S., & Sulisworo, D. (2021). Development of Mathematical Critical Thinking Test Instruments. *Alifmatika: Jurnal Pendidikan dan Pembelajaran Matematika*, 3(2), 159-169. https://doi.org/10.35316/alifmatika.2021.v3i2.159-169

## INTRODUCTION

All the skills an individual needs to successfully meet the challenges of this century are 21st century skills. The increasing quality of education in the 21st century must be balanced with the increasing quality of the instruments used to measure the various abilities of students. One of them is a test instrument for assessing students' critical thinking skills in the classroom. The capacity of an individual to link, modify and transform the information and experience that has been owned critically in making judgments to handle an issue at hand is referred to as critical thinking ability (As'ari, Ali, Basri, Kurniati, & Maharani, 2019).

The ability to think critically A test instrument is a device that measures higher-order thinking skills, i.e. thinking skills that are more than merely remembering (recalling), restating, or referencing without processing (reading). Assessment based on critical thinking skills in the context of evaluating learning outcomes aims to measure an individual's ability to transfer one concept to another, process and apply information that he already has and will have, properly seek information from multiple sources, and use the information to solve

problems. issues, as well as critically analyzing ideas or information. Azizah, Sulianto, & Cintang (2018) gives the critical thinking indicators namely asking questions, planning strategies, and evaluating decisions.

Indeed, discussions of students' HOTS abilities frequently involve studies on higher-order thinking capabilities. Cayani & Saltifa (2021) conducted the most recent study on the creation of the HOTS instrument in mathematics teaching. Researchers created HOTS questions regarding numbers for this investigation. The devised test instrument contains categories that are both valid and trustworthy. The development of mathematical HOTS instruments must be continual. Shalikhah, Purnanto, & Nugroho (2021) reveals in his research that math problems in school textbooks do not match the requirements for acceptable HOTS questions. As a result, he proposed the inclusion of HOTS questions in student textbooks to assist each student in acquiring HOTS talents. HOTS development is still a current trend in mathematics education research. For example, PeranginAngin, Panjaitan, Hutauruk, Manik, & Tambunan (2021) discovered that the most researched research subject in Yogyakarta State University's Journal of Mathematics Education Research (JRPM) in 2020 and 2021 is student high-level ability (HOTS). The widespread emergence of HOTS questions has not been generally appreciated by instructors in the Madrasah Aliyah atmosphere in the Sleman region.

The striking difference from the research conducted with previous research is the form of questions developed in the instrument. The questions developed require students to be able to collect relevant information in the internet world in order to solve the problems given. This means that questions can only be done if students are able to collect valid information on the internet and make assumptions so that they can be used in solving problems. The purpose of this study was to create a test instrument to assess students' thinking skills in the Madrasah Aliyah Miftahunnajah Sleman setting. The test instrument in question is one that includes questions based on critical thinking markers (Al Ghamdi & Deraney, 201); Schraw & Robinson, 2011). It is intended that the findings of the instrument test from this development may be used as a reference for academics to better research on the creation of assessment instruments that foster higher-order thinking skills above and beyond students' critical thinking abilities (Djawad, 2018). Thus, the study's questions will be how to construct critical thinking questions and whether the questions generated are valid, trustworthy, and have excellent distinguishing power.

## RESEARCH METHODS

Research and development is the research approach employed. Research and Development (R & D) is a process or process to produce a new product or improve an existing product that can be accounted for (Sugiyono, 2017). Information gathering, planning, initial product development, restricted trial, first product revision, field trial, and final product revision were all employed to produce the product (Hamzah, 2021).

At the initial stage, the researcher designed a test instrument based on indicators of critical thinking skills. During the development stage, researchers gather data on the degree of validity of the questions based on the evaluation of

material experts and media experts. During the restricted trial phase, the researcher gathered data on the degree of validity and reliability of the questions, as well as the time necessary to complete all of them. Furthermore, researchers will solicit recommendations for enhancements to the equipment in development. During the field trial stage, the instrument was tested on a broad scale to establish the level of validity, reliability, complexity, and discriminating power of questions. The following diagram depicts the steps of research carried out in this study.

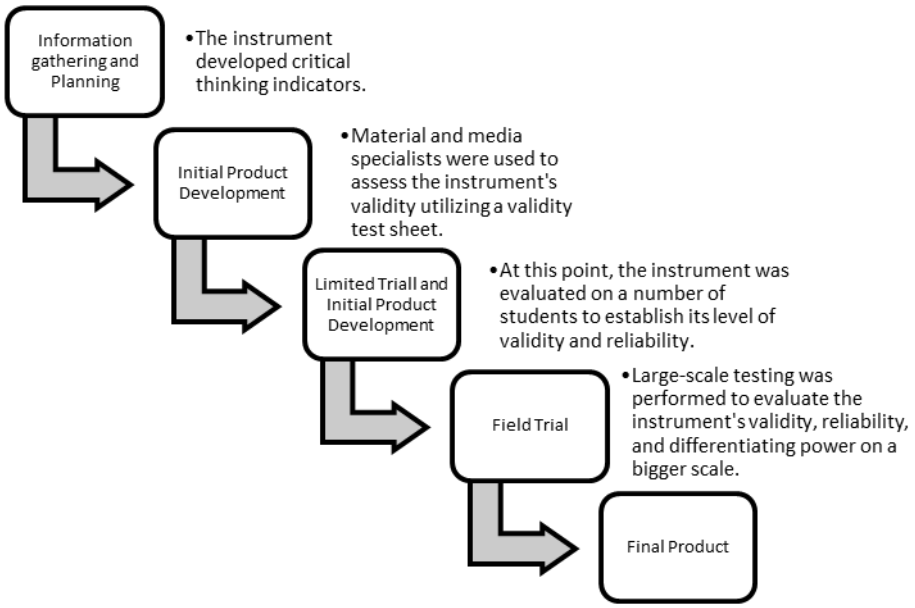


**Figure 1**. Instrument collection steps

The limitation of the research carried out is that the research is carried out on the subject of comparison and scale. According to the 2013 curriculum, one of the essential elements that junior high pupils must receive is comparison and scale. A material and media expert validation sheet will be utilized in the study to assess the level of validity of the questions generated in the test instrument, and the results will be assessed qualitatively. The validation of the test instruments used in this study attempts to evaluate the level of validity of the study's questions. The instrument validation procedure was carried out by five specialists in the field of mathematics education during the initial development stage.

After the validity of the instrument was established, it was tested on a limited basis with five non-experimental students. Meanwhile, field trials were held for 17 Madrasah Aliyah Miftahunnajah Sleman grade 11 pupils. The number of respondents is determined using a non-probability sampling method. The trial data were examined using traditional test theory parameters to empirically assess the instrument's quality and serve as the foundation for question revisions. The expert validation sheet's findings were reviewed to verify if the instrument's validation level value was consistent. To assess data in the form of the value of

each item from the expert evaluation, the content validity coefficient was derived using Aiken's V formula (Widyaningsih, Yusuf, Prasetyo, & Istiyono, 2021).

$$V = \frac{\sum S}{n(c-1)}$$

Where:

$\sum S$ : The difference between the expert's score and the minimal score is the total difference.

$n$ : The number of specialists who are involved.

$c$ : There is no limit to the number of score possibilities that can be provided.

The results of the study are then compared to Table 1 validity classification, which follows the class range.

Table 1. Validity Category for Experts (Arikunto, 2018)

| Range of Score | Validity of Categories | Remark |
|---|---|---|
| 0, 8 – 1,000 | Very high | Suitable for usage |
| 0.6 – 0.799 | High | Suitable for usage |
| 0.4 – 0.599 | Enough | Suitable for usage |
| 0.3 – 0.399 | Low | Not worth it to use |
| 0 – 0.199 | Very low | Not worth it to use |

Meanwhile, SPSS and Microsoft Excel were utilized to analyze data from the test results for the students. The analysis was carried out to determine the questions' validity, reliability, difficulty level, and discriminating capability. First, a validity study was done using a substantial threshold for the person's product-moment correlation. The Alfa Cronbach statistical test with the reliability criterion, on the other hand, was used in the reliability analysis (Yusup, 2018).

Using the derived difficulty level value, Table 2 was compared based on classification.

Table 2. Category of Difficulty

| Level of Difficulty | Category |
|---|---|
| Fewer than 0.3 | Very Difficult |
| 0.3-0.7 | Medium |
| More than 0.7 | Easy |

With the help of the same application, then the discriminatory power analysis is carried out using the following formula.

$$Discriminatory\ power = \frac{BA \times JB - BB \times JA}{JA \times JB}$$

Where:

*BA* : The percentage of participants in the top group who properly replied.

*BB* : The percentage of participants in the lower group who properly replied.

*JA* : The total number of people in the upper group.

*JB* : The total number of people in the lowest category

We compared the value of distinguishing power based on the categorization in Table 3 after getting the value of distinguishing power.

Table 3. Category of Distinguishing Power (Arikunto, 2018)

| Power Level Distinguishing | Category |
|---|---|
| 0.71-1.00 | Very well |
| 0.41-70 | Good |
| 0.21-0.40 | Enough |
| 0.00-0.20 | Not good |

Table 4 provides the value acquired by the student based on the scoring rubric as data from the test instrument to students in the form.

**Table 4**. Rubric for Scoring Instruments

| Score | Information |
|---|---|
| 2 | The answers provided are consistent with the meaning of the inquiry, and the computation process/source of information is valid. |
| 1 | The answers provided are in line with the purpose of the question, however there are still flaws in the calculating process/source of information. |
| 0 | The responses provided either do not reflect the intent of the query or do not respond at all. |

## RESULTS AND DISCUSSIONS

The gathering of data and product planning of the test instrument, which comprises six critical thinking questions on the theme of comparison and scale, is the first step in this study. The test instrument developed was based on critical thinking indicators, namely asking questions, planning strategies, and evaluating decisions. In other words, each developed question requires students to be able to ask a question that is used to identify important information in the problem, develop strategies based on the information that has been obtained to solve problems, and evaluate the work that has been done to obtain appropriate results. with the context of the problem (Azizah et al., 2018). Each question point developed is structured so that students are able to find relevant and valid information on the internet to support the problem solving process. For example, at one point, students were asked to look for valid information about the size of a tiger's body based on the results of other people's research.

The first stage of this study is product development. The trial phase was carried out to assess the questions' level of validity and reliability. The test is deemed to be legitimate if the findings fit the criteria, as in there are parallels between the test results and the criteria. A test's dependability relates to its level of stability, consistency, predictability, and correctness. High-reliability measures are those that can generate dependable data (Zhou, Zuo, Hou, & Zhang, 2017). At this point, the equipment was certified by investigators utilizing a validation checklist for materials specialists and media experts. SZN, a math instructor at SMP, completed a material and media expert review. Muhammadia Aya, Kebumen, ST as Math Teacher at SMK Muhammadia Mungkid, Magelang, and TDD as Math Teacher at Magelang and TDD Muhammadia Pakem, Sleman, CW as Math Teacher SMA IT Bina Ummah, Sleman as Math Teacher Miftahunnajah, and RD It also sends the tool's test script to the validator for validation. It was then examined using the Aiken formula V, and a coefficient of 1.00 was assigned to it. Having said that, test equipment is normally designed with user-friendliness in mind.

The restricted testing and early product modification stage follow. The instrument was then evaluated on a group of students using a nonprobability sampling approach to establish its level of validity and reliability (Vehovar, Toepoel, & Steinmetz, 2016). According to the findings of a five-student trial, the average time required for the five students to complete the test instrument was 60 minutes. The validity analysis was then rounded up to three decimal places using the SPSS tool, and data was taken from the outcomes of the student work evaluation. Based on the findings of the Pearson correlation analysis, the validity of the items was determined to be as follows.

Table 5. Data from the Limited Test Item Validity Analysis

| Question Items | Information |
|---|---|
| 1a | Valid |
| 1b | Valid |
| 2a | Invalid |
| 2b | Invalid |
| 3a | Valid |
| 3b | Valid |

Meanwhile, the Cronbach Alpha statistical test resulted in 0.91 in the reliability analysis. Based on the Cronbach alpha analysis results, the test instrument may be characterized as trustworthy. Field testing was carried out following restricted testing. However, prior to the test, items 2a, 2b were modified based on suggestions from a limited number of test participants. The suggestion is to add a given narrative description to the problem. Field tests are conducted after completion of subject improvement.

Field trials are the next step. The instrument that has been enhanced is evaluated on a broad scale at this stage to assess its validity, reliability, and differentiating power. The validity of the items, the instrument's dependability, the level of difficulty, and the discriminating power of the questions were all tested in field trials. Field experiments were conducted on 17 11th grade students from Madrasah Aliyah Miftahunnajah Sleman. Students are allocated 60 minutes of processing time at the stage of working on the test instrument, based on the average length of time for students in a restricted trial. Data on the scores attained for each item were acquired during on-the-job training. Item validity analysis was carried out utilizing the SPSS program and Pearson's statistical correlation test.

The conclusion of the item validity test findings is as shown in Table 6 based on the Pearson correlation analysis results.

Table 6. Data Item Validity Results Field Test Results

| Items for Discussion | Information |
|---|---|
| 1a | Valid |
| 1b | Valid |
| 2a | Valid |
| 2b | Valid |
| 3a | Valid |
| 3b | Valid |

Simultaneously, the Cronbach Alpha statistical test resulted in a value of 0.91 in the reliability analysis stage. Based on the Cronbach alpha analysis results, it is possible to determine that the test equipment is dependable. The next stage is to assess the problem's complexity and discriminating power. The analysis is carried by utilizing the Microsoft Excel program. Students are classified as senior or junior based on their average daily control point as compared to the Criteria for Complete Study (CCM) in Compulsory Mathematics. Table 7 summarizes the findings of the analysis.

Table 7. Data on the Difficulty Level and Distinguishing Power of Questions

| Items for Discussion | Level of Difficulty | Category | Distinguishing Characteristics | Category |
|---|---|---|---|---|
| 1a | 0.55 | Medium | 0.67 | Good |
| 1b | 0.60 | Medium | 0.73 | Good |
| 2a | 0.63 | Medium | 0.77 | Good |
| 2b | 0.53 | Medium | 0.65 | Good |
| 3a | 0.61 | Medium | 0.70 | Good |
| 3b | 0.55 | Medium | 0.63 | Good |

According to the chart above, 100 percent (all) of the test fixture pieces fall into the middle group. This is not an issue when building a tool to assess students' critical thinking capacity because questions that can test higher-order thinking abilities are not usually challenging (As'ari et al., 2019). The statistics, on the other hand, revealed that 100 percent (total) of the test tool's components had good discriminating power. Table 8 shows the overall statistical instrument test results based on the data acquired from the field test scores.

Table 8. Table of Overall Statistical Test Results

| Scale of Measurement | Amount | Interpretation |
|---|---|---|
| The average sign. Correlation test with two tails | 0.002 | Valid |
| Cronbach's Alpha (Cronbach's Alpha Value) | 0.91 | Reliable |
| Average level of difficulty | 0.59 | Medium |
| Average discerning ability | 0.69 | Good |

Based on the facts presented above, the average Sig is calculated. The values for the (two-sided) correlation test and Cronbach's alpha were 0.002 and 0.91, respectively. That is, beneficial elements are included in the constructed test equipment. Furthermore, the device tests developed are classified as reliable at the level of dependability. When the difficulty and features of the item were analyzed, the average difficulty and characteristics were 0.59 and 0.69, respectively. This indicates that products made on test devices generally have a moderate level of complexity and a high degree of distinguishability. Researchers did not get ideas for increasing the quality of their test instruments during field testing. As a result, the field test device was used as the end product of this development research.

## CONCLUSIONS AND SUGGESTIONS

The following findings are reached as a consequence of study and debate. The first step in this research is the collecting of material and the product planning of the test instrument, which includes six questions on critical thinking skills. Indicators of critical thinking skills are included in the tools created. During the original product development phase, researchers verified the instrument's validity with material and media specialists utilizing a validity test sheet. Validation actions are done out by providing the validator with the instrument test script. The data was then evaluated using Aiken's V method, yielding a coefficient of 1.00. In other words, test instruments, in general, include elements that are very easy to use. After determining that the instrument could be tested, the researcher went through a trial phase (restricted testing, first product revision, and field trial) to achieve the final product.

This study's ultimate outcome is a test instrument designed to assess students' critical thinking abilities. The acquired test instrument consisted of six questions in the form of a description. Based on the instrument analysis, the average sig. Correlation test (2-tailed) and Cronbach's alpha values were 0.002 and 0.91, respectively. That is, the test instrument was created with valid elements. Furthermore, the test instrument created is categorized as dependable at the level of dependability. The average level of difficulty and discriminating power of items is 0.59 and 0.69, respectively, when the level of difficulty and discriminating power are analyzed. This outcome indicates that the test instrument items have a moderate level of difficulty and good discriminating power.

**ACKNOWLEDGEMENT**

This study is dedicated to Indonesian teachers and educators. We would like to express our gratitude to MA Miftahunnajah Sleman for providing us with the venue and opportunity to conduct this research. Hopefully, the findings of this study would be beneficial to MA Miftahunnajah's family as well as all Indonesian instructors and educators.

**REFERENCES**

Al Ghamdi, A. K. H., & Deraney, P. M. (2013). Effects of Teaching Critical Thinking to Saudi Female University Students Using a Stand-Alone Course. *International Education Studies*, *6*(7), 176–188.

Arikunto, S. (2018). *Dasar-Dasar Evaluasi Pendidikan Evaluasi Pendidikan Edisi 3*. Jakarta: Bumi Aksara.

As'ari, A. R., Ali, M., Basri, H., Kurniati, D., & Maharani, S. (2019). Mengembangkan HOTS (Higher Order Thinking Skills) Melalui Matematika. *Universitas Negeri Malang*.

Azizah, M., Sulianto, J., & Cintang, N. (2018). Analisis keterampilan berpikir kritis Siswa sekolah dasar pada pembelajaran matematika kurikulum 2013. *Jurnal Penelitian Pendidikan*, *35*(1), 61–70.

Cayani, S., & Saltifa, P. (2021). Pengembangan Soal Higher Order Thinking Skill (HOTS) Materi Bilangan Di Sekolah Menengah Pertama. *Jurnal Equation: Teori Dan Penelitian Pendidikan Matematika*, *4*(1), 103–114.

Djawad, Y. A. (2018). Innovation In Learning Through Digital Literacy at Vocational School of Health. *International Conference on Indonesian Technical Vocational Education and Association (APTEKINDO 2018)*, *201*, 239–245. Atlantis Press.

Hamzah, A. (2021). *Metode Penelitian & Pengembangan (Research & Development) Uji Produk Kuantitatif dan Kualitatif Proses dan Hasil Dilengkapi Contoh Proposal Pengembangan Desain Uji Kualitatif dan Kuantitatif*. CV Literasi Nusantara Abadi.

PeranginAngin, R. B., Panjaitan, S., Hutauruk, A., Manik, E., & Tambunan, H. (2021). Arah dan Trend Penelitian Pendidikan Matematika di Jurnal Riset Pendidikan Matematika (JRPM). *Vygotsky: Jurnal Pendidikan Matematika Dan Matematika*, *3*(1), 49–62.

Schraw, G., & Robinson, D. H. (2011). *Assessment of higher order thinking skills*. IAP.

Shalikhah, N. D., Purnanto, A. W., & Nugroho, I. (2021). Soal Higher Order Thinking Skills (HOTS) Matematika Pada Buku Tematik Terpadu Kurikulum 2013. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *10*(2), 701–709.

Sugiyono, S. (2017). Metode Penelitian Bisnis: Pendekatan Kuantitatif, Kualitatif, Kombinasi, dan R&D. *Penerbit CV. Alfabeta: Bandung*.

Vehovar, V., Toepoel, V., & Steinmetz, S. (2016). Non-probability sampling. *The Sage Handbook of Survey Methods*, 329–345.

Widyaningsih, S. W., Yusuf, I., Prasetyo, Z. K., & Istiyono, E. (2021). The Development of the HOTS Test of Physics Based on Modern Test Theory: Question Modeling through E-learning of Moodle LMS. *International Journal of Instruction*, *14*(4), 51–68.

Yusup, F. (2018). Uji validitas dan reliabilitas instrumen penelitian kuantitatif. *Tarbiyah: Jurnal Ilmiah Kependidikan*, *7*(1), 17–23.

Zhou, P., Zuo, D., Hou, K.-M., & Zhang, Z. (2017). A decentralized compositional framework for dependable decision process in self-managed cyber physical systems. *Sensors*, *17*(11), 2580.