International Journal of Geography, Social, and Multicultural Education
https://journal.unesa.ac.id/index.php/ijgsme

# The Suitability of HOTS-Based Geography Questions in Terms of Validity, Reliability, and Effectiveness of Distractors

Andreas Sito Saputra[1] (Universitas Pendidikan Indonesia, Indonesia)
Siti Julpa[2] (Universitas Pendidikan Indonesia, Indonesia)
Mamat Ruhimat [3] (Universitas Pendidikan Indonesia, Indonesia)
Dina Siti Logayah [4] (Universitas Pendidikan Indonesia, Indonesia)
Igna Juli Triana[5] (Universitas Pendidikan Indonesia, Indonesia)

**Received:** 02-06-2025
**Accepted:** 08-01-2026
**Published:** 01-02-2026
**DOI:** https://doi.org/10.26740/ijgsme.v3n3.p14-22
**e-ISSN** : 2987-9140
**Volume** : 3 No. 3, (2026)
**Page** : 14-22

**Abstract**

This research is based on the importance of preparing evaluation instruments that are able to measure higher order thinking skills in a valid, reliable, and fair manner. Therefore, the purpose of this study is to analyze the quality of question items in Higher Order Thinking Skills (HOTS)-based Geography subjects that have been tested on class XI students of SMAN 2 Cikampek in the 2024/2025 academic year. The method used is descriptive with a quantitative approach, with a sample of 30 students selected randomly. The research instrument was in the form of 20 multiple choice questions that were done online through the Quizizz platform. Data analysis techniques include validity testing using product moment correlation, reliability testing using the Kuder-Richardson formula (KR-21), as well as analysis of difficulty level, differentiating power, and the effectiveness of exacerbators. The results showed that 14 out of 20 questions (70%) were valid with a high reliability value of 0.8122. Based on the level of difficulty, 65% of questions are classified as easy and 35% are classified as moderate, without any difficult category questions. Differentiating power analysis showed that 57% of the questions were categorized as very good, 21% as good, and the rest were classified as sufficient and insufficient. Meanwhile, most of the exemptions did not function optimally. These findings indicate that although most questions have met the criteria for good quality, improvements are still needed, especially in the aspect of exemptions.

**Keywords: Question items, HOTS, Validity, Reliability, Geography**

## 1. INTRODUCTION

In recent years, education systems in various countries have begun to shift from emphasizing lower-order thinking skills (LOTS) to higher-order thinking skills (HOTS) (Barak & Dori, 2009, in Suhendro, Sugandi, & Ruhimat, 2021). Higher Order Thinking Skills are a way of thinking at a higher level than simply memorizing facts or relaying information to someone (Heong et al., 2017 ).According to Anderson and Krathwohl (2001, in Jati, Ruhimat, and Logayah, 2024), learning evaluation is an activity that collects, processes, and displays measurement results in the form of students' mastery of material at a certain time and provides a score for the object being assessed (W3) states that HOTS questions no longer only test students' ability to recall, restate, or recite without processing, but also involves the ability to analyze, evaluate, and create. The primary objective of HOTS can be said to be an effort to develop students' thinking skills so that they are more critical in receiving various types of information and also think creatively in solving problems.

Corresponding Author: E-mail: anis.alghiffary@gmail.com

Learning is an effort made by educators to provide understanding to students in order to facilitate the achievement of expected learning objectives. According to Wini and Ruhimat (2018, p. 2), geography learning is an educational activity that involves the roles of both students and teachers, where students study geography comprehensively by observing various processes of interrelationships between natural factors and humans, which can be viewed through the lenses of natural and social sciences and then combined into an interaction between the earth and humans/Human-Earth system. Learning objectives have three aspects: knowledge, skills, and attitudes. Knowledge refers to the ability to remember and recognize previously studied material. To assess students' understanding in the knowledge aspect, evaluation activities are necessary.

Evaluation activities are essential in the learning process. Through evaluation, educators can determine the outcomes of the learning activities carried out by students. These outcomes then inform follow-up actions needed to improve the quality of the learning process. One form of evaluation activity to measure student learning outcomes is through the use of tests. Tests are measuring tools commonly used to measure students' learning achievements in cognitive areas such as knowledge, understanding, application, analysis, synthesis, and evaluation. Tests can also be defined as a set of questions or instructions in a certain number, either oral or written, which will be completed by students in the assessment process.

Item analysis is a process in which the answers to test questions and the test questions themselves are examined in their entirety (Siri and Freddano, 2011 in Logayah, Ruhimat, Arrasyid, and Islamy, 2024). According to Aiken (1994) in Depdiknas (2008), the purpose of item analysis is to improve the quality of test items and obtain diagnostic information about students. A high-quality question is one that provides accurate information, enabling the identification of students who have mastered the material and those who have not. A good evaluation test has characteristics such as validity, reliability, difficulty level, discriminative power, and the effectiveness of distractors on the data obtained. Therefore, the author conducted research on 20 geography questions that had been created and submitted to 30 respondents to determine the accuracy and quality of the 20 questions..

## 2. METHOD

The method used in this study is a descriptive method with a quantitative approach. A quantitative approach is a study based on phenomena or symptoms that have cause and effect obtained from a process of collecting data from a specific population and sample. Quantitative data analysis describes/illustrates the collected data without making general conclusions (Sugiono, 2015). The sample consisted of 30 students randomly selected from class XI of SMAN 2 CIKAMPEK for the 2024/2025 academic year. The data collection technique used was a Quizizz test. The test format analyzed was multiple choice or multiple response, where respondents were not limited by time and could complete it online from anywhere.

The analysis techniques used involved assessing validity, reliability, difficulty level, discriminative power, and the effectiveness of distractors on the obtained data. Each was calculated using Microsoft Excel. Quantitative data analysis used the raw product-moment correlation formula to calculate validity, the Kuder and Richardson (KR 21) formula to calculate test reliability, the difficulty level formula to calculate the difficulty level of each item, and Arikunto's (2018) formula to calculate the discriminating power of each item

## 3. RESULTS AND DISCUSSION

### Validity Test

After collecting the respondents' answers, the first step was to conduct a validity test with 20 geography questions based on the respondents' answers. This test was conducted manually using Excel, which then produced the calculated r (the calculated r is the correlation value of each question), which was then compared with the table r at a significance level of 0.05. Since the number of research subjects is 30, at a significance level of 0.05, n = 28, resulting in a value of 0.3610. A question is considered valid if the table r is smaller than the calculated r, and vice versa; if the table r is larger than the calculated r, the question is not valid.

Question validity refers to the alignment of the question with what is being assessed. The questions tested in this activity must align with the material taught by the teacher in previous learning activities. A question is considered valid if it aligns with the main topic presented to the students (Ida & Musyarofah, 2021). Based on the analysis of the 20 geography questions that were administered, 14 questions were valid, or 80% valid. This percentage indicates that the questions distributed were appropriate or on target. The following table shows a comparison of the number of valid questions with invalid questions.

Table 1. Details of the Validity of 20 Geography Questions

| N0 | Validity | Question Number | Total | Persentase |
|---|---|---|---|---|
| 1 | >0.3610 | 1,3,5,9,10,11,12, 14,15,16, 17, 18, 19, 20 | 14 | 80% |
| 2 | <0.3610 | 2,4,6,7,8,13 | 6 | 20% |
| **Total** | | | 20 | 100% |

According to Sudijono (2003), the validity of a test item is the accuracy of measurement in measuring what should be measured through that item. A valid question (80%) means that the question can perform its function, which is to measure what should be measured. However, there are still some invalid questions. This is similar to Gronlund's theory in

(Arifin, 2017), who states that the level of validity can be influenced by the instruments used in assessment, scoring, and test participants' answers. Therefore, the 20 geography questions created by the author need to be revised to fulfill their function in measuring what they are intended to measure. The following is a breakdown of valid and invalid questions based on their calculated r values.

Table 2. Distribution of geography questions based on validity

| No Question | r Calculate | Result |
|---|---|---|
| 1 | 0.3641 | Valid |
| 2 | 0.3323 | Not Valid |
| 3 | 0.6249 | Valid |
| 4 | 0.3463 | Not Valid |
| 5 | 0.5168 | Valid |
| 6 | 0.1871 | Not Valid |
| 7 | 0.3476 | Not Valid |
| 8 | 0.2415 | Not Valid |

| | | |
|---|---|---|
| 9 | 0.5012 | Valid |
| 10 | 0.4230 | Valid |
| 11 | 0.7769 | Valid |
| 12 | 0.6314 | Valid |
| 13 | 0.3113 | Not Valid |
| 14 | 0.3826 | Valid |
| 15 | 0.4602 | Valid |
| 16 | 0.7441 | Valid |
| 17 | 0.5777 | Valid |
| 18 | 0.4813 | Valid |
| 19 | 0.5911 | Valid |
| 20 | 0.6502 | Valid |

**Reliability Test**

After conducting a validity test, valid items are then selected to measure their reliability. The purpose of conducting a reliability test is to determine the level of confidence in the test instrument, so that the higher the score, the better the reliability. This confidence is related to the consistency of the instrument when repeated measurements are taken. Reliability measurement is conducted manually using Excel and the Kuder and Richardson (KR 21) formula. The results of these calculations are then interpreted using the criterion that if $\geq 0.70$, the item can be considered to have high reliability. The following is the calculation of the reliability level. From the above formula, a reliability value of 0.8122 was obtained, meaning that the fourteen questions have a high level of reliability.

**Testing The Level Of Difficulty**

In addition to validity and reliability, the quality of a learning assessment instrument is greatly influenced by the balance of question difficulty levels. The level of difficulty is the ratio of students who can answer questions correctly, thus illustrating how challenging the questions are for students, not for teachers or test makers. According to Arikunto (2013), a perfect question is not one that is entirely easy or entirely difficult, but rather one that has a balanced variation in difficulty—consisting of easy, medium, and difficult questions. The recommended composition is 30% easy questions, 40% medium questions, and 30% difficult questions. The purpose of this is to enable the questions to identify students with low, moderate, and high abilities. The determination of difficulty levels should be based on empirical analysis, i.e., based on actual data from students' answers, not just assumptions by the question creators. A question that is considered easy by the teacher may not necessarily be easy for students with different backgrounds of knowledge or understanding. If all questions are too easy, the evaluation tool will not be able to assess competence comprehensively. Conversely, if they are too challenging, students may become frustrated and the test results may be less accurate as a learning evaluation tool.

Therefore, it is crucial for teachers to analyze test items to obtain information about the level of difficulty, discrimination power, and efficiency of distractors in the test items (Arifin, 2017:28 in Astuti, Hapsan, Herianto, and Warsyidah 2024). With this approach, teachers not only function as test developers but also as evaluators and quality controllers of the learning

assessment process. In this way, the assessment tools used will objectively and accurately reflect students' abilities and contribute to improving the overall quality of education. The difficulty level analysis of the questions in this study uses a single formula.

Table 3.Table of interpretation scale for difficulty levels according to Witherington

| No | Interpretation Scale | Level of Difficulty |
|---|---|---|
| 1. | 0,00 - 0,30 | Difficult |
| 2. | 0,31 - 0,70 | Medium |
| 3. | 0,71 - 1,00 | Easy |

The results of the difficulty level analysis in this study are as follows:

Table 4. Table of test results for question difficulty

| Question | P = Number of correct answers / number of students | |
|---|---|---|
| | Index | Criteria |
| Question 1 | 0,73 | Easy Questions |
| Question 3 | 0,50 | Medium Questions |
| Question 5 | 0,73 | Easy Questions |
| Question 9 | 0,70 | Medium Questions |
| Question 10 | 0,83 | Easy Questions |
| Question 11 | 0,50 | Medium Questions |
| Question 12 | 0,73 | Easy Questions |
| Question 14 | 0,97 | Easy Questions |
| Question 15 | 0,80 | Easy Questions |
| Question 16 | 0,63 | Medium Questions |
| Question 17 | 0,77 | Easy Questions |
| Question 18 | 0,80 | Easy Questions |
| Question 19 | 0,50 | Medium Questions |
| Question 20 | 0,80 | Easy Questions |

From the results of the formula calculation on 14 valid questions that have been tested on students, the researcher concluded that there were 9 questions (65%) in the easy category with question numbers 1, 5, 10, 12, 14, 15, 17, 18, and 20. There were also 5 questions (35%) in the moderate category, meaning that the questions were neither too difficult nor too easy, with question numbers 3, 9, 11, 16, and 19. Meanwhile, there were 0 questions in the difficult category, meaning that there were no questions at all in the difficult category. This is certainly undesirable because difficult questions are also needed to train students' thinking skills. Therefore, the researcher acknowledges that a limitation of this study on question quality is that the difficulty level of the questions is still not balanced between easy, moderate, and difficult questions. The follow-up action for these questions is to select them by removing the very easy questions and using the difficult questions (Anatasia, in Nurinda et al., p. 79, in Susetyo 2020.) However, easy questions can be saved and reanalyzed, and their structure can be modified to make them difficult so that they can be reused by students as ready-to-use questions.

**Distinctivity Test**

Next, the valid items were selected and the discrimination power was measured manually using Excel by subtracting the average number of correct answers from the maximum score

for each item. The discrimination power of an item is necessary to distinguish between high-ability and low-ability students (Arikunto, 1999, p. 211). The following table shows the distribution based on discrimination power.

Table 5. Distribution of 14 questions based on discriminating power around

| No | Distinctive Power | Item Number | Total | % |
|---|---|---|---|---|
| 1 | 0,19 - Down (Not good) | 14 | 1 | 7 |
| 2 | 0,20 - 0,29(Fair)) | 15,18 | 2 | 15 |
| 3 | 0,30 - 0,39(Good) | 9,10,17, | 3 | 21 |
| 4 | 0,40 / > (Very Good) | 1,3,5,11,12,16, 19, 20 | 8 | 57 |
| Total | | | 20 | 100% |

From the table above, it can be seen that out of 14 valid questions, around 57% of the questions passed with a very good rating and 21% with a good rating. Meanwhile, around 15% of the questions were rated as adequate and around 7% were rated as poor. These questions need to be completely revised by investigating the causes of their shortcomings or failures. Questions with low distinctive power cannot be used as a reference in assessing students with high abilities and those with low abilities. Efforts that can be made to improve questions with low distinctive power are to improve the language of the questions and answer options so that they are not ambiguous and confusing to students with high abilities. The following are the details of the distinctive power of each question.

Table 6. Details Distinctive Power 14 valid questions

| Question Items | AVERAGE TOP | AVERAGE BOTTOM | DISTINCTIVE POWER |
|---|---|---|---|
| 1 | 0.93 | 0.53 | 0.40 |
| 3 | 0.73 | 0.27 | 0.47 |
| 5 | 1.00 | 0.47 | 0.53 |
| 9 | 0.87 | 0.53 | 0.33 |
| 10 | 1.00 | 0.67 | 0.33 |
| 11 | 0.87 | 0.13 | 0.73 |
| 12 | 0.93 | 0.53 | 0.40 |
| 14 | 1.00 | 0.93 | 0.07 |
| 15 | 0.93 | 0.67 | 0.27 |
| 16 | 0.87 | 0.40 | 0.47 |
| 17 | 0.93 | 0.60 | 0.33 |
| 18 | 0.93 | 0.67 | 0.27 |
| 19 | 0.73 | 0.27 | 0.47 |
| 20 | 1.00 | 0.60 | 0.40 |

**Power Test**

Table 7. Results of the power test

| Question Number | Frequency Answer Choice | | | | | Answer Key | Description |
|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | | Pengecoh A, B, dan D kurang baik |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | **23** | 1 | 5 | **C** | Distractor C is not good |
| 3 | 4 | 2 | 1 | **16** | 7 | **D** | Distractors C and E are not good |
| 5 | 2 | 5 | 0 | **23** | 0 | **D** | Distractors C and D are not good |
| 9 | 3 | 8 | 0 | 0 | **19** | **E** | Distractors B, D, and E are not good |
| 10 | 2 | 1 | **26** | 0 | 1 | **C** | Distractor E is not good |
| 11 | 4 | 2 | **15** | 8 | 1 | **C** | Distractor D dan E are not good |
| 12 | **23** | 4 | 2 | 1 | 0 | **A** | All distractions are not good. |
| 14 | **29** | 0 | 0 | 0 | 0 | **A** | Distractors B, C, and E are not good |
| 15 | **25** | 1 | 0 | 3 | 1 | **A** | Distactor B is not good |
| 16 | **21** | 0 | 2 | 3 | 4 | **A** | Distactors A, B, and D are not good |
| 17 | 1 | 1 | **22** | 1 | 5 | **C** | Distractors B and E are not good |
| 18 | 3 | 0 | **24** | 2 | 1 | **C** | Distractor E is not good |
| 19 | 4 | 3 | **17** | 5 | 1 | **C** | Distractors A, D, and E are not good |
| 20 | 1 | **24** | 5 | 0 | 0 | **B** | Distractors A, B, and D are not good |

Based on the available table of distractor effectiveness test results, a comprehensive analysis can be conducted on how effective each distractor is for each question. Distractor effectiveness is a crucial indicator of the quality of multiple-choice questions. An effective distractor is an answer choice other than the correct answer that is selected by a number of test takers, indicating that the option is sufficiently appealing and confusing for test takers who have not fully understood the material. Conversely, a distractor that is never selected or only selected by one test taker is considered an ineffective distractor and requires improvement. Of the 14 questions analyzed (questions 1, 3, 5, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, and 20), it is evident that many distractors are not functioning optimally. For example, question number 1 shows that neither distractor A nor B was selected at all (frequency 0), while only option C was selected the most (23 students). This indicates that the distractors are not sufficiently appealing and need to be improved.

In question number 5, no students chose distractor D, and only two students chose A. This shows that the distractors are not appealing and need to be changed or improved. A similar situation is seen in question number 14, where all distractors have very low frequencies (0–1 students), indicating that the question is too simple or the distractors are not confusing at all. This can reduce the ability of the questions to distinguish between students who truly understand the material and those who do not. However, not all questions reflect similar problems. Some questions show quite effective deception. For example, question number 20 has distractors A and D, which were chosen by 1 and 5 students, respectively, indicating that these distractors are quite attractive and function well. Question number 19 also shows that distractors B and D were chosen by more than one student, indicating their role as fairly effective distractors.

In general, this analysis concludes that many distractors in these questions are not functioning well. Distractors that are not chosen at all or are chosen by only one person indicate that the option is unable to deceive test takers and needs to be improved. Creating distractors that are more similar to the correct answer but still conceptually incorrect is crucial in creating high-quality questions. Assessing the appeal of incorrect answers should be an important element in question analysis to ensure that the evaluation tools used are truly valid, reliable, and capable of assessing students' abilities fairly and objectively.

## 4. CONCLUSION

This study found that of the 20 geography questions tested, 70% were declared valid and had high reliability with a KR-21 value of 0.8122, indicating that the test instrument was quite reliable. The majority of questions were at an easy (65%) and moderate (35%) level of difficulty, with no questions in the difficult category, indicating that the distribution of question difficulty was not balanced. In terms of distinctive power, most questions have a good ability to distinguish between high and low ability students, but there are some questions with less than optimal distinctive power. The analysis of distractors shows that many distractor options are ineffective because they are rarely or never chosen by participants, so improvements are needed to increase their function in misleading students who do not yet understand the material well. The limitations of the study include the relatively small number of respondents (30 students) and the use of manual data analysis with Excel, which may introduce errors. Therefore, it is recommended that future studies use a larger sample size and more advanced statistical analysis tools to ensure more valid and representative results. Additionally, more attention should be given to balancing the difficulty levels of the questions and improving the quality of distractors to make the test more comprehensive and effective as an evaluation tool for learning.

## REFERENCES

Arikunto, S. (2021). *Dasar-Dasar Evaluasi Pendidikan Edisi* 3. Bumi aksara.

Astuti, N. D., Hapsan, A., Herianto, M., Warsyidah, A. A., Riskawati, N. M., Febriana, B. W., & Toron, V. B. (2024). *Prinsip-prinsip Pengukuran dan Evaluasi Pendidikan: Disertai dengan Contoh Kasus*. CV. Ruang Tentor.

Ida, F., & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. AL-*Muarrib Journal Of Arabic Education*, 1(1), 34-44. https://doi.org/10.32923/al-muarrib.v1i1.2100

Jati, P., Ruhimat, M., & Logayah, D. S. (2023). Analisis Kualitas Butir Soal Geografi. *Jurnal Pendidikan Geosfer*, 8(2), 287-298.

Logayah, D. S., Ruhimat, M., Arrasyid, R., & Islamy, M. R. F. (2024). Item analysis of National Geography Olympiad multiple-choice questions (MCQs) in Indonesia. *Cogent Social Sciences*, 10(1), 2354971.

Munandar, A., Maryani, E., Rohmat, D., & Ruhimat, M. (2020). Establishing the Profesionalism of Geography Teacher through Authentic Assessment Field Study. *International Journal of Instruction*, 13(2), 797-818.

Rahmayati, D., Suwarni, N., & Miswar, D. (2013). Analisis Butir Soal Ujian Semester Ganjil Mata Pelajaran Geografi Siswa Kelas X SMA Bina Mulya Bandar Lampung Tahun Pelajaran 2011/2012. *JPG (Jurnal Penelitian Geografi*), 1(2).

Sugiyono, Cahyadi. "*Metode Penelitian dan Pendidikan." Bandung (Pendekatan Kuantitatif dan Kualitatif):* Alfabeta (2015).

Susetyo, A. M. (2020). Analisis butir soal ujian semester kelas VIII mata pelajaran bahasa Indonesia. *BELAJAR BAHASA: Jurnal Ilmiah Program Studi Pendidikan Bahasa dan Sastra Indonesia*, 5(2), 187-198

Winardi, Gunawan. (2002). *Panduan Mempersiapkan Tulisan Ilmiah*. Bandung: Akatiga.

Yee Mei Heong, Widad Binti Othman, Jailani Bin Md Yunos, Tee Tze Kiong, Razali Bin Hassan, and Mimi Mohaffyza Binti Mohamad, (2011). The Level of Marzano Higher Order Thinking Skills among Technical Education Students. *International Journal of Social Science and Humanity*. 1 (2). 121-125.