European Journal of Education Studies
ISSN: 2501 - 1111
ISSN-L: 2501 - 1111
Available online at: www.oapub.org/edu

DOI: 10.46827/ejes.v10i12.5089 Volume 10 | Issue 12 | 2023

# MEASURING HIGHER-ORDER THINKING SKILLS IN SCIENCE AMONG PRIMARY SCHOOL STUDENTS USING ITEM RESPONSE THEORY

**Sandag Gendenjamts**[i]
Head,
Department of Item Bank Development,
Education Evaluation Center,
Mongolia

**Abstract:**
Higher-order thinking skills (HOTS) are crucial competence in education. Higher-order thinking skills can help learners solve problems and decision making efficiently by anticipating connections between divergent ideas. The present study aims to develop reliable and valid instruments to assess higher-order thinking skills in science among primary school students. The study followed eight stages of developing a model adapted from a previous study. The total sample of this research comprised 428 fifth-grade students from six primary schools located in urban and rural areas in Mongolia. The gathered data were analyzed using SPSS 22.0 and STATA 16.0 to examine the item characteristics curve, test reliability, and item correlation. The study recommends developing creativity skills through exercise-based activities, so those item developers could produce reliable and valid instruments to assess HOTS.

**Keywords**: higher-order thinking skills, assessment, test instrument, measurement, reliability, validity

## 1. Introduction

In the 21st century, technological advancement and changes in the socioeconomic climate and workplace require future citizens to have a wide range of skills to face new challenges (OECD, 2015; Otgonbaatar, 2021a). To address these challenges, educators and international organizations have emphasized specific skills, such as critical thinking, creative thinking, problem-solving, and decision-making, encapsulated under the term "Higher-Order Thinking Skills (HOTS)" (Anderson & Krathwohl, 2001; Scully, 2017). However, these skills have been described with different terms, such as 21st-century skills (The Partnership for 21st Century Skills, 2009), transversal competencies (UNESCO, 2015), and social and emotional skills (OECD, 2015). The concept of HOTS connects to

[i] Correspondence: email gendenjamts01@gmail.com, tsengelee2000@gmail.com

 

Bloom's Taxonomy of Educational Objectives, and mainly corresponds with the top three levels of the taxonomy: analyzing, evaluating, and creating. (Anderson & Krathwohl, 2001; Nitko & Brookhart, 2007; Scully, 2017). Most countries report that these skills are not taught as separate subjects but incorporated across the curriculum, according to OECD (2015) and Ontario (2015). These studies identified the importance of developing skills in relation to specific subjects, rather than as topics for separate teaching. Thus, there is a call for education systems to intentionally emphasize and develop these specific through deliberate changes in curriculum design and pedagogical practice (Ontario, 2016; Otgonbaatar, 2021b). Students' HOTS are fostered through a more collaborative process across all subjects, which means that a person cannot develop these skills in isolation (Lawson, 1993; Shellens & Valcke, 2005).

Notably, primary and secondary educational reforms primarily referenced the poor results of fourth and eighth graders in the Trends in International Mathematics and Science Study (TIMSS)- 2011, which showed that Mongolian students performed very poorly in mathematics and natural sciences (39.6% for 4th graders, and 25.8% for eighth graders). These scores highlight the unacceptable quality of education and the inability of the education sector to meet labor market needs. From research conducted at the national level, we can see that learning achievement is not progressing at all, and the result is below 60% at all education levels (Education in Mongolia, a country report, 2019, p. 8). According to Brookhart (2010) and Tanujaya, Mumu & Margono (2017), there is a linear, positive, and robust relationship between HOTS and students' academic achievement, and if we can successfully assess higher-order thinking, we find that it increases student achievement.

## 2. Material and Methods

The study is conducted using a correlation research model. The researcher used existing research to develop approaches followed by the Borg & Gall Model (1983), which states that there are 10 steps in the test development process. The current study consists of eight stages of developing a model adapted from the Borg & Gall model: 1) Needs Analysis, 2) Planning, 3) Develop the Preliminary Form of the Product, 4) Field Testing, 5) Product Revision, 6) Operational Field Testing, 7) Final Product Revision, and 8) Dissemination and Implementation. The target population consisted of fifth-grade primary school students in Mongolia. The total sample of this research was 428 fifth-grade students from six primary schools, three located in Ulaanbaatar and three in Orkhon Province, Mongolia. Since 2013, the Mongolian government has been introducing and implementing a new curriculum nationwide, which these schools already implement. The question type selected was multiple-choice and open-ended, which Paul and Nosich (1992) argue is the best approach for assessing HOTS.

## 3. Data analysis and results

Data were collected from a pilot test. The pilot testing research data were gathered and input into SPSS 22.0. The reliability coefficient was 0.62. The sample size of the test was 58 fifth-grade students. Since the initial Cronbach's alpha of 0.62 was unacceptable, the researcher conducted a revision, after which Cronbach's alpha increased to 0.71 in the large-scale sample. Based on the results above, the reliability coefficient in the collected data of the sample was 0.71, within the acceptable range. Nunnally (1967) states that 0.70 – 0.80 is a good range useful for a classroom test.

A detailed analysis of descriptive statistics was conducted on large samples. The minimum, maximum, mean, median, mode, and standard deviation were calculated and are shown in Table 1.

**Table 1:** Descriptive statistics on student performance

| Item number | N | Min | Max | Mean | Median | Mode | Std. Deviation |
|---|---|---|---|---|---|---|---|
| Q1 | 428 | 0 | 1 | .58 | 1.00 | 1 | .494 |
| Q2 | 428 | 0 | 2 | 1.50 | 2.00 | 2 | .742 |
| Q3 | 428 | 0 | 2 | .67 | .00 | 0 | .764 |
| Q4 | 428 | 0 | 1 | .47 | .00 | 0 | .500 |
| Q5 | 428 | 0 | 2 | .75 | 1.00 | 1 | .566 |
| Q6 | 428 | 0 | 2 | .63 | 1.00 | 0 | .674 |
| Q7 | 428 | 0 | 1 | .36 | .00 | 0 | .480 |
| Q8 | 428 | 0 | 2 | 1.14 | 1.00 | 1 | .718 |
| Q9 | 428 | 0 | 1 | .27 | .00 | 0 | .443 |
| Q10 | 428 | 0 | 1 | .76 | 1.00 | 1 | .427 |
| Q11 | 428 | 0 | 2 | .98 | 1.00 | 1 | .802 |
| Q12 | 428 | 0 | 3 | .74 | .00 | 0 | .957 |
| Valid *N* (listwise) | 428 | | | | | | |

The item parameter is a fundamental concept of IRT. Item discrimination shows the ability of an item to differentiate between good and poor students based on how well an item can discriminate. The characteristic of a better test item is that high-ability students will answer it correctly more frequently than lower-ability students. The item discrimination parameter expresses how well an item can be differentiated among examinees with different ability levels. Satisfactory and good items usually have discrimination values ranging from 0.5 to 2. High discrimination indicates that higher-scoring candidates tend to answer the item correctly, while lower-scoring candidates tend to answer it incorrectly.

Item difficulty is one of the essential concepts in psychometrics and is the most useful item in analysis statistics. The item difficulty, known as the $[p]$ parameter, is essentially the percentage of examinees who answered the item correctly. The greater the difficulty of an item, the higher an examinee's ability must be to answer that item correctly. Items with greater difficulty are hard items, which low-ability examinees are

unlikely to answer correctly. If items with low difficulty are easy items, most examinees will get that item correct (Otgonbaatar, 2016).

According to Table 2, most of the items included a medium category of difficulty and they satisfied the discrimination index category. Item 12 (creating skill) has the highest item difficulty index (0.25), which means it is the hardest. Item 10 (analyzing skill) has the lowest item difficulty index (0.76), making it the easiest item. In addition, Items 5 (evaluating skill) and 9 (analyzing skill) have the lowest item discrimination index, meaning that they are least able to distinguish between examinees who are knowledgeable and those who are not.

**Table 2**: Item response theory parameters

| Item | Diff. *P*-value | Criteria | Disc *d*-value | Criteria | Answer | Alternate Weight | Means |
|---|---|---|---|---|---|---|---|
| 1 | 0.58 | Medium | 0.72 | Satisfactory | A | 1 | 8.27 |
| 2 | 0.75 | Easy | 0.44 | Satisfactory | A | 1 | 7.35 |
| | | | | | Correct idea | 1 | |
| 3 | 0.33 | Medium | 0.56 | Satisfactory | If not, estimate the cloud | 1 | 8.18 |
| | | | | | Correct graph | 1 | |
| 4 | 0.47 | Medium | 0.41 | Satisfactory | B | 1 | 8.38 |
| 5 | 0.37 | Medium | 0.15 | Poor | True | 1 | 8.1 |
| | | | | | Correct idea | 1 | |
| 6 | 0.31 | Medium | 0.41 | Satisfactory | Correct picture | 1 | 8.22 |
| | | | | | Correct idea | 1 | |
| 7 | 0.36 | Medium | 0.29 | Needs revision | B | 1 | 8.49 |
| 8 | 0.57 | Medium | 0.54 | Satisfactory | Mishel | 1 | 7.71 |
| | | | | | Correct idea | 1 | |
| 9 | 0.27 | Hard | 0.16 | Poor | A | 1 | 8.58 |
| 10 | 0.76 | Easy | 0.47 | Satisfactory | B | 1 | 8.09 |
| 11 | 0.49 | Medium | 0.69 | Satisfactory | Yes | 1 | 7.87 |
| | | | | | Correct idea | 1 | |
| 12 | 0.25 | Hard | 0.47 | Satisfactory | Correct response | 1 | 8.11 |
| | | | | | Any validate the method | 1 | |
| | | | | | Any validate the interpret | 1 | |
| | .25–.76 | | .15–.72 | | | 20 | |

Classical test theory-based item difficulties were found. The item difficulty index (p-value) is classified into three ranges: $p < 0.3$, too difficult; $0.31 \leq p \leq 0.7$, good or acceptable; and $p > 0.7$, too easy. The following formula (Güler, 2014) was used to calculate item difficulty indices for openended items:

$$\text{Item difficulty index} = \left(\frac{x - y}{z - y}\right)$$

x: Mean scores received from the item;
y: The minimum score receivable from the item;
z: The maximum score receivable from the item;

The item discrimination index (d-value) falls in the ranges $d \geq 0.40$, quite satisfactory; $0.30 \leq d \leq 0.39$, good; $0.20 \leq d \leq 0.29$, marginal and needs revision; and $d \leq 0.19$, poor.

According to Table 3, in the test instrument, five items were intended to measure analyzing skills. Since these items measure the same thing, the items should be correlated with each other. The intercorrelations of all items were analyzed by looking for the data result of each question from among the 428 samples. All three types of items had a high correlation, so it was determined that the items measure the same thing. The results are shown in Tables 3, 4, and 5.

**Table 3:** The Pearson correlations of analyzing skill items

| **Component of HOTS** | **Item number** | **Pearson Correlation** | | | | |
|---|---|---|---|---|---|---|
| | | **Q1** | **Q4** | **Q7** | **Q9** | **Q10** |
| Analyze (C4) | Q1 | 1 | .128** | .059 | .007 | .271** |
| | Q4 | .128** | 1 | −.041 | −.030 | .024 |
| | Q7 | .059 | −.041 | 1 | .014 | .063 |
| | Q9 | .007 | −.030 | .014 | 1 | .052 |
| | Q10 | .271** | .024 | .063 | .052 | 1 |

**Table 4:** The Pearson correlations of evaluating skill items

| | | **Q2** | **Q5** | **Q8** | **Q11** |
|---|---|---|---|---|---|
| Evaluate (C5) | Q2 | 1 | .245** | .273** | .273** |
| | Q5 | .245** | 1 | .240** | .208** |
| | Q8 | .273** | .240** | 1 | .378** |
| | Q11 | .273** | .208** | .378** | 1 |

**Table 5:** The Pearson correlations of creating skill items

| | | **Q3** | **Q6** | **Q12** |
|---|---|---|---|---|
| Create (C6) | Q3 | 1 | .300** | .217** |
| | Q6 | .300** | 1 | .245** |
| | Q12 | .217** | .245** | 1 |

Based on the tables above, we found the correlations between analyzing skills (Q1, Q4, Q10), evaluating skills (Q2, Q5, Q8, Q11), and creating (Q3, Q6, Q12) to be significant (p-value = 0.01). These questions can be used to measure the same skill and can serve as an instrument to measure higher-order thinking skills. In Figure 1 below, item characteristic curves (ICC) are shown for items of analyzing skills. The left-hand curve or q10 represents the easiest item. It shows that the probability of the correct answer is higher for low-ability students and closer to 1 for high-ability students. According to Figure 1, the ICC for analyzing the skill of an item is intended to measure ability. The probability of this defined success increases as the ability increases. The probability of correct

answers changes very quickly as the examinee's ability increases. This is an easier item and low-ability examinees should perform correctly on it.

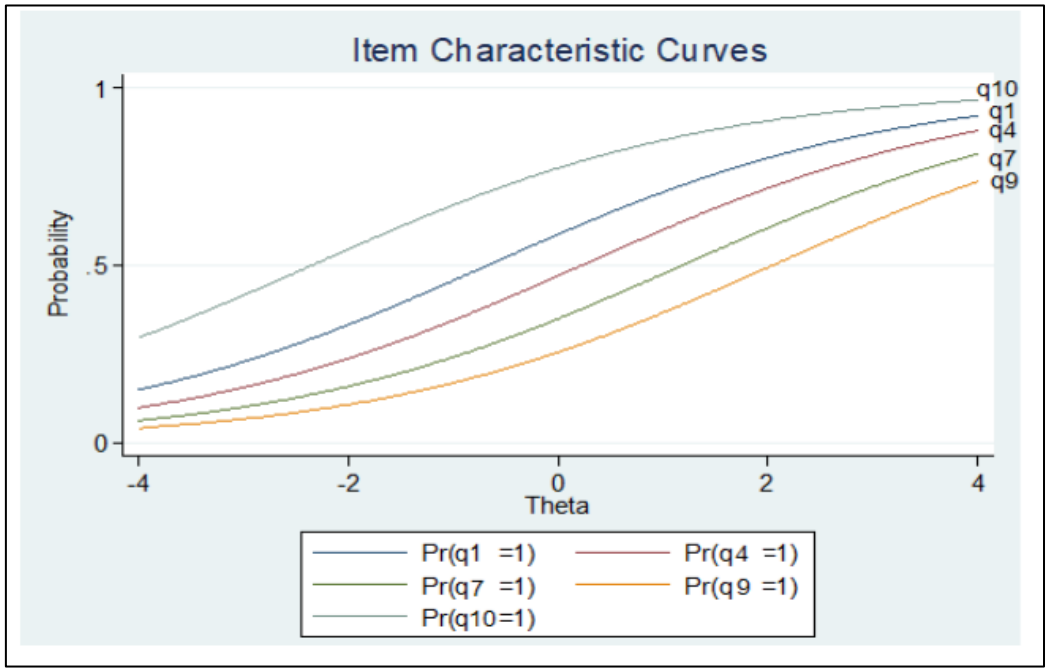


**Figure 1:** ICC of Analyzing skill

The range of HOTS tasks included in the survey assessment allows for describing six levels of problem-solving proficiency (Table 6).

**Table 6:** Relationship between items and student performance on a higher-order thinking scale (adapted from PISA 2012)

| | | | | |
|---|---|---|---|---|
| Items with relatively high difficulty and discrimination index | Level VI<br>Level V | Item 3<br>Item 12<br>Item 9<br>Item 11 | Student A, with relatively high proficiency | A student can successfully complete items up to level V and probably items at level VI as well |
| Items with moderate difficulty and discrimination index | Level IV<br>Level III | Item 6<br>Item 7<br>Item 4<br>Item 8 | Student B, with moderate proficiency | B student can complete items at levels I and II, and probably items at level III as well; but not items at levels V and VI, and probably not level IV either |
| Items with relatively low difficulty and discrimination index | Level II<br>Level I | Item 5<br>Item 1<br>Item 2<br>Item 10 | Student C, with relatively low proficiency | C student is unable to complete any items from level II to VI, and probably not level I either. |

The first level is the lowest described level, and it corresponds to an elementary level of higher-order thinking skills. The top level corresponds to the highest level of higher-order thinking skills. Students with a proficiency score within the range of the first level are expected to complete most elementary-level tasks successfully, but they are unlikely to be able to complete tasks at higher levels. Students with scores in the last level range

are likely to be able to complete all the tasks included in the survey assessment of problem-solving.

## 4. Discussion and Conclusion

The procedure described in this study to develop and validate the higher-order thinking skills test items was mainly in line with the checklist suggested for the preparation of multiple-choice and open-ended question constructions (Paul & Nosich,1992; Haladyna, 1997; TIMSS 2011 items).

The reliability coefficient in the collected data of the sample was 0.71, which is acceptable. Therefore, the test items developed in this study accurately measure higher-order thinking skills among the target population.

All items were developed to measure the three components of higher-order thinking skills, and the level of Bloom's taxonomy was higher than applying. The items did not use ambiguous sentences or words, such as item, stem, table, or figure. All items were intercorrelated, and all items converged on the same construct. Therefore, it is believed that the items used in this study have high content and construct validity. Item analysis revealed that items for analyzing skills have moderate item validity coefficients, while those for evaluating and creating skills have higher validity coefficients. It can be judged that the items are valid and tend to measure the same skill.

Results show that students' performance on the HOTS test was below the expected average score. Notably, performance on the creating skill tasks was lower than on the analyzing and evaluating tasks. Finally, the test instruments used to measure higher-order skills are reliable and valid for the purpose of this study. The performance of higher-order thinking skills at the national level was low among the target population. The students are less trained in solving HOTS-related tasks. The fifth-grade students' skills are weak in creativity to solve a problem, intellectual analysis, making assumptions, and ability to execute independent actions. Similar findings were reported in a study that examined creativity among Mongolian students (Otgonbaatar, 2020). School students are less trained in solving items of higher-order thinking. This may have multi-faceted reasons. One of the causal factors is that the students might be unfamiliar with the item formats and how the questions were posed. Mongolian primary school students do not receive much training in solving higher-order thinking items or demanding higher-order thinking activities.

This study determined that the level of Mongolian fifth-graders' HOTS and the status of implementation of the new curriculum appear to be low, showing that the quality of reform implementation still has challenges.

**Acknowledgements**
I would like to thank the Japan International Cooperation Agency (JICA) for their support throughout this research paper. I am extremely grateful to my main supervisor, Shimizu

Kenya, for his support, wisdom, and encouragement along the way. He always offered guidance and valuable advice when needed.

**Conflict of Interest Statement**
The author declares no conflicts of interest.

**About the Author(s)**
Gendenjamts Sandag is Head of Department at Education Evaluation Center in Mongolia. He earned a B.Sc from National University of Mongolia and M.Sc from Hiroshima University, Japan. His research interests include educational assessment and development of higher-order thinking skills through curriculum. He can be contacted at email: gendenjamts01@gmail.com

**References**

Anderson, L. W. (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. London: Longman.

Borg, W. R., & Gall, M. D. (1983). Educational Research: An Introduction. New York: Longman.

Güler, N. (2014). Analysis of open-ended statistics questions with Many Facet Rasch Model. Eurasian Journal of Educational Research, 55, 73-90. https://doi.org/10.14689/ejer.2014.55.5.

Haladyna, T. M. (1997). Writing Test Items to Evaluate Higher Order Thinking. Oakland, CA: Pearson.

Lawson, A. E. (1993). At what levels of education is the teaching of thinking effective? Theory Into Practice, 32(3), 170-178, DOI: 10.1080/00405849309543593

MECSS & MIER. (2019). Education in Mongolia, A Country Report (pp. 7-8).

Nitko, A. J., & Brookhart, S. M. 2007. How to Assess Higher-Order Thinking Skills in Your Classroom http://mpi.uinsgd.ac.id/wp-content/uploads/2018/07/

Nunnally, J. C. (1967). Psychometric Theory. New York: McGraw-Hill (pp. 172-235).

Paul, R., & Nosich, R. (1992). A model for the national assessment of higher-order thinking. (ERIC Document Reproduction Service No. ED 353 296).

OECD, (2015), Skills for Social Progress: The Power of Social and Emotional Skills, OECD Skills Studies, OECD Publishing, Paris, https://doi.org/10.1787/9789264226159-en.

Ontario. (2016). 21st-Century Competencies. Retrieved 17 April 2021, from http://www.edugains.ca/resources21CL/21stCenturyLearning/21CL_21stCentuyCompetencies.pdf

Otgonbaatar, K. (2020). Examining Mathematical Creativity Among Mongolian Ninth-Grade Students Using Problem-Posing Approach. Journal of Education and Practice, 11(27), 69-75. https://doi.org/10.7176/jep/11-27-08

Otgonbaatar, K. (2021a). Effectiveness of anchoring vignettes in re-evaluating self-rated social and emotional skills in mathematics. International Journal of Evaluation and Research in Education (IJERE), 10(1), 237. https://doi.org/10.11591/ijere.v10i1.20716

Otgonbaatar, K. (2021b). The development of a theoretical framework and tools to measure social and emotional skills in mathematics in the Mongolian lower secondary education (Doctoral dissertation, 広島大学).

Partnership for 21st Century Skills. (2009). "P21 framework definitions", Retrieved 17 April 2020, from https://eric.ed.gov/?id=ED519462

Scully, Darina. (2017). Constructing multiple-choice items to measure higher-order thinking. Practical Assessment, Research, and Evaluation, 22, 4.

Shellens, T., & Valcke, M. (2005). Collaborative learning in asynchronous discussion groups: What about the impact on the cognitive process? Computers in Human Behavior, 21(6), 957-975.

Tanujaya, B., Mumu, J., & Margono, G. (2017). The relationship between higher order thinking skills and academic performance of students in mathematics instruction.

TIMSS & PIRLS. (2011). Study report.

UNESCO. (2015). Transversal competencies in education Policy and Practice. (S. Strandberg, Ed.). Paris: UNESCO. Retrieved from http://unesdoc.unesco.org/images/0023/002319/231907E.pdf.

Оттонбаатар, Х. (2016). Даалгаврын хариултын онол (Item response theory)-ын 3 параметртэй загварыг ашиглаж тестийн даалгаврууд шинжилгээ хийх нь. *Proceedings of the Mongolian Academy of Sciences*, *56*(2), 24–30. https://doi.org/10.5564/pmas.v56i2.707