Evaluation of a Higher Order Thinking Skills Test for the Topic of Cell Division and Subtopic of Gametogenesis (UKBATG)

Wan Nasriha Wan Mohamed Salleh
Che Nidzam Che Ahmad
Norhamidah Hussain
Endang Setyaningsih
Saidatul Ainoor Shaharim
Rohani Hashim

DOI: https://doi.org/10.37178/ca-c.23.1.192

Wan Nasriha Wan Mohamed Salleh, *Department of Biology, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, Malaysia*

Che Nidzam Che Ahmad, *Department of Biology, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, Malaysia*
Email: nidzam@fsmt.upsi.edu.my

Norhamidah Hussain, *Department of Biology, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, Malaysia*

Endang Setyaningsih, *Biology Education Department, Faculty of Teacher Training, University of Muhammadiyah Surakarta, Indonesia*

Saidatul Ainoor Shaharim, *School of Educational Studies, Universiti Sains Malaysia, 11800 Georgetown, Pulau Pinang, Malaysia*

Rohani Hashim, *Department of Biology, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, Malaysia*

Abstract

This quantitative study aimed to evaluate the quality of a Higher Order Thinking Skills (HOTS) Test instrument for the topic of Cell Division and the subtopic of Gametogenesis (UKBATG). UKBATG contains contextual question and problem-solving question which is 25 items in the form of multiple-choice and 6 subjective items that can measure students' Higher Order Thinking Skills (HOTS). The process of validity was conducted by five experts in the field of Biology and HOTS and two experts in the field of Malay language and English language. The reliability of UKBATG was determined by calculating the KR-20 coefficient, Cronbach's alpha coefficient, and inter-rater reliability (IRR) by determining the intra-class correlation coefficient (ICC). The quality of UKBATG was also determined by calculating the difficulty index, p and discrimination index, D for each item. Findings showed that all UKBATG items have a good validity exceeding 70% agreement among experts. In terms of reliability, it was found that the KR-20 coefficient was 0.774, with Cronbach's alpha of 0.844 and the ICC coefficient in the range of 0.635 to 0.841. The values of p and D indicated that the

UKBATG items were at a moderate level and were accepted as good items. In conclusion, this study successfully developed an instrument that is good in terms of validity and reliability and also has good item quality. The implication is that the UKBATG can be used to measure students' HOTS and increase the number of HOTS test instruments, especially in Biology.

Keywords: *Validity, reliability, difficulty index, discrimination index, biology.*

INTRODUCTION

The Malaysia Education Blueprint 2013-2025 [1] changes the landscape of the education system in Malaysia to ensure the effectiveness of the Malaysian education system and further improve the quality of education to be on par with international standards. The main goal is to put Malaysia in the top three in terms of performance based on the international assessment of Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA). The non-maximisation and instability of achievement in TIMSS and PISA, according to the Organisation for Economic Cooperation and Development (OECD) [2], is caused by the unresolved problem-solving questions in the real-world context by students as students still lack higher-order thinking skills, creativity, and innovation, which are the characters of 21st-century skills. In the next three years, based on data from the OECD [3], the performance results based on the TIMSS and PISA assessments would still be low. Therefore, to achieve these aspirations and results, students need to master HOTS, which is a component in increasing student performance scores from an early age . The emphasis on thinking skills in all disciplines directly makes thinking skills, especially Higher Order Thinking Skills (HOTS), more significant in today's education system. This effort is an initiative that supports transformation efforts to produce students who master thinking skills, that is, cognitive skills including reasoning and critical, creative and, innovative thinking.

Higher-order thinking skills are an emerging trend in education and one of the cognitive abilities of students that can be trained and developed in 21st-century learning. Students HOTS must be habituated and trained with higher-order thinking test items by giving questions in the form of problem-solving, creative thinking, critical thinking and metacognitive abilities [4-7]. This HOTS assessment can encourage students to think broadly and deeply about a current problem in their real life and situations related to learning materials.

Therefore, assessment instruments based on High Order Thinking Skills (HOTS) principles are required to measure and improve students' HOTS achievement by providing HOTS questions that require students to use their thinking skills to respond to those questions. These HOTS questions can be applied as practice questions, quizzes, test questions, and examinations. The emphasis on these HOTS questions is also in line with the goal of science education in Malaysia, which is to produce students who can apply scientific knowledge in decision making and problem-solving in life [8]. As such, one way to increase the focus on HOTS and prepare students for a higher level of thinking that is in line with the 21st-century demands the development of an instrument capable of measuring students' HOTS performance by providing quality HOTS questions. A HOTS item requires the ability to apply higher-order thinking as the item is presented using a stimulus from daily-life real problems [4]. Through HOTS-based test items, creative and critical thinking skills can be built through practices in this problem-solving.

OBJECTIVE

The purpose of this study is to evaluate a test instrument that has satisfactory validity and reliability in addition to obtaining the quality level of test items through item analysis. This test instrument was developed to measure students' HOTS for Biology subjects for the topic of Cell Division and subtopics of Gametogenesis covering the four highest cognitive level skills based on the revised Bloom's Taxonomy by Anderson and Krathwohl [9] namely skills of apply, analyse, evaluate, and create. Therefore, test instruments that have validity, reliability, and quality need to be developed so that students' HOTS performance can be measured effectively and continuously.

The objectives of this study are:

1. To measure the validity value of the Higher Order Thinking Skills Test for the topic of Cell Division and the subtopic of Gametogenesis (UKBATG).
2. To measure the reliability value of UKBATG.
3. To determine the quality of the item based on the value of the item difficulty index and the item discrimination index.

PROBLEM STATEMENT

The mastery of HOTS by students needs to be given due attention. However, the performance of student achievement in tests in the form of HOTS implemented by international institutions such as TIMS and PISA or through research that has been conducted indicates that Malaysia and Indonesia student achievement is at a level below satisfactory. Findings from the results of an international study, TIMSS and PISA, found that the thinking skills of students are still at a low level [3]. In general, the ability of students is deficient in understanding complex information, theory, analysis, and problem solving, use of tools procedures and problem-solving, as well as conducting investigations. In addition, the findings of HOTS studies on students are generally not commendable. The higher order thinking skills levels among the students were at low level [10-17]. Students in higher education and secondary education in science learning are still lagging in terms of the use of Higher Order Thinking Skills (HOTS) [15, 18-24]

The strategy for applying knowledge such as problem-solving was least introduced by science teachers[15, 25]. Students are given less exposure related to problem-solving issues of daily life, especially through the HOTS test. Sometimes, the tests given to students are narrative statements that are unable to stimulate students' high-order thinking skills. The question items given are directed more to lower-order thinking skills of memorisation and understanding the basic concepts in science rather than higher-order thinking skills [25-29] due to lack of ability in developing instrument assessment towards HOTS by teachers [30]. This proves that HOTS has been poorly trained and accommodated to the students, especially in HOTS-based tests. This situation causes students to have difficulty in relating information and implementation of strategies to solve HOTS test items and cause students' low interest in HOTS test items [13, 31].

Besides, the format of public examination questions in Malaysia such as *Sijil Pelajaran Malaysia* (SPM) is simple and more focused on diagrams and tables that do not require high-order thinking skills to give answers [32]. This indicates the structure of SPM questions does not encourage students to think more critically[32] . The SPM biology questions cover all Form 4 and 5 Biology topics which measure low level and high-level cognitive domains according to the percentage determined by the Malaysia Ministry of Education. Figure 1 shows the increase in the percentage of HOTS questions in SPM examination papers from 2014 until 2021.

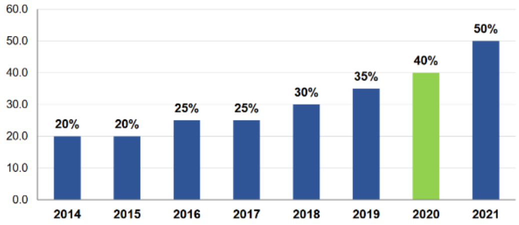


Figure 1 Percentage of HOTS questions in SPM examination papers
Source: Kementerian Pendidikan Malaysia (2021)

In addition, the instrument to test HOTS in the existing market is not accurate to measure HOTS because it does not go through a systematic development process, not based on the characteristics of HOTS items proposed by MOE and not through the process of validity, reliability and item analysis. This is because a measurement tool needs to be developed based on local context supported by content experts[33]. This shows that lack of instruments that can be used to measure students' HOTS performance [34].

To address issues and challenges, an instrument that can train students' higher order thinking and measure HOTS by standards HOTS test should be developed. The HOTS test is an instrument that can be used for nurturing, strengthening, improving, and measuring skills [35]. The development of HOTS test contributes to the number of existing HOTS instruments that can be used by teachers to train, assess and measure the current performance of their students' HOTS, especially for Biology subject. The development of HOT questions can also reduce the burden on teachers to provide quality HOTS questions in Biology learning. This is an added value and contribution of researchers to the teachers, schools and also to the government.

Therefore, the Higher Order Thinking Skills Test was developed to properly test students' HOTS by providing 100% high-level cognitive domain questions for the topics of Cell Division and subtopic of Gametogenesis namely UKBATG. These topics has been difficult topic for students since a long time ago [36-42] because of many abstract concepts[43, 44]. Students' conception about these topics is often not in line with scientific concepts [45] because these concepts are difficult to understand and various terms are difficult to describe. Thus, it is important for students to master HOTS in Biology learning because these skills are essential to understand abstract biological concepts and be able to solve many biological questions. Students who are trained and familiar with HOTS items will show high performance in learning and future career.

The UKBATG instrument developed based on the HOTS items criteria as suggested by the Malaysia Ministry of Education and based on the Bloom Taxonomy of cognitive domain. The UKBATG instruments are also evaluated in terms of validity, reliability and item quality through the values of the difficulty index and discrimination index. The high quality of UKBATG instrument enabling it to measure student HOTS in topic of Cell Division and Subtopic Gametogenesis. The UKBATG instrument can also be used by Biology teacher in both of evaluating the process and as guiding in forming the test level of HOTs.

METHODOLOGY

This quantitative study was conducted to obtain the validity, reliability and item analysis for UKBATG. A total of five experts were involved in the UKBATG validation process. Two experts were involved in the verification language UKBATG comprising a Malay language expert and an expert in English. A total of 37 students who took Biology as an elective subject were involved in this study to obtain the value of UKBATG reliability in addition to measuring the value of difficulty index, p and discrimination index, D for UKBATG items.

Validity of UKBATG

UKBATG validity was assessed through face validity and content validity. The determination of UKBATG validity was determined by calculating the percentage of expert or respondent agreement[46]stated that an agreement percentage equal to or more than 70% indicates accepted consent. Determination of face validity and content validity was done using the UKBATG Instrument Validity Form constructed by the researcher based on the study highlights and guidelines for the construction of HOTS items by the Malaysian Examinations Board, Ministry of Education. The Instrument Validity Form consisted of two parts, namely Part A - Face Validity and Part B - Content Validity. According to [46], a test is said to have face validity when the set of tests presented appears to measure what is to be measured. The face validity of a test describes the extent to which the test appears relevant, important, and interesting. Content validity is the most important process in the development of an achievement test[47].

Content validity is the most important process in the development of an achievement test[48, 49]. Content validity is an assessment of the content of a test to determine the meaning of the test scores and whether or not the domain of behaviour being measured represents the entire content of a domain [49]. This means that the test items constructed represent the items to be tested or measured. In the context of this study, UKBATG items are said to have content validity if they have four characteristics as suggested by the Malaysian Examinations Board, namely conformity to the content domain (Cell Division and Gametogenesis), meet the learning and teaching objectives (content standards and learning standards) as stated in Curriculum and Assessment Standard Documents of Biology, conform to the domain cognitive, which is an item that tests the skills of apply, analyse, evaluate and create and finally possesses the characteristics of HOTS items, which include extensive stimuli, various levels of thinking, unusual context, real situations in everyday life and items that are not recurring. Content validity can be improved by evaluating test items based on the Test Specification Table review by a panel of reference experts who review the items and comment on whether the items have covered all the content to be tested[50]. The checking of the suitability of items was done according to the method proposed by [51], which is to use the code 1 = suitable item, 2 = doubtful and 3 = inappropriate. The characteristics of HOTS items in UKBATG are summarised in Figure 2.

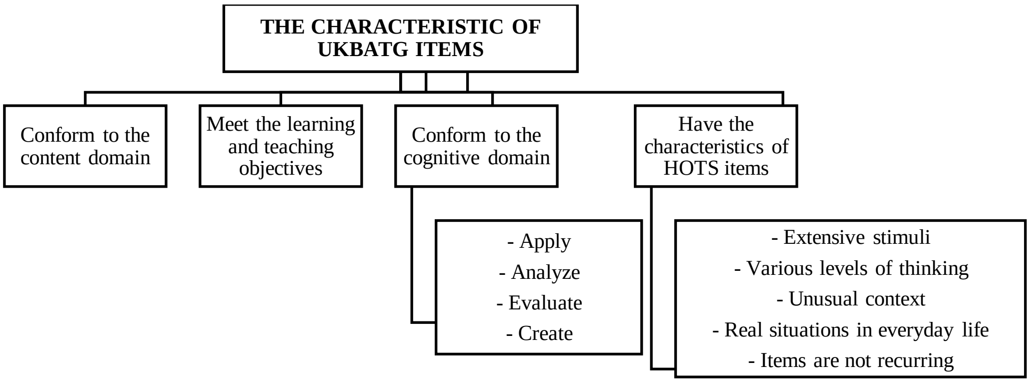

THE CHARACTERISTIC OF UKBATG ITEMS
Conform to the content domain
Meet the learning and teaching objectives
Conform to the cognitive domain
Have the characteristics of HOTS items
- Apply
- Analyze
- Evaluate
- Create
- Extensive stimuli
- Various levels of thinking
- Unusual context
- Real situations in everyday life
- Items are not recurring

Figure 2 The characteristics of HOTS items in UKBATG

Reliability of UKBATG

[52]suggested that any classroom tests must always undergo reliability testing. The reliability of UKBATG was determined by calculating the KR-20 coefficient, Cronbach's alpha coefficient, and inter-rater reliability (IRR) by determining the intra-class correlation coefficient (ICC). KR-20 coefficient provides relatively conservative estimates of the coefficient of equivalence. According to [53], high test reliability refers to the consistency, accuracy and precision of the measurements made. The Cronbach's alpha coefficient is the determination of reliability using a single test against a group of students and subjectively test items that include the total score of each item that is not dichotomous. According to[54], Cronbach's alpha coefficient calculation method is important for tests that have non-uniform scores including subjective items. By calculating Cronbach's alpha coefficient, the internal consistency and uniformity of each item in the instrument can be estimated. Cronbach's alpha compares the variance of each item to total test variance.

[54] suggested that value of the KR20 coefficient in the range of .70 and .80 are sufficient to represent the reliability of the test questions. [55] argued that the reliability coefficient is usually in the range of 0.5 to 0.70 and a value of 0.4 is considered low. Whereas for teacher-constructed tests for classroom testing, a Cronbach's alpha coefficient value of at least 0.70 is adequate [56-58]. Cronbach's alpha reliability coefficient provided by Cohen et al. [60] were used in this research shown by Table 1.

Table 1

Cronbach's alpha reliability coefficient

Coefficient	Description
>0.90	Very highly reliable
0.80 – 0.90	Highly reliable
0.70 – 0.79	Reliable
0.60 – 0.69	Marginally/minimally reliable
<0.60	Unacceptably low reliability

According to [58], inter-rater reliability (IRR) is particularly important if subjective items are evaluated by two or more examiners. Williams et al. [59], suggested that at least three examiners are required to examine each student's answer sheet. This way, it is possible to avoid examiner bias, reduce measurement errors and improve instrument reliability. [41]added that through the IRR method, any bias by any examiner can be eliminated. Therefore, three examiners among the Paper 2 examiners for the SPM Biology subject were appointed to check the UKBATG answers of the students. The definitions of the intra-class correlation coefficient (ICC) by Hallgren, [62] shown by Table 2.

Table 2

The definitions of the intra-class correlation coefficient (ICC)

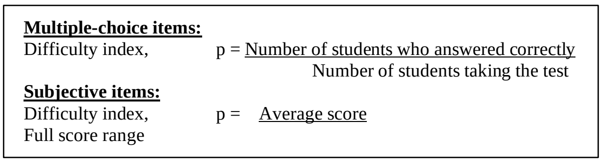
Coefficient	Description
1.00	Excel
.75 to .99	Very good
.60 to .74	Good
.40 to .59	Satisfactory
<.40	Less satisfactory
0	Random

Item analysis

Item analysis is the process of analysing all test items that have been statistically formulated. Two important analyses for the items were the determination of the difficulty index and the discrimination index. By calculating the difficulty index and the discrimination index, the quality of each item can be determined empirically [60]. Item analysis is performed aiming to help the item drafter to refine the test whether by storing, using directly, purifying, or getting rid of an item. Items with easy and difficult difficulty levels as well as items with low discriminatory power will be refined, modified, or removed [61]. The difficulty index, p and discrimination index, D for each item were calculated using Microsoft Excel software.

a. Item difficulty Index

The item difficulty index is defined as the percentage or ratio of students who answered correctly out of the total number of students who answered the item. The difficulty index is an indicator of the difficulty of an item. The difficulty index is calculated using different formulas according to the type of item, multiple-choice items and subjective items. According to[61], the formulas are respectively as in Figure 2.



Multiple-choice items:

Difficulty index, p = Number of students who answered correctly / Number of students taking the test

Subjective items:

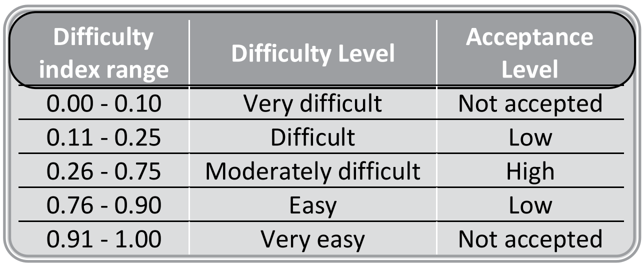
Difficulty index, p = Average score
Full score range

Figure 2 Difficulty index, p calculation formula for multiple-choice items and subjective items

Interpretation of the value of p obtained is a method to find out the difficulty of test items administered to students. The more difficult an item is, the fewer students give the correct answer. Statistically, p has a value between 0.0 and 1.0. The larger the value of p, the easier the item is and the smaller the value of p, the harder the item[62]. For this study, the interpretation by [63] was used as in Table 3. For a good level of difficulty, [63]recommend choosing items that are in the medium level, which is in the middle between difficult and easy levels. (0.26 p ≤ 0.75).

Table 3

Item Difficulty Level Interpretation

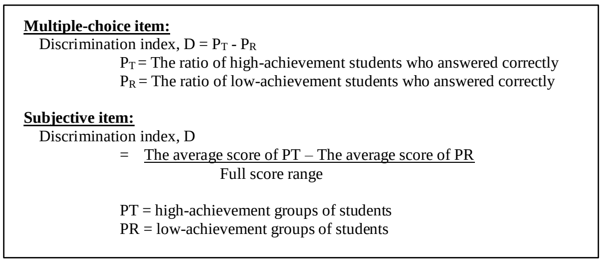


Difficulty index range	Difficulty Level	Acceptance Level
0.00 - 0.10	Very difficult	Not accepted
0.11 - 0.25	Difficult	Low
0.26 - 0.75	Moderately difficult	High
0.76 - 0.90	Easy	Low
0.91 - 1.00	Very easy	Not accepted

b. Item discrimination index

An important feature of the next test instrument is that it has discriminatory power [64]. The discrimination index is symbolised by D, which is used to differentiate achievement between the group of high-achievement students and the group of low-achievement students. A good item is an item that can differentiate between high-achievement students and low-achievement students. The selection procedure for high-achievement and low-achievement student groups started by sorting the test score data from the highest score to the lowest score. Then, the group of high-achievement students and the group of low-achievement students were determined. There are various opinions about the breakdown or percentage of the selection of the group of students. For example, Reynolds, [64]suggested selecting 27% of all students who sat the test to high-achievement students and low-achievement students and leave 46% of students between the two groups. Measurement experts agreed that the best prediction of discrimination index is obtained when involving 27% of the low-achievement group of students and 27% of the high-achievement group of students [69]. The value of 27% is a sufficient value to be analysed in a normal distribution [70]. According to [62], any value between 25% to 33% is adequate. For this study, 27% of 37 students were taken, which is equivalent to 10 people as a group of high-achievement students and 10 people as a group of low-achievement students. This is because this percentage value can maximise the difference in the normal distribution to provide sufficient cases for analysis.

The discrimination index, D was calculated using different formulas according to the types of items, namely multiple-choice items and subjective items. The value of D is between -1.00 and +1.00. The value of D = +1.00 means that all high-achievement students gave the correct answer and all low-achievement students gave the wrong answer. The higher the D value, the more groups of high-achievement students answered correctly for an item. This suggests that the item discriminates against the group of high-achievement students who must answer more items correctly compared to the group of low-achievement students. While a negative D value for an item indicates that the score of low-achievement students is higher than the score of high-achievement students, this item fails to differentiate the group of students at the same time; thus, it is appropriate for such items to be removed. According to Reynolds et al. [59], the item discrimination index calculation formula for multiple-choice items and subjective items are as respectively shown in Figure 3.



Multiple-choice item:

Discrimination index, D = $P_T - P_R$

P_T = The ratio of high-achievement students who answered correctly

P_R = The ratio of low-achievement students who answered correctly

Subjective item:

Discrimination index, D

$$= \frac{\text{The average score of PT – The average score of PR}}{\text{Full score range}}$$

PT = high-achievement groups of students

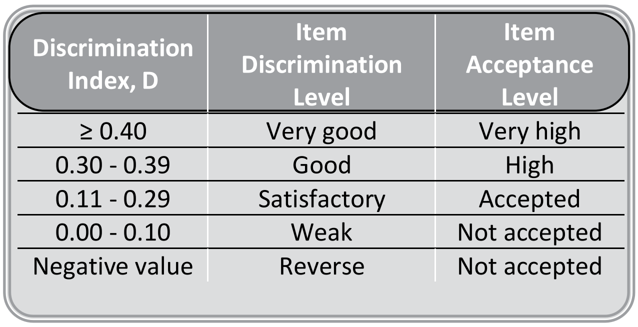
PR = low-achievement groups of students

Figure 1 Difficulty index, D calculation formula for multiple-choice items and subjective items

According to [65], the interpretation of discrimination index, D values is as shown in Table 4.

Table 4

Item Analysis Guide based on Item Discrimination



Discrimination Index, D	Item Discrimination Level	Item Acceptance Level
≥ 0.40	Very good	Very high
0.30 - 0.39	Good	High
0.11 - 0.29	Satisfactory	Accepted
0.00 - 0.10	Weak	Not accepted
Negative value	Reverse	Not accepted

RESULT AND DISCUSSION

The results of the analysis of the findings are discussed based on the objectives of the study stated.

Objective 1: To measure the validity value of Higher Order Thinking Skills Test for the topic of Cell Division and the subtopic of Gametogenesis (UKBATG).

Validity is the most important process in the construction of a test to determine its quality and relevance [66]. The draft UKBATG and scoring scheme have gone through a validation process by five experts who have performed face validation and content validation. These experts are experienced in the field of Biology Education and HOTS. The experts are also directly involved with the drafting of the Biology Curriculum and Assessment Standard Document and the drafting of Biology questions for the Selangor state level Malaysian Certificate of Education. Table 5 shows the profiles of the experts involved in the validation of the UKBATG instrument.

The results of this study found that the overall average for the percentage of expert agreement of face validity by the expert was 97.78%. The percentage of expert agreement in each item of face validity by the expert was ranged from 80% to 100% and exceeded the minimum value of 70%. All (100%) experts agreed for each item submitted except for Item 10 and Item 18 where 80% of the experts agreed there were no spelling errors and the answer space corresponded to the expected response.

Therefore, the aspects of spelling and answer space provided need to be improved by the researcher. This study concluded that the UKBATG instrument was able to measure what should be measured and that the UKBATG instrument has good face validity. Table 6 shows the findings for face validity for the UKBATG.

Table 5

Experts' Profile

No	Expert	Position/ Grade	Academic Qualifications	Years of Experience	Areas of Expertise/ Contributions
1	A	DG54	Degree	> 20 years	Biology Education, Excellent Biology Teacher, Paper 2 of Biology Examiner, Drafter of DSKP Biology Form 4 KSSM
2	B	DG44	Degree	11 – 15 years	Biology Education, Paper 2 of Biology Examiner
3	C	DG52	PhD	> 20 years	Biology Education, Assessment and Measurement, HOTS
4	D	DG44	Degree	11 – 15 years	Biology Education, Paper 2 of Biology Examiner, Selangor Biology Question Drafter
5	E	DG48	Masters	16 – 20 years	Biology Education, Paper 2 of Biology Examiner, Selangor Biology Question Drafter

Table 6

Face Validity of UKBATG (test questions)

No.	Statement	Agreement			
		Yes		No	
		Frequency	Percentage (%)	Frequency	Percentage (%)
1	The format corresponds to the skill being measured	5	100	0	0
2	The instructions given are clear	5	100	0	0
3	The meaning of each verse is clear	5	100	0	0
4	The language used is easy to understand	5	100	0	0
5	The language used suitably with the user level	5	100	0	0
6	The terms used are appropriate	5	100	0	0
7	Consistent use of the term	5	100	0	0
8	Adjust the biology/science used accordingly	5	100	0	0
9	The grammar used is appropriate according to the level of the user	5	100	0	0
10	There are no spelling mistakes	4	80	1	20
11	The punctuation used is correct	5	100	0	0
12	Appropriate font size	5	100	0	0
13	The font size is easy to read	5	100	0	0
14	The text used is clear	5	100	0	0
15	The text used is easy to read.	5	100	0	0
16	The order of the sentences is appropriate	5	100	0	0

17	The paragraph structure is appropriate	5	100	0	0
18	The answer space corresponds to the expected response	4	80	1	20
	Average		97.78		2.22

The results of the expert review of UKBATG found that the percentage of expert agreement on conformity to the content domain was 100% for all Part A questions (multiple-choice) and Part B questions (subjective) except items 5 d (ii), 5 d (iii) and 6 f. The percentage of agreement of items 5 d (ii), 5 d (iii) was above 70% and only Item 6 f displayed 60%. However, the results of the researcher's review on Curriculum and Assessment Standard Documents of Biology in KSSM [72] clearly showed that the item was included in the content of Curriculum and Assessment Standard Documents of Biology for the topic of Cell Division and Gametogenesis. This indicates that the UKBATG items complied with the content domain, that is, they all revolve around a predetermined topic.

The percentage of expert agreement on conformity to the exact objectives of teaching and learning was 100% for all questions in Section A (multiple-choice) except Item 6 and all items of questions of Section B (subjective). However, the percentage of agreement of Item 6 has exceeded 70%. This indicates that each UKBATG item has met the objectives of teaching and learning as expected in the DSKP. The percentage of expert agreement on conformity to cognitive domains testing the skills of apply, analyse, evaluate and create for all Part A questions (multiple-choice) except items 1 and 17 and all Part B question items (subjective) was over 80%. Only two items on the Part A questions (objective) namely Item 1 and Item 17 showed 60% agreement. These items were refined according to the suitability of a predefined cognitive domain.

The last aspect assessed by the experts was whether or not the UKBATG item met the characteristics of the HOTS item. The five characteristics of HOTS items according to the Malaysian Examinations Board [36] are having extensive stimuli, various levels of thinking, unusual contexts, non-repetitive items and real situations in daily life. The results of the expert review found some items in the opinion of experts that have not met the characteristics of HOTS items. There were three items, namely Item 1, Item 14, and Item 22 in the questions of Section A (multiple-choice), which only achieved a 60% agreement percentage; less than the minimum value of 70%. The refinement of these items was done considering all the views of experts. In conclusion, all items on Question Part B (multiple-choice) except Item 23 and all items on question Part B (subjective) were accepted by the experts. Item 23 was refined according to expert recommendations. All items that did not reach the 70% agreement level were refined to ensure that all UKBATG items meet the characteristics of HOTS items as outlined by the Malaysian Examination Board [36].

In conclusion, the validity of the UKBATG was indicated by the experts' judgment showing that the UKBATG is suitable to be used in the aspects of contents, format, and language. Table 7 and Table 8 display the findings for content validity of the UKBATG questions in Part A (multiple-choice) and Part B (subjective).

Table 7

Content Validity of UKBATG (Part A: Multiple-choice Question)

Item	Content Validity														
	Conform to the content domain			Meet the learning and teaching objectives			Conform to the domain cognitive			Has the characteristics of HOTS items			Item Determination (Accepted)		
	Freq (f)	Percentage (%)	Level of acceptance	Freq (f)	Percentage (%)	Level of acceptance	Freq (f)	Percentage (%)	Level of acceptance	Freq (f)	Percentage (%)	Level of acceptance	Freq (f)	Percentage (%)	Level of acceptance
1	5	100	Accepted	5	100	Accepted	3	60	Accepted	3	60	Accepted	4	80	Accepted
2	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted
3	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted	5	100	Accepted
4	5	100	Accepted	5	100	Accepted	4	80	Accepted	4	80	Accepted	4	80	Accepted
5	5	100	Accepted	5	100	Accepted	4	80	Accepted	4	80	Accepted	5	100	Accepted
6	5	100	Accepted	4	80	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted
7	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted	4	80	Accepted
8	5	100	Accepted	5	100	Accepted	4	80	Accepted	4	80	Accepted	4	80	Accepted
9	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted	5	100	Accepted
10	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted	5	100	Accepted
11	5	100	Accepted	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted
12	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted
13	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted
14	5	100	Accepted	5	100	Accepted	4	80	Accepted	3	60	Accepted	4	80	Accepted
15	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted	5	100	Accepted
16	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted	4	80	Accepted
17	5	100	Accepted	5	100	Accepted	3	60	Accepted	4	80	Accepted	4	80	Accepted
18	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted
19	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted
20	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted
21	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted
22	5	100	Accepted	5	100	Accepted	5	100	Accepted	3	60	Accepted	3	60	Accepted
23	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted	5	100	Accepted
24	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted
25	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted

Table 8

Content Validity of UKBATG (Part B: Subjective Question)

Item	Content Validity														
	Conform to the content domain			Meet the learning and teaching objectives			Conform to the domain cognitive			Has the characteristics of HOTS items			Item Determination (Accepted)		
	Freq (f)	Percentage (%)	Level of acceptance	Freq (f)	Percentage (%)	Level of acceptance	Freq (f)	Percentage (%)	Level of acceptance	Freq (f)	Percentage (%)	Level of acceptance	Freq (f)	Percentage (%)	Level of acceptance
1															
a	5	100	Accepted	5	100	Accepted	4	80	Accepted	4	80	Accepted	4	80	Accepted
b	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
c i	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted			
c ii	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
d	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted			
e	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
f	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
2															
a i	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted	4	80	Accepted
a ii	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted			
a iii	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
b	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
3															
a	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted	4	80	Accepted
b i	5	100	Accepted	5	100	Accepted	4	80	Accepted	4	80	Accepted			
b ii	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
c	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
d	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted			
4															
a i	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted
a ii	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
b	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted			
c	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
d	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
e	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
f	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			

5															
a	5	100	Accepted	5	100	Accepted	4	80	Accepted	4	80	Accepted	5	100	Accepted
b i	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
b ii	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
c	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
d i	5	100	Accepted	5	100	Accepted	5	100	Accepted	4	80	Accepted			
d ii	4	80	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
d iii	4	80	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
d iv	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
6															
a	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted
b	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted			
c i	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
c ii	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
d	5	100	Accepted	5	100	Accepted	5	100	Accepted	4	80	Accepted			
e	5	100	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
f	3	60	Accepted	5	100	Accepted	5	100	Accepted	5	100	Accepted			
g	5	100	Accepted	5	100	Accepted	4	80	Accepted	5	100	Accepted			

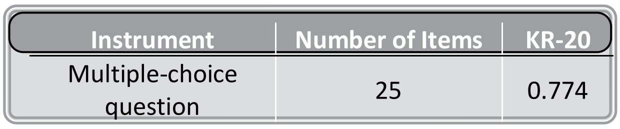
Objective 2: To measure the reliability value of UKBATG.

To ensure that these items comply with the characteristics of HOTS items, repetition of the same item to the same respondent was not allowed. This is because the item tested on a respondent must be a new item [36]. Therefore, the determination of the degree of reliability of the UKBATG instrument was conducted in one session to the same respondents. The reliability of the UKBATG instrument was determined using two methods, namely by calculating the KR-20 value for multiple-choice items and Cronbach's alpha coefficient to measure the internal consistency of subjective items in the HOTS instrument as well as determining inter-rater reliability (IRR) by finding the intra-class correlation coefficient (ICC) for each subjective item.

The reliability of multiple-choice items of a dichotomous nature can be calculated using the Kuder-Richardson 20 (KR-20) formulas. Kuder-Richardson 20 (KR-20) analysis was used as the difficulty level of the test items was not homogeneous and differ in level according to Bloom's taxonomic level. From the KR-20 analysis, it was found that the result was r11 = 0.774. This indicates that the part A question item (multiple-choice) has a high-reliability value. The value of KR-20 is shown in Table 9.

Table 9

Reliability of multiple-choice questions.

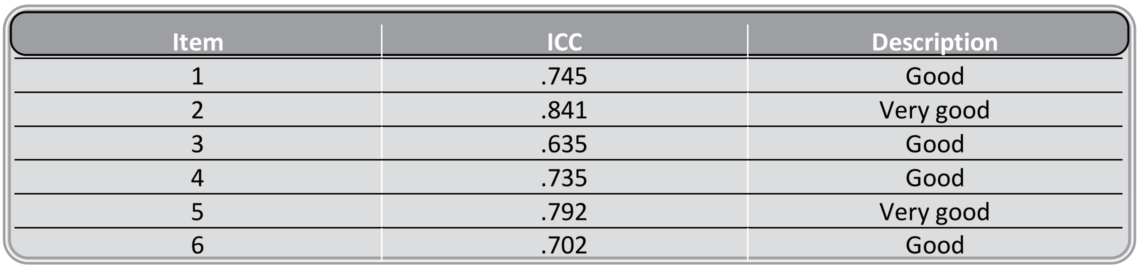


Instrument	Number of Items	KR-20
Multiple-choice question	25	0.774

The reliability of subjective items is shown by Cronbach's alpha value and ICC coefficient value. The value of Cronbach's alpha coefficient for the entire subjective item was 0.844. This indicates that subjective item has a high-reliability value [60] and this value is in the accepted range above 0.70 [48, 59]. Meanwhile, the reliability between examiners demonstrated that the value of the ICC coefficient for all items was in the range of 0.635 to 0.841. Items 2 and 5 were at a very good level, while items 1, 3, 4 and 6 were at a good level. This value gives the impression that there is only a very small error between the examiner and the score obtained by the student indicating the actual level of HOTS tested. Table 10 shows the values of the ICC coefficients for each item and the overall scores.

Table 10

ICC Reliability Index Values for Each Item and UKBATG Score



Item	ICC	Description
1	.745	Good
2	.841	Very good
3	.635	Good
4	.735	Good
5	.792	Very good
6	.702	Good

These findings indicate that the UKBATG instrument has fulfilled the requirement for reliability. The UKBATG items were consistent and have uniformity in measuring the students'. According to [41, 59], high internal consistency values describe UKBATG items as homogeneous that measure the same domain. This also proves that the reliability of the UKBATG instrument is good as the UKBATG was able to measure students' HOTS consistently.

Objective 3: To determine the quality of the item based on the value of the item difficulty index and the item discrimination index

The results of item analysis showed an average value of difficulty index, p = 0.39 for multiple-choice items and 0.35 for subjective items. This means that on average, UKBATG is at a moderately difficult level (0.26 ≤ p ≤ 0.75), thus indicating that the level of item acceptance is high. For multiple-choice items, Item 11 (p = 0.76) and Item 24 (p = 0.76) were at an easy level with a low level of item acceptance. Item 18 has a p-value of 0.05 since being on hold is very difficult as this item is at the level of analysis on Bloom's taxonomy resulted in many students unable to answer correctly.

For subjective items, all items except Item 4 were categorised with moderately difficult, hence allowing these items to be accepted. Item 4 has a value of p = 0.25 meaning that this item has a difficult level[67] suggested that items with p values between 0.11 - 0.25 and 0.76 - 0.90 must be modified and retested before being used in actual testing. Accordingly, items 11, 18 and 24 for multiple-choice questions and Item 5 for subjective questions need to be refined. No items were removed. This is because according to [64] the process of drafting an item takes a long time, so it is best for items that are too difficult or easy to be purified and maintained. All items of Part A (multiple-choice) and items of Part B (subjective) UKBATG were analysed individually. Table 11 shows the difficulty index, p for each UKBATG item.

Table 11

Difficulty Index, p for Each UKBATG item

Part of question	Item Difficulty index (p)	Item Difficulty Level	Item Acceptance Level	Item number	Total
Part A (Multiple-choice)	0.00 - 0.10	Very difficult	Not accepted	18	1
	0.11 - 0.25	Difficult	Low	2, 7, 16, 17, 21,	5
	0.26 - 0.75	Moderately difficult	High	1, 3, 4, 5, 6, 8, 9, 10, 12, 13, 14, 15, 19, 20, 22, 23, 25	17
	0.76 - 0.90	Easy	Low	11, 24	2
	0.91 - 1.00	Very easy	Not accepted	-	-
	Total				25
Part B (Subjective)	0.00 - 0.10	Very difficult	Not accepted		
	0.11 - 0.25	Difficult	Low	4	1
	0.26 - 0.75	Moderately difficult	High	1, 2, 3, 5, 6	5
	0.76 - 0.90	Easy	Low		
	0.91 - 1.00	Very easy	Not accepted		
			Total		6

Table 11 shows that 17 items (68%) in multiple-choice questions have a moderately difficult level. Meanwhile, 5 items (83%) in subjective questions have a moderately difficult level. This indicates that UKBATG items have a high level of acceptance based on the difficulty level of the items.

The results of item analysis for this study showed the range of discrimination index, D items in the UKBATG instrument between 0.0 and 0.8. In detail, the items on the multiple-choice questions, namely items 2, 8, and 18 were at a good level of discrimination (0.30 ≤ D ≤ 0.39), while items 1, 3, 4, 6, 7, 9, 10, 12, 13, 14, 17, 19, 20, 21, 22 and 23 were at a very good level of discrimination (D ≥ 0.40). The level of good and very good item discrimination indicates that the items have a high and very high level of acceptance. This indicates that these items can distinguish the group of high-achieving students and the group of low-achieving students. However, items 11, 16, and 25 were at a satisfactory level of discrimination (0.11 ≤ D ≤ 0.29). This indicates that these items are still acceptable. For items 5, 15, and 24, the results of the analysis showed that the items were at a weak level of discrimination (0.00 ≤ D ≤ 0.10). This means that these items are less discriminatory between the high-achievement group and the low-achievement group.

For subjective question items, items 2, 3, 4, 5, and 6 were at a satisfactory level of discrimination (0.11 ≤ D ≤ 0.29) and the level of acceptance was acceptable. Lewis and Smith [73] stated that a value of D between 0.10 and 1.00 illustrates that the item is still applicable. For item 1, the value of D was 0.09, which is at a weak level. This means that this item is less discriminatory between the high achievement group and the low achievement group. Discrimination Index, D for each UKBATG item is shown in Table 12.

Table 12

Item Difficulty Index, p and Item Discrimination Index, D for Each Item of UKBATG

Part of question	Discrimination Index, D	Item Discrimination Level	Item Acceptance Level	Item number	Total
Part A (Multiple-choice)	≥ 0.40	Very good	Very high	1, 3, 4, 6, 7, 9, 10, 12, 13, 14, 17, 19, 20, 21, 22, 23	16
	0.30 - 0.39	Good	High	2, 8, 18	3
	0.11 - 0.29	Satisfactory	Accepted	11, 16, 25	3
	0.00 - 0.10	Weak	Not accepted	5, 15, 24	3
	Negative value	Reverse	Not accepted	-	
		Total			25
Part B (Subjective)	≥ 0.40	Very good	Very high		
	0.30 - 0.39	Good	High		
	0.11 - 0.29	Satisfactory	Accepted	2, 3, 4, 5, 6	5
	0.00 - 0.10	Weak	Not accepted	1	1
	Negative value	Reverse	Not accepted		
			Total		6

Table 12 presents those 16 items (64%) in multiple-choice questions have a very good discrimination level. Meanwhile, 5 items (83%) in subjective questions have a satisfactory discrimination level. This indicates that UKBATG items have an acceptable level of acceptance based on the discrimination level of the items. Overall, the values of the difficulty index and discrimination index for each UKBATG item are shown in Table 13.

Table 13

Item difficulty index and item discrimination index of UKBATG

Part of question	Item no.	Item Difficulty index, p	Item Difficulty Level	Item Acceptance Level	Discrimination Index, D	Item Discrimination Level	Item Acceptance Level
A	1	0.73	Moderate difficult	High	0.50	Very good	Very high
	2	0.16	Difficult	Low	0.30	Good	High
	3	0.43	Moderate difficult	High	0.40	Very good	Very high
	4	0.59	Moderate difficult	High	0.80	Very good	Very high
	5	0.32	Moderate difficult	High	0.00	Weak	Not accepted
	6	0.29	Moderate difficult	High	0.50	Very good	Very high
	7	0.24	Difficult	Low	0.50	Very good	Very high
	8	0.51	Moderate difficult	Low	0.30	Good	High
	9	0.54	Moderate difficult	Low	0.40	Very good	Very high
	10	0.27	Moderate difficult	Low	0.60	Very good	Very high
	11	0.76	Easy	Low	0.20	Satisfactory	Accepted
	12	0.38	Moderate difficult	High	0.50	Very good	Very high
	13	0.49	Moderate difficult	High	0.60	Very good	Very high
	14	0.29	Moderate difficult	High	0.40	Very good	Very high
	15	0.70	Moderate difficult	High	0.10	Weak	Not accepted
	16	0.16	Difficult	Low	0.20	Satisfactory	Accepted
	17	0.19	Difficult	Low	0.50	Very good	Very high
	18	0.05	Very difficult	Not accepted	0.30	Good	High
	19	0.29	Moderate difficult	High	0.50	Very good	Very high
	20	0.29	Moderate difficult	High	0.40	Very good	Very high
	21	0.19	Difficult	Low	0.50	Very good	Very high
	22	0.35	Moderate difficult	High	0.50	Very good	Very high
	23	0.38	Moderate difficult	High	0.70	Very good	Very high
	24	0.76	Easy	Low	0.10	Weak	Not accepted
	25	0.27	Moderate difficult	High	0.20	Satisfactory	Accepted
Average		0.39	Moderate difficult	High	0.40	Very good	Very high
B	1	0.31	Moderate difficult	High	0.09	Weak	Not accepted

	2	0.38	Moderate difficult	High	0.26	Satisfactory	Accepted
	3	0.41	Moderate difficult	High	0.18	Satisfactory	Accepted
	4	0.25	Difficult	Low	0.23	Satisfactory	Accepted
	5	0.34	Moderate difficult	High	0.23	Satisfactory	Accepted
	6	0.44	Moderate difficult	High	0.16	Satisfactory	Accepted
Average		0.35	Moderate difficult	High	0.19	Satisfactory	Accepted

Findings from the difficulty index and discrimination index of the UKBATG items are at a moderate difficulty level and have a good discrimination index. However, the researcher was aware that some items have a high difficulty index with discrimination index at a satisfactory and weak level. Unsatisfactory item analysis results may be due to learning and instruction problems implemented. Moreover, the topic of Cell Division is considered the most difficult. [33-35] Students involved in this test may not have mastered this topic, which involves Higher Order Thinking Skills. So, most high achievement students are unlikely to answer those items correctly while low achievement students may only answer those items correctly by chance. Thus, the items were retained but with revisions and improvements. These included the format of the writing, completeness of the stimulus texts, clearer pictures, and suitability with the Form 4 students' level as suggested by the experts. The researcher took steps to refine, improve these items and maintain them in the final instrument of UKBATG as suggested by [68]. [68]recommend that an item should not be removed but purified preferably. According to them, drafting an item involves a long process and the drafter should not waste an item that has been built but instead take appropriate purification measures. Getting rid of items is a detrimental thing. This is supported by [69]who says that purifying these items will improve the quality of a test.

CONCLUSION AND IMPLICATION

This study has successfully produced a test to measure HOTS, which has good validity and reliability for the topic of Cell Division and subtopic of Gametogenesis named UKBATG. In addition, the quality of each item in UKBATG was calculated based on item difficulty index, p and item discrimination index. Recommendation further research by developing HOTS tests on other Biology topics and other subjects to increase the number of tests that have good validity and reliability. In addition, further research is conducted by using the Rasch Model to evaluate and analysing the validity and reliability of test items as previously conducted by [74] as the more modern method. In conclusion, this instrument was feasible for users to identify the level of HOTS for the four sub-constructs of Higher Order Thinking Skills for the topics involved. In addition, the UKBATG instrument can also be used by teachers and students to assess the current performance for the topic of Cell Division and the subtopic of Gametogenesis.

ACKNOWLEDGEMENT

The authors would like to acknowledge the Scholarship and Financing Division, Ministry of Education Malaysia for the sponsorship of the "Hadiah Latihan Persekutuan" (HLP) that enabled this study to be implemented successfully. The authors also acknowledge the Department of Biology, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris (UPSI), Biology Education Department, Faculty of Teacher Training, University of Muhammadiyah Surakarta, Indonesia and School of Educational Studies, Universiti Sains Malaysia and those who extended their help in contributing to the completion of this manuscript.

REFERENCES

1. Don, Y., et al., *Educational leadership competencies and Malaysia education development plan 2013-2025.* Humanities and Social Sciences Review, 2015. **4**(3): p. 615-625.
2. Macinko, J., B. Starfield, and L. Shi, *The contribution of primary care systems to health outcomes within Organization for Economic Cooperation and Development (OECD) countries, 1970–1998.* Health services research, 2003. **38**(3): p. 831-865.DOI: https://doi.org/10.1111/1475-6773.00149.
3. Mensah, C.N., et al., *Technological innovation and green growth in the Organization for Economic Cooperation and Development economies.* Journal of Cleaner Production, 2019. **240**: p. 118204.DOI: https://doi.org/10.1016/j.jclepro.2019.118204.
4. Hamdi, S., I.A. Suganda, and N. Hayati, *Developing higher-order thinking skill (HOTS) test instrument using Lombok local cultures as contexts for junior secondary school mathematics.* REiD (Research and Evaluation in Education), 2018. **4**(2): p. 126-135.DOI: https://doi.org/10.21831/reid.v4i2.22089.
5. Adnan, M., et al., *Construction of the Form One Mathematics Higher Level Thinking Skills item for the topic of Fractions.* Malaysian Journal of Science and Mathematics Education, 2018. **8**(1): p. 46-54.DOI: https://doi.org/10.37134/jpsmm.vol8.1.4.2018.
6. Lestari, N.A., et al. *A Preliminary Study of Environmental Learning to Improve Students' Higher Order Thinking Skills in Physics*. IOP Publishing.DOI: https://doi.org/10.1088/1742-6596/1805/1/012033.
7. Raflee, S.S.M. and L. Halim, *The Effectiveness of Critical Thinking in Improving Skills in KBAT Problem Solving.* Jurnal Pendidikan Sains dan Matematik Malaysia, 2021. **11**(1): p. 60-76.
8. Abd Ghani, A., *The teaching of indigenous Orang Asli language in Peninsular Malaysia.* Procedia-Social and Behavioral Sciences, 2015. **208**: p. 253-262.DOI: https://doi.org/10.1016/j.sbspro.2015.11.201.
9. Anderson, L.W. and D.R. Krathwohl, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, p. 323-331*. 2001: Longman.
10. Heong, Y.M., et al., *The level of marzano higher order thinking skills among technical education students.* International Journal of Social Science and Humanity, 2011. **1**(2): p. 121.DOI: https://doi.org/10.7763/IJSSH.2011.V1.20.
11. Kiong, T.T., et al., *Thinking skills for secondary school students in Malaysia.* Journal of Research, Policy & Practice of Teachers and Teacher Education, 2012. **2**(2): p. 12-23.
12. Merta Dhewa, K., et al., *The development of Higher Order Thinking Skill (Hots) instrument assessment in physics study.* IOSR Journal of Research & Method in Education (IOSR-JRME), 2017. **7**(1): p. 26-32.DOI: https://doi.org/10.9790/7388-0701052632.
13. Hadi, S., et al., *The difficulties of high school students in solving higher-order thinking skills problems.* Problems of Education in the 21st Century, 2018. **76**(4): p. 520.DOI: https://doi.org/10.33225/pec/18.76.520.
14. Peng, C.F. and Z.H. Hamad, *Higher Order of Thinking Skills in Teaching and Learning Malay Language through Questioning Technique.* Jurnal Pendidikan Bahasa Melayu, 2018. **8**(2): p. 1-12.
15. Ramdiah, S., et al., *Understanding, planning, and implementation of HOTS by senior high school biology teachers in Banjarmasin-Indonesia.* International Journal of Instruction, 2019. **12**(1): p. 425-440.
16. Ichsan, I.Z., et al., *HOTS-AEP: Higher Order Thinking Skills from Elementary to Master Students in Environmental Learning.* European Journal of Educational Research, 2019. **8**(4): p. 935-942.DOI: https://doi.org/10.12973/eu-jer.8.4.935.
17. Sanip, F.A. and C.N. Che Ahmad, *Awareness of metacognitive strategies and high -level thinking skills among Biology students.* Journal of Educational Research, 2014. **15**.
18. Chinedu, C.C., O.S. Olabiyi, and Y. Kamin, *Strategies for improving higher order thinking skills in teaching and learning of design and technology education. 7(2).* 2015.
19. Saido, G.M., et al., *Higher order thinking skills among secondary school students in science learning. MOJES: Malaysian Online Journal of Educational Sciences, 3(3), 13-20.* 2015.
20. Innocenti, B., et al., *Development and validation of a wear model to predict polyethylene wear in a total knee arthroplasty: a finite element analysis.* Lubricants, 2014. **2**(4): p. 193-205.DOI: https://doi.org/10.3390/lubricants2040193.
21. Abidin, M.Z.Z. and K. Osman, *The level of knowledge, understanding, skills and implementation of science teachers on high -level thinking skills (HLTS). Journal of Advanced Research in Social and Behavioral Sciences, 8 (1), 97-113.* 2017.

22. Peng, C.F. and Z.H. Hamad, *Higher Order of Thinking Skills in Teaching and Learning Malay Language through Questioning Technique.* Journal of Malay Language Education, 2018. **8**(2): p. 1-12.
23. Kassim, N. and E. Zakaria, *Integration of high -level thinking skills in mathematics teaching and learning: An analysis of teacher needs. Journal of Mathematics Education, 3 (1), 1–12.* 2015.
24. Saido, G.A.M., et al., *Teaching strategies for promoting higher order thinking skills: A case of secondary science teachers.* MOJEM: Malaysian Online Journal of Educational Management, 2017. **3**(4): p. 16-30.
25. Retnawati, H., et al., *Teachers' knowledge about higher*-order *thinking skills and its learning strategy*. Problems of Education in the 21st Century, 2018. **76**(2): p. 215.DOI: https://doi.org/10.33225/pec/18.76.215.
26. Tandas, J.B., *Development of a Valid and Reliable Summative Test in Plane Trigonometry.* Journal of Science and Mathematics Letters, 2021. **9**(1): p. 46-59.
27. Rahman, S.M.A. and S. Ya'acob, *Gamified learning to improve higher order thinking among Malaysian students.* Open International Journal of Informatics (OIJI), 2018: p. 9-21.
28. Chun, T.C. and M.N.L.Y.B. Abdullah, *The teaching of higher order thinking skills (HOTS) in Malaysian schools: Policy and practices.* MOJEM: Malaysian Online Journal of Educational Management, 2019. **7**(3): p. 1-18.DOI: https://doi.org/10.22452/mojem.vol7no3.1.
29. Nor, M., et al., *Application of high -level thinking skills (HLTS) in the Design and Technology (RBT) curriculum of primary schools.* International journal of education and training, 2017. **3**(2): p. 1-7.
30. Seman, S.C., W.M.W. Yusoff, and R. Embong, *Teachers challenges in teaching and learning for higher order thinking skills (HOTS) in primary school.* International Journal of Asian Social Science, 2017. **7**(7): p. 534-545.DOI: https://doi.org/10.18488/journal.1.2017.77.534.545.
31. Subiantoro, A.W. and D.F. Treagust, *Development and validation of an instrument for assessing high*-*school students' perceptions of socio*-*scientific issues-based learning in biology.* Learning Environments Research, 2021. **24**(2): p. 223-237.DOI: https://doi.org/10.1007/s10984-020-09332-z.
32. Aisyah, A., et al. *Eliciting Elements of Higher Order Thinking Skills in the Higher Secondary Examination Question Structure in Japan and Malaysia*. Springer.DOI: https://doi.org/10.1007/978-981-13-0203-9_42.
33. Agarwal, S., et al., *Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist.* bmj, 2016. **352**.DOI: https://doi.org/10.1136/bmj.i1174.
34. Sulaiman, T., A.F.M. Ayub, and S. Sulaiman, *Curriculum change in English language curriculum advocates higher order thinking skills and standards-based assessments in Malaysian primary schools.* Mediterranean Journal of Social Sciences, 2015. **6**(2): p. 494-494.DOI: https://doi.org/10.5901/mjss.2015.v6n2p494.
35. Morkel, C. and M. Ramasobama, *Measuring the effect of evaluation capacity building initiatives in Africa: A review.* African Evaluation Journal, 2017. **5**(1).DOI: https://doi.org/10.4102/aej.v5i1.187.
36. Tekkaya, C., Ö. Özkan, and S. Sungur, *Biology concepts perceived as difficult by Turkish high school students.* Hacettepe üniversitesi eğitim fakültesi dergisi, 2001. **21**(21).
37. Atilla, C. and imer, *What makes biology learning difficult and effective: Studentsâ€™ views.* Educational research and reviews, 2012. **7**(3): p. 61-71.
38. Ozcan, T., et al., *Identifiying and comparing the degree of difficulties biology subjects by adjusting it is reasons in elemantary and secondary education.* Procedia-Social and Behavioral Sciences, 2014. **116**: p. 113-122.DOI: https://doi.org/10.1016/j.sbspro.2014.01.177.
39. Haris, N. and K. Osman, *The effectiveness of a virtual field trip (VFT) module in learning biology.* Turkish Online Journal of Distance Education, 2015. **16**(3): p. 102-117.DOI: https://doi.org/10.17718/tojde.13063.
40. Gungor, S.N. and M. Ozkan, *Evaluation of the concepts and subjects in biology perceived to be difficult to learn and teach by the pre-service teachers registered in the pedagogical formation program.* European Journal of Educational Research, 2017. **6**(4): p. 495-508.DOI: https://doi.org/10.12973/eu-jer.6.4.495.
41. Fauzi, A. and M. Mitalistiani, *High school biology topics that perceived difficult by undergraduate students.* Didaktika Biologi: Jurnal Penelitian Pendidikan Biologi, 2018. **2**(2): p. 73-84.DOI: https://doi.org/10.32502/dikbio.v2i2.1242.
42. Salleh, W.N.W.M., C.N.C. Ahmad, and E. Setyaningsih, *Difficult topics in Biology from the view point of students and teachers based on KBSM implementation.* EDUCATUM Journal of Science, Mathematics and Technology, 2021. **8**(1): p. 49-56.
43. Harrison, R.G., *A textbook of human embryology.* Academic Medicine, 1964. **39**(9): p. 868.

44. Anwar, A.H., N.Y. Rustaman, and W. Purwianingsih. *Development of three-tier diagnostic test instruments for detecting students' conception*. IOP Publishing.DOI: https://doi.org/10.1088/1742-6596/1318/1/012064.
45. Kazeni, M.M.M., *Development and Validation Test of the Integrated Science Process Skills for the Further Education and Training Learner's, University of Pretoria, South Africa.* Retrieved June, 2005. **10**: p. 2019.
46. Caruth, G.D., *Demystifying mixed methods research design: A review of the literature*. Online Submission, 2013. **3**(2): p. 112-122.DOI: https://doi.org/10.13054/mije.13.35.3.2.
47. Cohen, R.J., M.E. Swerdlik, and S.M. Phillips, *Psychological testing and assessment: An introduction to tests and measurement 7th edition.* 1996: Mayfield Publishing Co.
48. Anastasi, A. and S. Urbina, *Psychological testing 7th edition* 1997: Prentice Hall/Pearson Education.
49. Popham, W.J. and T.R. Husek, *IMPLICATIONS OF CRITERION-REFERENCED MEASUREMENT 1, 2.* Journal of Educational Measurement, 1969. **6**(1): p. 1-9.
50. Haladyna, T.M., *Writing Test Items to Evaluate Higher Order Thinking*. 1997: ERIC.
51. Morrison, S. and K.W. Free, *Writing multiple-choice test items that promote and measure critical thinking.* Journal of Nursing Education, 2001. **40**(1): p. 17-24.DOI: https://doi.org/10.3928/0148-4834-20010101-06.
52. Popham, W.J., *Classroom assessment: What teachers need to know*. 1999: ERIC.
53. Naglieri, J.A., et al., *Psychological testing on the Internet: new problems, old issues.* American Psychologist, 2004. **59**(3): p. 150.DOI: https://doi.org/10.1037/0003-066X.59.3.150.
54. Kryklywy, J., *Kaplan, RM & Saccuzzo, DP (2009) Psychological Testing Principles, Applications, and Issues. –abbreviated paperback edition printed especially for this course.(Belmont, CA.: Wadsworth).(Note: the words Psychology 2080 A/B appear on the cover of this book.).*
55. Czerwinski, F., A.C. Richardson, and L.B. Oddershede, *Quantifying noise in optical tweezers by allan variance.* Optics express, 2009. **17**(15): p. 13255-13269.DOI: https://doi.org/10.1364/OE.17.013255.
56. Nunnally, J.C., *Psychometric Theory 2nd edition (New York: McGraw).* 1978.
57. Moore, G.C. and I. Benbasat, *Development of an instrument to measure the perceptions of adopting an information technology innovation.* Information systems research, 1991. **2**(3): p. 192-222.DOI: https://doi.org/10.1287/isre.2.3.192.
58. Wilson, F.R., W. Pan, and D.A. Schumsky, *Recalculation of the critical values for Lawshe's content validity ratio.* Measurement and evaluation in counseling and development, 2012. **45**(3): p. 197-210.DOI: https://doi.org/10.1177/0748175612440286.
59. Cohen, L., L. Manion, and K. Morrison, *Research methods in education, 6th edition.* . 2002: routledge.DOI: https://doi.org/10.4324/9780203224342.
60. Popham, W.J., *Modern educational measurement: Practical guidelines for educational leaders, 3rd edition.* 2000: Pearson College Division.DOI: https://doi.org/10.1111/j.1745-3984.1969.tb00654.x.
61. Aiken, L.R., *Psychological testing and assessment, 9th edition.* . 2009: Pearson Education India.
62. Cohen, R.J. and R. Jay, *Exercises in psychological testing and assessment.* Vol. 2. 2005: McGraw-Hill New York, NY, USA:.DOI: https://doi.org/10.1037/1040-3590.17.3.375.
63. Dorset, D.L., *X-ray diffraction: a practical approach.* Microscopy and microanalysis, 1998. **4**(5): p. 513-515.DOI: https://doi.org/10.1017/S143192769800049X.
64. Loewenthal, K.M., *An Introduction to Psychological Tests and Scales (2nd Ed.). University of London: Psychology Press.* 2001.
65. Onyefulu, C., *Assessment practices of teachers in selected primary and secondary schools in Jamaica.* Open Access Library Journal, 2018. **5**(12): p. 1-25.DOI: https://doi.org/10.4236/oalib.1105038.
66. Kerlinger, F.N., H.B. Lee, and D. Bhanthumnavin, *Foundations of behavioral research: The most sustainable popular textbook by Kerlinger & Lee (2000).* Journal of Social Development, 2000. **13**: p. 131-144.
67. Brasselet*, E., et al., *Light-induced nonlinear rotations of nematic liquid crystal droplets trapped in laser tweezers.* Molecular Crystals and Liquid Crystals, 2009. **512**(1): p. 143-1989.DOI: https://doi.org/10.1080/15421400903050780.
68. Lewis, A. and D. Smith, *Defining higher order thinking.* Theory into practice, 1993. **32**(3): p. 131-137.DOI: https://doi.org/10.1080/00405849309543588.
69. Aziz, N.F.A., H. Ahmad, and I.M. Nashir, *Validation of technical and vocational teachers' competency evaluation instrument using the rasch model.* Jurnal Pendidikan Sains Dan Matematik Malaysia, 2019. **9**(1): p. 18-25.DOI: https://doi.org/10.37134/jpsmm.vol9.1.3.2019.