Characteristics of National Standardized School Examination Test Items on Biology Subject in High School

Murti Ambarwati Hidayah[1*], Heri Retnawati[2] , Eri Yusron[3] 🆔

[1,2] Educational Research and Evaluation Graduate School, Yogyakarta State University,Yogyakarta, Indonesia
[3] Elementary Teacher Education Study Program, Indonesia University of Education, Bandung, Indonesia

A B S T R A K

Perangkat tes yang baik digunakan untuk mengukur kemampuan akademik siswa. Bagi siswa SMA kelas akhir, kemampuan ini diukur dengan soal Ujian Sekolah Berstandar Nasional (USBN). Penelitian ini bertujuan untuk menganalisis karakteristik butir soal USBN biologi dan menganalisis butir soal yang sulit dan tidak baik. Jenis penelitian ini adalah deskriptif eksploratif dengan pendekatan kuantitatif dan kualitatif. Analisis data kuantitatif dilakukan menggunakan teori tes klasik dan teori respon butir dengan model 1 parameter logistik. Analisis data kualitatif dilakukan dengan mendeskripsikan butir soal sulit dan tidak baik menurut teori tes klasik. Data dikumpulkan melalui dokumentasi respon jawaban dari 193 siswa. Hasil menunjukkan bahwa menurut teori tes klasik, butir soal memiliki rata-rata tingkat kesukaran 0,525 dikategorikan sedang dan rata-rata daya beda butir 0,124 dikategorikan baik. Berdasarkan teori respon butir, rata-rata tingkat kesukaran diperoleh -0,00029 dikategorikan sedang. Estimasi uji reliabilitas sebesar 0,51 dikategorikan sedang. Hasil analisis kualitatif menunjukkan butir yang termasuk kategori sulit yaitu pada materi Jaring makanan dalam ekosistem, sistem saraf, dan sistem ekskresi manusia.

A B S T R A C T

A good test instrument is required to measure students' academic abilities. For final-year students, this ability is measured by USBN tests. This study aimed to analyses the characteristics of biology test items in the National Standardized School Examination (USBN) and analyses item tests categorized as difficult and poor. Type of study was descriptive exploratory with quantitative and qualitative approaches. Quantitative data analysis was performed using the classical test and item response theories with 1 logistic parameter model. Qualitative data analysis described why such question items are categorized as difficult and poor according to the classical test theory. The data were collected by using documentation of the answers of 193 students. The study results show that, according the classical test theory, Biology USBN item tests had the average difficulty index of 0.525 (moderate category) and the average of discrimination index obtained is equal to 0.124 (good category). Based on the item response theory, the mean difficulty index was -0.00029 (moderate category). The reliability test estimation result was 0.51, categorized as moderate. The qualitative analysis result showed that items categorized as difficult are the food web in the ecosystem, nervous system, and human excretion system.

1. INTRODUCTION

Assessment is a vital part of education. Assessment and measurement are closely related to evaluation. Assessment in education is the process of data collection through measurement using measuring apparatus and information processing to determine the learning outcome achievements of students. The assessment aims to discover the ability and success of students in achieving predetermined learning outcomes (Almaleki, 2021; Anderson et al., 2001). Education goals are achieved if the implementation follows the government provisions. One of which is by following the student assessment regulation issued by the government. Assessment following the Regulation of the Minister of Education and Culture (Permendikbud) No. 23 of 2016 is that learning outcomes are assessed in tests, observation, assignments, or other necessary forms. Educators' learning outcome assessment is used to measure and discover the competence achievement of students, improve the learning process, and rearrange the improvement report of daily, mid-term, end-term, end-year, and end-grade learning outcomes (Hussin et al., 2018; Kurniawati & Sundawa, 2019). The government's learning outcome assessment is conducted in a national exam or other necessary forms. Previous researcher suggests that an integrated assessment process can determine student abilities and monitor developments in achieving learning outcomes (Elvira

Corresponding author.
E-mail addresses: mahidayah1@gmail.com (Murti Ambarwati Hidayah)

& Hadi, 2016). Assessment is provided as a reference to improve learning quality (Laila Puspita, 2019; Putra et al., 2021). One of the assessment methods in learning is evaluating learning outcomes.

Learning outcome evaluation is an effort to measure education quality. A means of evaluation is by holding the National-Standardized School Examination (USBN) from primary school to high school and equivalent levels in 2018. The National-Standardized School Examination (USBN) was held to fulfill the mandate of Law No. 20 of 2003 concerning the National Education System. USBN aims to nationally assess graduate competencies (in Indonesian, Mathematics, and Science subjects), comparability between schools, regions, and years in the subjects tested. The USBN score is a consideration for mapping the education unit quality, selecting the next education level, and determining student graduation from an education unit (Herkusumo, 2011; Sappaile & Pristiwaluyo, 2019). USBN functions as the graduation determinant of students from an education unit. The measurement instrument commonly used in education comprises tests and non-tests. Tests of USBN questions are measurement instruments that provide information on cognitive characteristics. According to previous research measurement using tests can estimate possible errors from participant responses on the said test (Kartowagiran et al., 2018; Santoso, 2018). In ensuring the test to illustrate the actual subject's achievement or ability, quality tests are necessary.

Quality tests are challenging to produce. It takes special abilities from the test maker and other measurement techniques to help produce quality tests. In examining a test series quality in various aspects, question item analysis is required. The primary goal of question item analysis is to have information on the characteristics of each question item or empirical analysis (Erguven, 2013; Jannah et al., 2021; Santoso, 2018). CTT utilizes measures of item characteristics, item difficulty and item discrimination, the values of which are dependent upon the distribution of examinee proficiency within a sample. The quality of the test as a whole will be determined on the basis of some coefficients for reliability and validity. In IRT framework the item characteristics are decided based on values taken by the parameters of the model chosen for the item response. The parameters are estimated from the samples chosen for the item analysis. Based on the values taken by the parameters for each item, the quality of the item will be decided (Ashraf, 2020; Hu et al., 2021; O. A. & E. R. I., 2016).

As an education unit, high school will have its success observed from the evaluation outcomes of its students. One of the learning outcome evaluation forms in high school is the National-Based School Exam. The exam score illustrates the achievement of competency standard completeness and learning quality implemented by teachers. Therefore, the test instrument should be representative in measuring each student's achievement aspect. One of the subjects in USBN is Biology. Biology is a science studying living things and their life processes from various aspects or symptoms displayed by nature. (Asmat & Naz, 2020; Etobro & Fabinu, 2017). Students' learning difficulties in Biology subjects are also caused by low learning intensity, teacher-centered teaching strategies and facilities and learning facilities and infrastructure that do not support the Biology learning process (Fauzi & Mitalistiani, 2018; Muspikawijaya et al., 2017). In its implementation, the National-Standardized School Examanition (USBN) is created shortly, where teachers do not master an excellent development technique of quality questions regarding the arrangement of a good instrument, particularly the emphasis on discrimination index and difficulty index. Hence, in discovering quality USBN questions, a study is required concerning item characteristics or Item analysis is carried out based on certain procedures and steps to identify which items are effective and of good quality to be used as an assessment tool in the National-Strandardized School Examination.

Studies on characteristics and analyses items in biology subjects had been carried out, reviewing question items from the difficulty level and power difference of each question (Aldyza et al., 2018; Kusnani et al., 2016; Muzayanah, 2015). Similar studies focusing on reviewing final exam questions had been carried out, both on the National Exam (UN) and National-Standardized School Examination (USBN), such as research conducted by (Argianti & Retnawati, 2020; Sappaile & Pristiwaluyo, 2019). The final exam question quality can be examined from the question validity level and reliability coefficient index (Budianingsih et al., 2017; Muyasaroh, 2014; Quaigrain & Arhin, 2017; Sari et al., 2019). The current article answers the USBN question item quality problem in high school on Biology subject. The question item quality in this study focuses on the analysis of difficulty index and discrimination index of each question item with the classical test and item response theories using the 1 logistic parameter model (1 PL) and describing why such a question is poor or difficult. So, this study aimed to analyses the characteristics of biology test items in the National Standardized School Examination (USBN) and analyses item tests categorized as difficult and poor in the high school for Biology subject in Yogyakarta City.

2. METHODS

This study is a type of descriptive exploratory using quantitative and qualitative approaches. Quantitative approach is used to describe the characteristics of biology test items using the classical test theory and item response theory. Qualitative approach was performed by describing the question as poor or difficult. Respondens in this study were 193 class XII high school students, who took biology national standardized school exam in Yogyakarta. Data were collected by documenting student responses. The data source comes from the test of Biology national standardized school exam, with consist of 35 multiple choice question , answer key and 193 student responses. Question items in the test were objective tests with five possible answers, where one answer is the key answer and the other answers are the distractors. Quantitative analysis was conducted using the classical test theory to calculate discrimination index and difficulty index using Anbuso 4.0 application. Meanwhile, analysis using the item response theory used the 1 logistic parameter by the Quest program to calculate difficulty index and reliability. Qualitative data analysis was performed by describing the question as poor or difficult. The qualitative analysis steps include identifying items with a serious difficulty, outlining the question completion procedure, and analyzing factors that make these items difficult for students. Based on the classical test theory analysis, discrimination index of a test item in this study was classified using two categories. The discrimination coefficient value > 0, included in the good category, and < 0 are the poor category (Kolte, 2015). Difficulty index is one of the key parameters of item analysis. The interpretation of the item difficulty index (Karim et al., 2021; Sharma, 2021), are shown in Table 1.

Table 1. Classification of Difficulty Index Level

Criteria	Category
≤ 0,3	Difficult
> 0,3 - ≤ 0,7	Moderate
< 0,7	Easy

Based on the item response theory analysis using Quest program, criteria a good item with a difficulty index ranging from -2 to +2, a difficulty index more than -2 is categorized as an easy question, while a difficulty index of less than +2 categorize as a difficult question (Sainuddin, 2018; Syahrul Sarea & Hadi, 2015). Classification of reliability coefficient (Kaltakci-Gurel et al., 2017; Suparman, 2020), are presented in Table 2.

Table 2. Classification of Reliability Coefficient

Reliability coefficient	Category
0,8-1,00	Very high
0,60-0,80	High
0,40-0,60	Moderate
0,20-0,40	Low
-1,00-0,20	Very low

3. RESULT AND DISCUSSION

Results

Question item analysis in this study is based on two approaches. First, the classical test theory results in question item characteristics, including the difficulty index and discrimination index. Subsequently, the item response theory will be analyzed using the 1 logistic parameter (1 PL) to calculate difficulty index and reliability. The results of difficulty level analysis of the USBN Grade XII of State High Schools in Yogyakarta City based on the classical test theory are presented in Table 3.

Table 3. Difficulty Index Coefficient of Biology USBN Test Based on Classical Test Theory

No	p	Category	No	p	Category	No	p	Category
1	0.218	Difficult	13	0,430	Moderate	25	0.497	Moderate
2	0.560	Moderate	14	0,943	Easy	26	0.461	Moderate
3	0.420	Moderate	15	0,155	Difficult	27	0.974	Easy
4	0.503	Moderate	16	0,073	Difficult	28	0.404	Moderate
5	0.316	Moderate	17	0,409	Moderate	29	0.896	Easy
6	0.150	Difficult	18	0,953	Easy	30	0.969	Easy

No	p	Category	No	p	Category	No	p	Category
7	0.456	Moderate	19	0,383	Moderate	31	0.404	Moderate
8	0.834	Easy	20	0,891	Easy	32	0.839	Easy
9	0.176	Difficult	21	0,731	Easy	33	0.166	Difficult
10	0.637	Moderate	22	0,642	Moderate	34	0.736	Easy
11	0.228	Difficult	23	0,420	Moderate	35	0.497	Moderate
12	0.845	Easy	24	0,161	Difficult			

Information:
p = Difficulty Index Coefficient

Table 3 shows that, from 35 questions, eight items or 22.86% were categorized as difficult, 16 items or 62.85% were categorized as moderate, and 11 items or 31.43% were categorized as easy. The dominant category was moderate in the analyzed Biology USBN questions. Based on the classical test theory, the mean difficulty index coefficient score of USBN test items Grade XII of State High Schools in Yogyakarta was 0.525, categorized as moderate. The results of question item analysis based on the discrimination index category following the classical test theory are presented in Table 4.

Table 4. Discrimination Index Coefficient of Biology USBN Test Based on Classical Test Theory

No	r_{pbis}	Category	No	r_{pbis}	Category	No	r_{pbis}	Category
1	0.030	Good	13	0.157	Good	25	0.099	Good
2	0.132	Good	14	0.044	Good	26	0.112	Good
3	0.119	Good	15	0.056	Good	27	0.152	Good
4	0.303	Good	16	-0.019	Poor	28	-0.150	Poor
5	0.089	Good	17	0.191	Good	29	0.317	Good
6	-0.040	Poor	18	0.222	Good	30	0.169	Good
7	0.073	Good	19	0.124	Good	31	0.199	Good
8	0.281	Good	20	0.035	Good	32	0.338	Good
9	-0.035	Poor	21	0.226	Good	33	0.258	Good
10	0.203	Good	22	0.227	Good	34	0.163	Good
11	-0.105	Poor	23	0.138	Good	35	0.220	Good
12	-0.017	Poor	24	0.034	Good			

Information:
r_{pbis} = Discrimination Index Coefficient

Based on Table 4, the discrimination index classification results of 35 USBN questions showed that 29 (82.86%) were categorized as good, and six (17.14%) were categorized as poor. It concludes that question items with good power differences dominated the question items. Based on the classical test theory, the mean discrimination index coefficient score of USBN test items Grade XII of State High Schools in Yogyakarta was 0.124, categorized as good. Then, the results of difficulty index analysis of the USBN based on the item response theory are presented in the following Table 5.

Table 5. Difficulty Index Coefficient of Biology USBN Test Based on Item Response Theory

No	p	Category	No	p	Category	No	p	Category
1	1.60	Moderate	13	0.56	Moderate	25	0.26	Moderate
2	0.10	Moderate	14	-2.64	Easy	26	0.43	Moderate
3	0.61	Moderate	15	2.02	Difficult	27	-3.46	Easy
4	0.25	Moderate	16	2.88	Difficult	28	0.66	Moderate
5	1.07	Moderate	17	0.64	Moderate	29	-1.98	Moderate
6	2.06	Difficult	18	-2.85	Easy	30	-3.45	Easy
7	0.45	Moderate	19	0.76	Moderate	31	0.66	Moderate
8	-1.46	Moderate	20	-1.98	Moderate	32	-1.47	Moderate
9	1.86	Moderate	21	-0.81	Moderate	33	1.94	Moderate
10	-0.33	Moderate	22	-0.38	Moderate	34	-0.81	Moderate
11	1.54	Moderate	23	0.61	Moderate	35	0.28	Moderate
12	-1.51	Moderate	24	1.97	Moderate			

Information:
p = Difficulty Index Coeffient

Based on the difficulty index analysis with the item response theory, from 35 questions, three items or 8.57% were categorized as difficult, 28 items or 80% were categorized as moderate, and four items or 11.43% were categorized as easy. The difficulty index from easy to difficult ranged between -3.46 to 2.88. Based on the item response theory, the mean difficulty level coefficient score of USBN test items Grade XII of State High Schools in Yogyakarta was -0.00029, categorized as moderate. The reliability score of Biology USBN test items was obtained calculated using the Quest program is presented in Table 6.

Table 6. Reliability of Biology USBN Test

Reliability	Score
Internal Consistency	0,51

Based on the data presented in Table 6, the internal consistency reliability score was 0.51, categorized as moderate. The qualitative analysis revealed that question items were categorized as difficult according to the classical test theory and item response theory approaches were questions number 6, 15, and 16.
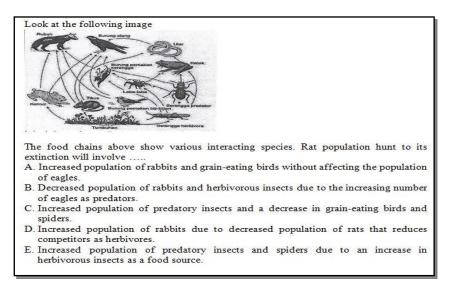

Figure 1. Test Item of USBN No. 6

Base on Figure 1, the question presents an image of food webs showing various interacting species in an ecosystem. Students are asked to answer what will happen if a living population in the image is missing. The question item is difficult and poor since it generates a negative power difference coefficient; thus, it fails to measure the actual student ability or not smart students cannot answer such a question. In answering the question, students should comprehend the concept of food webs in an ecosystem.

Figure 2. Test Item of USBN No. 15

Base on Figure 2, the question is a health article on the danger of using ecstasy on pregnant women towards infants during pregnancy. From this article, students are asked to answer the reason for delayed fetal motoric development.

A man visits a doctor after experiencing excessive fatigue, hunger, frequent urination, thirst, dry mouth with ketone smell, and difficulty in healing wounds.
The results of the urine test in the laboratory are as in the following table:

Component	Concentration in urine	Note
Water	+++++	+ = present
Na+	+++	- = absent
Cl-	+++	
Albumine	-	
Amino acids	-	
Glucose	++	
Ammonia	+++	

Based on the symptoms and urine test, the patient has problems at the….
A. Henle loop.
B. distal tubule.
C. proximal tubule.
D. Glomerulus.
E. collecting tubule.

Figure 3. Test Item of USBN No. 16

Base on Figure 3, the question presents a problem of a man with health complaints mentioned above, then displaying the urine test result. From this question, students are asked about the problem. This question item number 16 is the most difficult question based on the classical test theory and a poor question since it generates a negative power difference coefficient. That is, this question does not differentiate the actual student ability.

Discussion

Tests such as USBN play an important role in providing feedback to teachers regarding the actions taken so that the quality of the test is a critical issue. After administering and grading a test, a teacher needs to know how well the test items are and whether they reflect student performance in the course in relation to the specific learning objectives taught over a certain period of time (Koçdar et al., 2016; Quaigrain & Arhin, 2017). The quality of a test can be done by analyzing items based on the characteristics of the items that can be used to ensure that items have an appropriate standard for inclusion in the test, or the item requires improvement. Characteristics of the USBN Biology items were analyzed quantitatively based on the classical test theory and item response theory (Preston, 2018; Syahrul Sarea & Hadi, 2015). Based on the results of the classical test theory analysis to calculate the difficulty indexs, it was found that the most items were in the moderate category as many as 16 items or 62.85% with average difficulty index of 0.525 included the moderate category, and for the results of calculating the discrimination indexs of items, the results were good categories as many as 29 or 82.86% and 6 or 17.14% items in the poor category. It can be said that there are more items with good discrimination index than the existing items, that shows with the average of discrimination index obtained is equal to 0.124 belong to good category. Meanwhile, based on the analysis of the response theory, the item difficulty index was the most in the moderate category, which was 28 items or 80% with the mean difficulty index was -0.00029 included moderate category. Based on these results, the average of difficulty index of USBN items analyzed from classical test theory and item response theory is in the moderate category (Pakpahan et al., 2020; Rahmayani et al., 2013).

The item difficulty index shows how difficult or easy the item is. According to previous researcher the difficulty index of the questions is the proportion of the number of participants who correctly answered a question to all test participants so that categories of questions can be obtained which include easy, moderate and difficult (Maulida et al., 2015). Questions with a good level of difficulty are questions that are neither too easy nor too difficult. In addition, the item discrimination index is an important factor in the effectiveness of distractors. Most of the test items are able to distinguish between high and low abilities (Etobro & Fabinu, 2017; Jailani & Almukarramah, 2020; Retnawati, 2017). Therefore, it can be said that great item discrimination can lead to the effectiveness of distractors (Maharani & Putro, 2020;

Okoye, 2014). Based on this statement, the Biology USBN questions studied included questions with a good difficulty index. Besides that, they also had good discrimination index of items as indicated by the results of the criteria for more good items. Based on the analysis of discrimination index of the classical test theory, the items in the poor category are about the basic material of food webs in ecosystems, the mechanism of muscle contraction, plant tissue, functions of the nervous system, diseases of the human excretory system, and metabolism. This is indicated by the value of the negative differential power coefficient. A total of 6 items with negative coefficient values are included in the category of bad questions, because they cannot measure the actual ability of students or smart students cannot answer the question and students who are not smart can answer the question. The categories of questions are not good, namely items number 6, 9, 11, 12, 16, and 28. In line with the opinion of previous research that state items that have negative discrimination index should not be used in the next test (Sudijono, 2011).

Based on a qualitative analysis, items that are in the difficult category are taken based on the identification of the results of quantitative analysis including items number 6, 15 and 16. USBN Biology questions which are included in the difficult category are about food web material in ecosystems, nervous system disorders, and diseases on the human excretory system. The difficult questions are dominated by analytical items but there are also questions of understanding (Karagöz & Çakir, 2011). The main factor that causes difficult questions and many students get the wrong answer can be because some biology topics are considered difficult by students including abstract concepts, too many concepts, various biological events that cannot be seen directly by the eye, there are many foreign/latin words that are difficult understood, difficulty understanding the content of the lesson (Almroth, 2015; Çimer, 2012; Tekkaya, Ceren, 2001). As a result of the difficulty of biological material understood, many students only memorize biological facts to learn them (Anggani et al., 2016; Margalita et al., 2014). The way that can be done to overcome these difficulties is that there must be the participation of the teacher, where the teacher must have a strategy in teaching and learning activities, including often discussing questions, solving various problems related to the biology topic being taught. This is because when students work on or complete questions/problems, students can make a review of the topics they have just learned so that learning biology becomes more effective (Çimer, 2012; Retnawati, 2017).

National-Standardized School Examination (USBN) is one of the determinants of graduation for high school students. USBN is managed and created by teachers or subject teacher forums (MGMP) with a scope according to the provisions of each region. Based on this, the quality of the USBN test does not need to be doubted, but in this study the reliability test estimatiton with the Quest program obtained an internal consistency (Alpha) reliability result of 0.51 categorized as moderate. It indicates that USBN questions are less reliable. However, it can be improved. Items that become unreliable can be caused by several things, including the situation when the test was carried out and the state of the test itself (Argianti & Retnawati, 2020; Basuki & Hariyanto, 2014). Reliability decisions are made by confirming the reliability coefficient of the calculation results with certain limit criteria. According to previous researcher to obtain a reliable picture is indeed difficult, because the element of the human psyche is not fixed (Arikunto & Suharsimi, 2009). Moreover other researcher state that this problem can be overcome by increasing the insight and skills of teachers in making lesson plans, preparing and conducting biology assessments, by conducting training for biology teachers (Iswari et al., 2010).

This study can be used as an evaluation material for learning in High School at Yogyakarta City. In addition, this research will be very useful for education parties, both students, teachers and education stakeholders related to information on national standardized school examination test items on biology subject in high school. This research is still quite limited, one of which is on the subject of research that only involves one school. It is hoped that future research will be able to deepen and broaden the scope of research related to national standardized school examination test items.

4. CONCLUSION

The National-Standardized School Examination (USBN) on biology subject in high school at Yogyakarta City, based on the classical test theory analysis had the average of difficulty index is 0.525 included in the moderate category, and the average of discrimination index obtained is equal to 0.124 belong to good category. Based on the item response theory analysis, the mean difficulty index was -0.00029 included the moderate category and the reliability test estimation result was 0.51, categorized as moderate. The qualitative analysis results showed that difficult questions discuss food web in the ecosystem, nervous system, and human excretion system. Difficult and poor questions require comprehension and analysis, leading students not to guess the answer. Abstract biological materials provide information that is relatively difficult to be captured by the five senses directly. Therefore, students struggle to answer the question because they do not understand and master the concept, cannot

differentiate particular characteristics, are not familiar with working on analysis questions, and are distracted by distractors resembling the key answer.

5. REFERENCES

Aldyza, N., Afkar, & Andani, D. (2018). Analisis Tingkat Kesukaran Butir Soal Ujian Nasional (UN) IPA SMP Tahun Ajaran 2014 / 2015 di Kabupaten Aceh Tenggara. *Pendidikan Almuslim*, *VI*(1), 2014–2016. http://jfkip.umuslim.ac.id/index.php/jupa/article/view/330.

Almaleki, D. (2021). Examinee Characteristics and their Impact on the Psychometric Properties of a Multiple Choice Test According to the Item Response Theory (IRT). *Engineering, Technology & Applied Science Research*, *11*(2), 6889–6901. https://doi.org/10.48084/etasr.4056.

Almroth, B. C. (2015). The Importance of Laboratory Exercises in Biology Teaching; Case Study in An Ecotoxicology Course. *Pedagogical Development and Interactive Learning*, *september*, 1–11. https://pil.gu.se/digitalAssets/1550/1550056_carney-almroth-laboratory-studies-in-teaching.pdf.

Anderson, L. W., Krathwohl Peter W Airasian, D. R., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom's Taxonomy of Educational Objectives* (Abridged E). Ney York: Longman.

Anggani, A., Paidi, P., & Triharjana, T. (2016). Identifikasi Kesulitan Belajar Struktur Fungsi Jaringan Hewan pada Siswa Kelas XI SMA Negeri 1 Muntilan. *Jurnal Pendidikan Biologi*, *5*(4), 39–50. https://journal.student.uny.ac.id/index.php/jeb/article/view/4537.

Argianti, A., & Retnawati, H. (2020). Characteristics of Math National-Standardized School Exam Test Items in Junior High School: What Must be Considered? *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(2), 156–165. https://doi.org/10.21831/pep.v24i2.32547.

Arikunto, & Suharsimi. (2009). *Dasar-dasar evaluasi pendidikan*. PT Bumi Aksara.

Ashraf, Z. A. . K. J. (2020). Classical and Modern Methods in Item Analysis of Test Tools. *International Journal of Research and Review*, *7*(5), 5. https://doi.org/10.52403/ijrr.

Asmat, A., & Naz, F. (2020). Validation and Analysis of An Achievement Test in the Subject of Biology at the Secondary Level. *Journal of Science Education*, *2*(2), 73–90. http://journal.aiou.edu.pk/journal1/index.php/jse/article/view/1152.

Basuki, I., & Hariyanto. (2014). *Asesmen Pembelajaran*. PT Remaja Rosdakarya.

Budianingsih, R., Utama, I. D. G. B., & Sutama, I. M. (2017). Validitas dan Reliabilitas Soal UN Bahasa Indonesia Tahun 2016 untuk Jurusan IPS. *Journal Jurusan Pendidikan Bahasa Dan Sastra Indonesia, Undhiksa*, *7*(2), 1–10. https://doi.org/10.23887/jjpbs.v7i2.11612.

Çimer, A. (2012). What Makes Biology Learning Difficult and Effective: Students' Views. *Educational Research and Reviews*, *7*(3), 61–71. https://doi.org/10.5897/ERR11.205.

Elvira, M., & Hadi, S. (2016). Karakteristik Butir Soal Ujian Semester dan Kemampuan Siswa SMA di Kabupaten Muaro Jambi. *Jurnal Evaluasi Pendidikan*, *4*(1), 58–68. https://journal.student.uny.ac.id/index.php/jep/article/view/2172.

Erguven, M. (2013). Two Approaches to Psychometric Process : Classical Test Theory and Item Response Theory. *Journal of Education*, *2*(2), 23–30. https://jebs.ibsu.edu.ge/jms/index.php/je/article/view/84.

Etobro, A. B., & Fabinu, O. E. (2017). Students' Perceptions of Difficult Concepts in Biology in Senior Secondary Schools in Lagos State. *Global Journal of Educational Research*, *16*(2), 139. https://doi.org/10.4314/gjedr.v16i2.8.

Fauzi, A., & Mitalistiani, M. (2018). High School Biology Topics that Perceived Difficult by Undergraduate Students. *DIDAKTIKA BIOLOGI: Jurnal Penelitian Pendidikan Biologi*, *2*(2), 73. https://doi.org/10.32502/dikbio.v2i2.1242.

Herkusumo, A. P. (2011). Penyetaraan (Equating) Ujian Akhir Sekolah Berstandar Nasional (UASBN) dengan Teori Tes Klasik. *Jurnal Pendidikan Dan Kebudayaan*, *17*(4), 455. https://doi.org/10.24832/jpnk.v17i4.41.

Hu, Z., Lin, L., Wang, Y., & Li, J. (2021). The Integration of Classical Testing Theory and Item Response Theory. *Psychology*, *12*, 1397–1409. https://doi.org/10.4236/psych.2021.129088.

Hussin, W. N. T. W., Harun, J., & Shukor, N. A. (2018). Problem Based Learning to Enhance Students Critical Thinking Skill via Online Tools. *Asian Social Science*, *15*(1), 14. https://doi.org/10.5539/ass.v15n1p14.

Iswari, R., Mulyani, S., Saptono, S., Peniati, E., & Purwantoyo, E. (2010). Upaya Meningkatkan Kemampuan Guru SMA Mengajar Biologi dengan Metode Pembelajaran Aktif, Kreatif, Inovatif, Dan Menyenangkan di MGMP Biologi Kabupaten Demak. *Jurnal Abdimas*, *14*(1), 25504. https://doi.org/10.15294/abdimas.v14i1.22.

Jailani, J., & Almukarramah, A. (2020). Upaya Peningkatan Kualitas Pembelajaran Biologi melalui Pembelajaran Bermakna dengan Menggunakan Peta Konsep. *Jurnal Biology Education*, *8*(2), 122–130. https://doi.org/10.32672/jbe.v8i2.2371.

Jannah, R., Hidayat, D. N., Husna, N., & Khasbani, I. (2021). An Item Analysis on Multiple-Choice Questions : A Case of a Junior High School English Try-Out Test in Indonesia. *Leksika*, *15*(116), 9–17. https://doi.org/10.30595/lks.v15i1.8768.

Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Research in Science & Technological Education*, *35*(2), 238–260. https://doi.org/10.1080/02635143.2017.1310094.

Karagöz, M., & Çakir, M. (2011). Problem Solving in Genetics: Conceptual and Procedural Difficulties. *Educational Sciences: Theory & Practice*, *11*(3), 1668–1674. https://eric.ed.gov/?id=EJ936343.

Karim, S., Sudiro, S., & Sakinah, S. (2021). Utilizing Test Ttems Analysis to Examine the Level of Difficulty and Discriminating Power in A Teacher-Made Test. *Edulite*, *6*(2), 256–269. https://doi.org/10.30659/e.6.2.256-269.

Kartowagiran, B., Munadi, S., Retnawati, H., & Apino, E. (2018). The Equating of Battery Test Packages of Mathematics National Examination 2013-2016. *SHS Web of Conferences*, *42*(July 2019), 00022. https://doi.org/10.1051/shsconf/20184200022.

Koçdar, S., Karadag, N., & Sahin, M. (2016). Analysis of the Difficulty and Discrimination Indices of Multiple-Choice Questions According to Cognitive Levels in an Open and Distance Learning Context. *Journal of Educational Technology*, *15*(4), 16–24. https://eric.ed.gov/?id=EJ1117619.

Kolte, V. (2015). Item Analysis of Multiple Choice Questions in Physiology Examination. *Indian Journal of Basic and Applied Medical Research*, *4*(4), 320–326. https://www.ijbamr.com/assets/images/issues/pdf/september 2015 320-326.pdf.pdf.

Kurniawati, S., & Sundawa, D. (2019). An Analysis of National Standard School Examination Items Based on the Characteristics of Hots (Higher Order Thinking Skills) Questions for the Main Items of K13-071 Academic Year 2016/2017 in Karawang Regency. *International Journal Pedagogy of Social Studies*, *3*(2), 100–112. https://doi.org/10.17509/ijposs.v3i2.15793.

Kusnani, Muldayanti, N. D., & Rahayu, H. M. (2016). Analisis Butir Soal Ulangan Akhir Semester Ganjil Pada Mata Pelajaran Biologi Kelas X Mia SMA Negeri 1 Sungai Raya Tahun Ajaran 2014/2015. *Jurnal Bioeducation*, *3*(2), 42–52. https://doi.org/10.29406/185.

Laila Puspita. (2019). Pengembangan modul berbasis keterampilan proses sains sebagai bahan ajar dalam pembelajaran biologi Module development based on science process skills as teaching materials in biological learning. *Jurnal Inovasi Pendidikan IPA*, *5*(1), 79–87. https://doi.org/10.21831/jipi.v5i1.22530.

Maharani, A. V., & Putro, N. (2020). Item Analysis of English Final Semester Test. *Indonesian Journal of EFL and Linguistics*, *5*(2), 491–504. https://doi.org/10.21462/ijefl.v5i2.302.

Margalita, S., Rachmadiarti, F., & Prastiwi, M. (2014). Analisis Miskonsepsi Tertinggi Materi Ekologi pada Siswa Sekolah Menengah Atas (SMA). *BioEdu*, *3*(3), 571–579. https://jurnalmahasiswa.unesa.ac.id/index.php/1/article/view/13431.

Maulida, Muhibbudin, & Yusrizal. (2015). Analisis Indeks Kesukaran Dalam Pengembangan Item Tes pada Konsep Sel Tingkat Sekolah Menengah Atas. *Jurnal Edubio Tropika*, *3*(1), 42–45. http://www.jurnal.unsyiah.ac.id/JET/article/view/5276.

Muspikawijaya, Iswari, R., & Marianti, A. (2017). Analisis Kesulitan Peserta Didik SMA/MA Kabupaten Luwu Timur dalam Memahami Konsep pada Materi Metabolisme Sel. *Journal of Innovative Science Education*, *6*(2), 252–263. https://journal.unnes.ac.id/sju/index.php/jise/article/view/15439.

Muyasaroh, S. J. (2014). Validitas Soal-Soal Ujian Nasional Mata Pelajaran Bahasa Indonesia Untuk SMP/MTs. *Jurnal VARIDIKA*, *26*(1), 56–68. https://doi.org/10.23917/varidika.v26i1.733.

Muzayanah, U. (2015). Kualitas Butir Soal PAI pada Ujian Sekolah Berstandar Nasional. *Smart*, *1*(1), 3–4. https://doi.org/10.18784/smart.v1i1.234.

O. A., A., & E. R. I., A. (2016). Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. *European Scientific Journal, ESJ*, *12*(28), 263. https://doi.org/10.19044/esj.2016.v12n28p263.

Okoye, M. D. (2014). Authentic Assessment and Evaluation: Paramount Means for the Maximization of Teaching and Learning. *Journal of Educational and Social Research*, *4*(7), 31–40. https://doi.org/10.5901/jesr.2014.v4n7p31.

Pakpahan, T. R., Hernawati, D., & Ardiansyah, R. (2020). Analisis Miskonsepsi Peserta Didik pada Materi Sistem Saraf Menggunakan Four-Tier Diagnostic Test. *Bioeduscience*, *4*(1), 27–36. https://doi.org/10.29405/j.bes/4127-364844.

Preston, C. (2018). Food Webs : Implications for Instruction. *The American Biology Teacher*, *80*(5), 331–338. https://doi.org/10.1525/abt.2018.80.5.331.

Putra, E. N., Syamsurizal, Lufri, & Zulyusri. (2021). The Teacher-Made Multiple-Choice Tests Quality for Natural Science Subject. *Journal of Educational Research and Evaluation*, *5*(3), 422–428. https://doi.org/10.23887/jere.v5i3.33219.

Quaigrain, K., & Arhin, A. K. (2017). Using Reliability and Item Analysis to Evaluate A Teacher-Developed Test in Educational Measurement and Evaluation. *Cogent Education*, *12*(1). https://doi.org/10.1080/2331186X.2017.1301013.

Rahmayani, Sinambela, M., & Rosida. (2013). Analisis Faktor Kesulitan Belajar Siswa pada Materi Pokok Sistem Ekskresi Manusia Kelas XI MIA SMA Negeri 16 Medan. *Pelita Pendidikan*, *5*(2), 55–65. https://doi.org/10.24114/jpp.v5i2.8415.

Retnawati, H. (2017). Why are the Mathematics National Examination Items Difficult and What Is Teachers ' Strategy to Overcome It ? *International Journal of Instruction*, *10*(3), 257–276. https://doi.org/10.12973/iji.2017.10317a.

Sainuddin, S. (2018). Analisis Karakteristik Butir Tes Matematika Berdasarkan Teori Modern (Teori Respon Butir). *Jurnal Penelitian Matematika Dan Pendidikan Matematika*, *1*(1), 1–12. http://www.journal.uncp.ac.id/index.php/proximal/article/view/842.

Santoso, A. (2018). Karakteristik Butir Tes Pengantar Statistika Sosial Berdasarkan Teori Respon Butir. *Jurnal Pendidikan Matematika Dan Sains*, *6*(2), 158–168. https://doi.org/10.21831/jpms.v4i1.10111.

Sappaile, B. I., & Pristiwaluyo, T. (2019). Analisis Butir Soal Ujian Sekolah Berstandar Nasional dengan Pendekatan Klasik dan Teori Respon Butir Mata Pelajaran Matematika. *Proceedings of National Seminar Research and Community Service Institue Universitas Negeri Makassar*, *0*(15), 246–253. https://ojs.unm.ac.id/semnaslemlit/article/view/11384.

Sari, S. S., Aziz, A., & Nurhayati, N. (2019). Validitas dan Reliabilitas Soal USBN (Ujian Sekolah Berstandar Nasional) Tahun 2019 SMA Negeri Se-Kota Makassar. *Seminar Nasional LP2M UNM*, *3*, 443–446. https://ojs.unm.ac.id/semnaslemlit/article/view/11461.

Sharma, L. R. (2021). Analysis of Difficulty Index , Discrimination Index and Distractor Efficiency of Multiple Choice Questions of Speech Sounds of English. *International Reasearch Journal of MMC*, *2*(1), 15–28. https://doi.org/10.3126/irjmmc.v2i1.35126.

Sudijono, A. (2011). *Pengantar Evaluasi Pendidikan*. Rajawali Pers.

Suparman. (2020). Menemukan Karakteristik Butir Menggunakan Quest. *Komunikasi Dan Pendidikan Islam*, *9*, 83–104. https://doi.org/10.36668/jal.v9i1.134.

Syahrul Sarea, M., & Hadi, S. (2015). Analisis Kualitas Soal Ujian Akhir Semester Mata Pelajaran Kimia SMA di Kabupaten Gowa. *Jurnal Evaluasi Pendidikan*, *3*(1), 35–43. http://journal.student.uny.ac.id/ojs/index.php/jep%0D.

Tekkaya, Ceren, et Al. (2001). Biology Concepts Perceived as Difficult by Turkish High School Students. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *21*(21), 145–150. https://dergipark.org.tr/en/download/article-file/87982.