Development of A Test Instrument to Measure Hots and Cognitive Knowledge in Biology

V Septiani¹ and Paidi²

¹Biology Education, Postgraduate Program, Universitas Negeri Yogyakarta, Jl. Colombo No. 1, Yogyakarta, Indonesia
²Biology Education, Faculty of Mathematics and Natural Science, Universitas Negeri Yogyakarta, Jl. Colombo No. 1, Yogyakarta, Indonesia

*vinal1septiani@gmail.com

Abstract. SKL of primary and secondary education set each graduate have HOTS (analysis, evaluation, and create) and cognitive knowledge (factual, conceptual, procedural, and metacognitive). The aim of this research is to develop test instrument of multiple choice for measuring how the ability to HOTS and cognitive knowledge of biology of senior high school students on the subject of human digestive system. The type of this research is Research and Development (R&D) that based on the development model of Oriondo & Dallo-Antonio and Wilson. The field testing was be done in March-April 2020 involving 280 around the five state senior high school in Yogyakarta. The data based on analysis measurement Rasch Model and use Quest program to find out goodness of fit, difficulty index, and reliability. Result of analysis shows that test instrument of HOTS and Cognitive Knowledge were valid and reliable. The goodness of fit test for the overall test and each item shows fit with the Rasch model. Difficulty indexes for the item ranged -3.44 to +1.36. Reliabilities of item estimate were 0.99.

1. Introduction
The superior human resources can compete in this 21st century. Because the most valuable treasure for the nation is human resources. All nations can thrive and prosper, get quality and human resources to support it [1]. Thus, the government sets the primary and secondary education competency standards (SKL) to compete in the 21st century. At SKL, every graduate must possess knowledge competency (factual, conceptual, procedural, and metacognitive) and skills (creative, productive, critical, independent, collaborative, and communicative). When observed by the dimension of skills, creative and critical thinking are included in thinking skills [2, 3, 4]. Thinking skills are included in Higher Order Thinking Skills (HOTS) [5, 6, 7, 8, 9, 10, 11, 12]. So, elementary and secondary education graduates are required to have factual, conceptual, procedural, metacognitive knowledge, and HOTS.

The competence above refers to Bloom's taxonomy which was revised by Anderson & Krathwohl. Anderson & Krathwohl [13] divides two cognitive dimensions into process dimensions and cognitive knowledge. Dimensions of cognitive processes include: remembering/C1 (recognizing and recalling), understanding/C2 (interpreting, modeling, classifying, including, summarizing, comparing, and explaining), applying/C3 (executing and implementing), analyzing/C4 (differentiating, organizing, and attributing), evaluating/C5 (checking and criticizing), and creating/C6 (formulating, planning, and producing). Dimensions of cognitive knowledge include factual/K1, conceptual/K2, procedural/K3, and
metacognitive/K4. HOTS from the cognitive process dimension is defined as the top three levels, they are C4, C5, and C6.

By this time, the revised Bloom's taxonomy on the 2013 Curriculum hasn’t gone as expected, especially in assessing student learning outcomes. Like research conducted by Amirulloh, Rustaman & Sriyati [14] revealed that, the quality of biology SNMPTN questions in 2010-2012 only developed cognitive process dimensions: C1 (24.45%), C2 (62.22%), C3 (13.33%) and cognitive knowledge dimensions: factual (6.66%), conceptual (91.12%), procedural (2.22%). Zehlia, Luzyawati, & Hamidah [15] revealed that, the competency test questions of biology book publisher Erlangga cell division material only covers the dimensions of process and cognitive knowledge: C1 K1 (8%), C2 K1 (10%), C1 K2 (30%), C2 K2 (12%), C3 K2 (10%), C4 K2 (14%), C5 K2 (16%). Whereas, in material heredity patterns only covers the dimensions of process and cognitive knowledge: C1 K1 (11%), C1 K2 (9%), C2 K2 (7%), C4 K2 (2%), C5 K2 (13%), C3 K3 (60%). The latest research was carried out by Haryati [16] revealed that, the quality of biology UN in 2019 is 75% of Lower Order Thinking Skill (LOTS) and 25% of HOTS questions. The distribution of cognitive process dimensions: C1 (7.50%), C2 (27.50%), C3 (32.50%), C4 (25.00%), C5 (7.50%), dan C6 (2.50%). Based on the three results of the study it can be concluded that the developed test instruments have not measured the competencies expected in the SKL.

Paidi et al. [17] have developed a process assessment instrument and cognitive biology products from C1-C4 and K1-K4. However, instruments that measure cognitive processes and products from simple to complex levels are less effective at measuring HOTS. Therefore, we need an instrument that only focuses on measuring HOTS and students' cognitive knowledge. The author compiles the test instrument based on cognitive dimensions by Anderson & Krathwohl. HOTS measured include C4 (differentiating, organizing, and attributing) and C5 (checking). Cognitive knowledge measured includes K2 and K3. This study aimed to produce HOTS test instruments and cognitive knowledge of good quality so that it can be utilized for schools, teachers, and students. Benefits for schools as reference material in developing HOTS test instruments and cognitive knowledge consist of two, for teachers as an alternative to measuring HOTS and students' cognitive knowledge, for students as a learning resource to train and improve HOTS and cognitive knowledge.

2. Methods
The type of this research is Research and Development (R&D) based on the development model of Oriondono & Dallo-Antonio [18] and Wilson [19]. The stages of developing the test are shown in Figure 1.

![Figure 1. Test development stages.](image-url)
The design step of the test was carried out in July 2019 - February 2020. Competencies tested were knowledge on the topic of the human digestive system. The instrument was developed using a cognitive dimension table by Anderson & Krathwohl, where each item shows an intersection between HOTS and cognitive knowledge. The test instrument was made in the form of multiple choices. Logical validation was carried out by two expert judges, who act as experts on the assessment instruments and content experts. The distribution of the number of items based on the intersection between HOTS and cognitive knowledge can be shown in Table 1.

Table 1. The distribution of the number of items based on the intersection between HOTS and cognitive knowledge.

| No. | HOTS          | Cognitive knowledge     | Number of items |
|-----|---------------|-------------------------|-----------------|
| 1.  | Analysis (C4)|                         |                 |
| a.  | Differentiating | Conceptual knowledge (K2)| 1 |
|     |                | Procedural knowledge (K3)| 1 |
| 1.  | Organize      | Conceptual knowledge (K2)| 1 |
|     |                | Procedural knowledge (K3)| 2 |
| c.  | Attributing   | Conceptual knowledge (K2)| 1 |
|     |                | Procedural knowledge (K3)| 1 |
| 2.  | Evaluation (C5)|                       |                 |
| a.  | Checking      | Conceptual knowledge (K2)| 1 |
|     |                | Procedural knowledge (K3)| 2 |
|     | Total         |                         | 10              |

The trial phase of the test was conducted in March - April 2020. Regarding the sample size, some experts stated that the specific sample size for the Rasch model is between 30-300 people [20]. In this study, it was first determined that public high schools in the city of Yogyakarta have high, low, and average in biology UN scores. Five of the 11 public high schools were selected as test samples, one school had a high average score, three schools had a medium average score, and one school had a low average grade. The five public high schools include SMAN 5 Yogyakarta, SMAN 6 Yogyakarta, SMAN 7 Yogyakarta, SMAN 8 Yogyakarta, and SMAN 10 Yogyakarta. Field test results data are used as a guide to revising test items.

Data analysis of this research with the Rasch model used the QUEST program. Data analysis was carried out on several aspects: (1) goodness of fit, (2) difficulty index, and (3) reliability. The goodness of fit testing is carried out for the whole test or for each item. The goodness of fit test for the overall test is based on the mean value of Infit Mean Square, Outfit Mean Square, Infit t, and Outfit t along with their standard deviations. If the mean Infit Mean Square and Outfit Mean Square are close to 1.00 and the standard deviation is close to 0.00 or the average Infit t and Outfit t is close to 0.00 and the standard deviation is close to 1.00, then the overall test is fit with the Rasch model [21]. Testing the goodness of fit for each item's test based on the value of Infit Mean Square and Outfit t. If the Infit Mean Square value is between 0.77 - 1.30 and the Infit t value is between -2.0 to 2.0, then the test item is fit with the Rasch model [22]. An item is said to be good if the difficulty index is \(-2 \leq b \leq 2\) [23]. The higher the estimated item spacing index, the more precisely the whole item is analyzed according to the Rasch model [24].

3. Results and discussion

3.1. Goodness of fit
The goodness of fit-test for the overall test based on the mean value of Infit Mean Square, Outfit Mean Square, Infit t, and Outfit t along with their standard deviations, shows the overall fit test with the Rasch model (Table 2). Infit Mean Square ranges from 0.89 to 1.14 and the range of Infit t start from -1.5 to 1.3, which means that each test item is fits with the Rasch model.
Table 2. Overall goodness of fit test results.

| No. | Aspect                        | Item estimates | Case estimates |
|-----|-------------------------------|----------------|----------------|
| 1.  | Mean ± SD Infit Mean Square   | 1.00 ± 0.09    | 0.99 ± 0.56    |
| 2.  | Mean ± SD Outfit Mean Square  | 1.13 ± 0.37    | 1.13 ± 1.70    |
| 3.  | Mean ± SD Infit t             | -0.17 ± 0.96   | -0.02 ± 0.99   |
| 4.  | Mean ± SD Outfit t            | 0.28 ± 1.22    | 0.16 ± 0.85    |

3.2. Difficulty index
The threshold value revolves from -3.44 to 1.36, which means that there are items that have a very easy difficulty index, which is items C4 (attributing) - K2. To be clearer, the graph of the distribution item difficulty indexes on each HOTS and cognitive knowledge shown in Figure 2.

![Figure 2](image_url)

Figure 2. The distribution item difficulty indexes on each HOTS and cognitive knowledge.

Based on Figure 2, it can be seen that the sequence of item difficulty index at the consecutive trial stage is C4 (attributing) - K2, C4 (differentiating) - K3, C4 (organizing) - K2, C4 (organizing) - K3, C5 (checking) - K2, C5 (checking) - K3, C4 (differentiating) - K2, and C4 (attributing) - K3.

3.3. Reliability
The estimated value of item reliability is 0.99. This shows that overall items are precisely analyzed according to the Rasch model.

3.4. Discussion
Good instrument results can be trusted and able to measure what should be measured. The quality of the instruments arranged can be discovered by analyzing the instruments, which are in the form of validity and reliability [25]. In this study, the HOTS instrument and cognitive knowledge were analyzed with the Rasch model using the QUEST program. Sumintono & Widhiarso [26] revealed the strengths of the Rasch model, such as fulfilling the five principles of the measurement model, including: (1) being able to provide a linear scale with the same interval, (2) being able to predict the missing data, (3) being able to provide a more precise estimate, (4) able to detect inaccurate models, and (5) produce replicable measurements.

Logical validity is carried out with two expert judgments, which are tasked with assessing the content, construction, and language aspects. Empirical validity is evidenced by the goodness of fit on the Rasch model. The goodness of fit test for the overall test based on the mean value of Infit Mean
Square, Outfit Mean Square, Infit t, and Outfit t along with their standard deviations, shows the overall fit-test with the Rasch model. That is, the empirical validity of HOTS test instruments and Cognitive Knowledge is valid. The validity of the instrument is also supported by the value of Infit Mean Square and Infit t, where each item is compatible with the Rasch model. Then based on the reliability value, the HOTS test instrument and Cognitive Knowledge are precisely analyzed according to the Rasch model.

HOTS test instruments and Cognitive Knowledge developed is valid and reliable. This is caused by several things: (1) items are developed according to the correct instrument development procedures, (2) items are developed from indicators derived from HOTS, Cognitive Knowledge and material of the human digestive system, (3) tests consisting of 10 items have passed the logical validity test with two expert judgments, and (4) the respondents who tested were actually working on the test, because researchers promised rewards for students who scored above the KKM [27].

Thus, the developed test instrument can be used to measure HOTS and Cognitive Knowledge of students' biology on the material of the human digestive system. In case viewed from the difficulty index, there is one problem that needs to be revised, because it is included in the category of items that is very easy.

4. Conclusion
HOTS test instruments and Cognitive Knowledge was developed in the form of multiple choices in the ability to analyzing/C4 (differentiating, organizing, and attributing), the ability to evaluating/C5 (checking), conceptual knowledge, and procedural knowledge for human digestive system material. HOTS test instruments and Cognitive Knowledge have fulfilled validity and reliability. One of the 10 question items needs to be revised because it is included in the category of very easy items.

Acknowledgment
Researchers want to say thank you to SMAN 5 Yogyakarta, SMAN 6 Yogyakarta, SMAN 7 Yogyakarta, SMAN 8 Yogyakarta, dan SMAN 10 Yogyakarta, for supporting this research.

References
[1] Muhlisin A, Susilo H, Amin M and Rohman F 2016 APFSLT 17 2
[2] Wagner T 2008 The Global Achievement Gap (New York: Basic Books)
[3] Fadel C K and Trilling B 2009 21st Century Skills: Learning for Life in Our Times (San Francisco: Jossey-Bass)
[4] Greenstein L 2012 Assessing 21st Century Skill, A Guide to Evaluating Mastery and Authentic Learning (USA: Corwin A Sage Company)
[5] Tan A 2000 The J. of Creative Behavior 34 260
[6] Moseley D, Baumfield V M, Elliott J G, Gregson M, Higgins S E, Miller J and Newton D P 2005 Frameworks for Thinking: A Handbook for Teaching and Learning (New York: Cambridge University Press)
[7] Barak M, Ben-Chaim D and Uri Z 2007 Res. Sci. Educ. 37 354
[8] Brookhart S M 2010 How to Assess Higher-order Thinking Skills in Your Classroom (Alexandria,VA: ASCD)
[9] King F J, Goodson L and Rohani F 2010 Assessment & Evaluation Educational Services Program: Higher-order Thinking Skills (Washington, DC: A publication of the Educational Services Program)
[10] Bagarukayo E, Weide T, Mbarika V and Kim M 2012 IJEDICT 8 120
[11] Conklin W 2012 Higher Order Thinking Skills to Develop 21st Century Learners (Huntington: Shell Education Publishing, Inc)
[12] Hargrove R A 2013 Int. J. Technol. Des. Educ. 23 490
[13] Anderson L W and Krathwohl D R 2001 A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom’s Taxonomy of Education Objectives (New York: Wesley Longman, Inc)
[14] Amirulloh D, Rustaman N and Sriyati S 2014 Formica Educ. Online 1 1
[15] Haryati M 2020 J. Educ. & Development 8 92-93
[16] Zehlia A, Luzyawati L and Hamidah I 2019 Gema Wiralodra 10 168
[17] Paidi, Djukri, Yulaikah S and Alfindasari D 2017 IJSE 12 1719
[18] Oriondo L L and Dallo-Antonio E M 2008 Evaluation of Educational Outcomes (Manila: Rex Printing Company, Inc)
[19] Wilson M 2005 Constructing Measures: An Item Response Modeling Approach (Mahwah: Lawrence Erlbaum Associates, Inc)
[20] Bond T G and Fox C M 2007 Applying the Rasch Model: Fundamental Measurement in The Human Sciences 2nd ed. (Mahwah, New Jersey: Lawrence Erlbaum Associates)
[21] Adams R J and Kho S 1996 Acer Quest Version 2.1. (Camberwell, Victoria: The Australian Council for Educational Research)
[22] Wright B D and Masters G N 1999 Rating Scale Analysis (Chicago, IL: MESA Pres)
[23] Hambleton R K 1991 Fundamentals of Item Response Theory (Newbury Park: Sage Publications)
[24] Krishnan S and Idris N 2018 IJERE 7 313
[25] Khaerudin 2015 J. Madaniyah 2 214
[26] Sumintono B and Widhiarso W 2014 Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial (edisi revisi) (Bandung: Tim Komunikata Publishing House)
[27] Istiyono E, Mardapi D and Suparno 2014 JPEP 18 10