Developing critical thinking skills assessment for pre-service elementary school teacher about the basic concept of science: validity and reliability

S Patonah1 *, S B Rahardjo2, Cari2, Sajidan2

1Doctoral Program Natural Science Teacher Education, Universitas Sebelas Maret, Surakarta, Indonesia
2Doctoral Program Natural Science Teacher Education, Universitas Sebelas Maret, Surakarta, Indonesia

*Corresponding author: sitifatonah@upgris.ac.id, ¶sentotbr@staff.uns.ac.id

Abstract. One of learning outcomes in higher education is the ability to apply logical, innovative, and critical thinking in the science and technology learning. But the instruments to measure it are still few, so this study was aimed to develop critical thinking skills instruments for elementary school pre-service teachers. The mixed-method was used in this research. It consisted of 3 stages, namely, 1) the initial study, 2) design and review, and 3) try out. 122 students were selected as samples. The instrument consisted of 20 items about the Basic Science Concept Development. The items developed according to the indicators of critical thinking skills by Facione. The results showed that all items have a fit in the range of 0.77 - 1.20; mean INFIT MNSQ of 0.99, SD of 0.11; and internal consistency of 0.58. Overall, all items were declared valid and reliable.

1. Introduction

According to guidelines for the preparation of higher education curricula in Indonesia [1], that learning outcomes for high school students comprised four aspects, knowledge, attitudes, general skills, and special skills. One of the general skills to be mastered by students is to “apply logical, critical, systematic, and innovative thinking in the developing or implementing science and technology. And pays attention to and applies humanist principles in their fields of expertise.” Critical thinking becomes a crucial skill for all people who pursue tertiary education in the 21st-century era [2-4].

For the prospective teacher, Borja II [5] mentions nine characteristics to be mastered by them: a) the adapter, b) the communicator, c) the learner, d) visionary, e) the leader, f) the model, g) the collaborator, and h) the risk-taker. To carry out their duties, the prospective teachers must be able to adapt to the development of the era, adapt to the situation of their students, and adapt to their colleagues. Communication becomes an art that must be mastered by prospective teachers. Good communication facilitates the success of the learning process inside and outside the classroom. To adapt and communicate well, the prospective teachers must instil in themselves as the learner all the
time. The learning process can be conducted not only in the classroom but can throughout various activities, the prospective teacher has the ability and willingness to continue learning. The prospective teacher must also have an advanced view, ready to be a leader, can be the role model, can work together and dare to take the risks. Those nine characteristics can be obtained and trained by promoting critical thinking skills in the classroom [6-7].

Facione [9] has divided critical thinking skills into six aspects: interpretation, analysis, inference, evaluation, explanation, and self-regulation. Interpretation is the ability to understand and express the meaning of various experiences, situations, data, events, judgments, rules, procedures, and criteria. From the various variables being studied, students can express the meaning of each data got or generated during the learning process. The analysis is the ability to identify inferential relationships between the variables. Inference relates to the ability to identify and determine the elements needed to draw reasonable conclusions. Students can infer the events by paying attention to the pattern of events and then predict what events that have or will occur. Evaluation is the ability to judge things. The evaluation can be done by assessing their work and the other which believed to be a tentative truth. The explanation is the ability to explain reasoning based on conceptual, procedural, methodological considerations, and to convey reasoning as convincing arguments. Students are guided to have full confidence in conveying the truth which they believed to be based on a strong foundation. The last is self-regulation, the skill to monitor their cognitive actions by applying various skills in the previous phase.

Various instruments to assess or improve critical thinking skills have been produced, and many subject matters from each field of science has led to explosive and dynamic development. Those instruments included: the California Critical Thinking Dispositions Inventory (CCTDI); the Collegiate Assessment of Academic Proficiency (CAAP); the Cornell Critical Thinking Test (CCTT), the Critical Thinking Assessment Battery (CTAB); and the California Critical Thinking Skill Test (CCTST) written by Sumarni [5]. Those tests were intended to assess critical thinking skills in general settings. Some subject-specific instruments have been developed, such as Bellugi [5] has developed critical thinking assessment instruments for Civic Education. Kong [10] has developed instruments for junior secondary school which encompass nine different topics. Mabruroh [11] has developed the instrument of critical thinking skills for Acid and Base concept in high school. Meanwhile, the instruments to assess critical thinking skills for prospective elementary school teachers was still low in number. Therefore, this article will discuss the development of instruments to assess critical thinking skills for prospective elementary school which encompass four themes: sound, light, electricity, and magnetism.

2. Methods

2.1. Procedure
The study was conducted in the second semester of 2018/2019 academic year at 2 universities: PGRI University of Semarang (UPGRIS) and Sebelas Maret University (UNS). 122 students were participated to test the instrument with 20 questions. The instrument was developed using 4D type development model which consists of Define, Design, Develop, and Disseminate.

2.2. Participant
This research was conducted on the pre-service elementary school teachers, pre-service physics teachers, and pre-service science education teacher. The participants have enrolled at the semesters 4 and 6 programs at PGRI University of Semarang and Universitas Sebelas Maret 2018/2019. The demography of participants were shown at Table 1.
### Table 1. Participants’ profiles

| Demography | Category       | Number of participants (T=122) |
|------------|----------------|---------------------------------|
| Department of Education | Elementary School | 45                             |
|            | Physic         | 21                             |
|            | Natural Science| 63                             |
| Sex        | Male           | 17                             |
|            | Female         | 112                            |

2.3. Data Analysis

Data were analyzed through two stages: the expert judgment (content validation), and instruments testing. For expert judgment, it uses Aiken’s Index obtained from five validators with doctoral degrees in Physics, Biology, Science, and Chemistry. The Aiken Validity index was obtained using following formula:

\[
V = \frac{\Sigma s}{n (c-1)}
\]  

Where the V was the agreement index regarding item validity; s was the score for each validator minus the lowest score in the category, n was the number of validators, and c was the number of categories chosen by the validators. The results were categorized as following: 1 = not relevant; 2 = less relevant; 3 = quite relevant; 4 = relevant; and 5 = very relevant. The Aiken index ranges from 0 to 1. If \( V \leq 0.4 \) the validity was lacking, if \( 0.4 < V \leq 0.8 \) the validity was moderate, and if \( V > 0.8 \) was very valid [13]. The revisions were made based on the results of the validation from the experts, and the instrument entered the trial phase. The trial results were analyzed using the Quest program to obtain different reliability, difficulty levels, and power. The categories of questions were determined based on the degree of difficulty and distinguishing powers as shown in Table 2.

### Table 2. Point biserial and b category

| Pt-biserial | Category | b       | Category |
|-------------|----------|---------|----------|
| 0.30 ≤ pt-biserial ≤ 0.70 | Good    | <0.30   | Hard     |
| 0.20 ≤ pt-biserial ≤ 0.29 | Moderate | 0.3 − 0.70 | Medium   |
| Pt-biserial ≤ 0.70     | Poor    | 0.70    | Easy     |

3. Results and Discussion

3.1. Instrument development process

What kind of and how the instrument to be developed was determined at the Define phase. It included need analysis and supportive literature studies. Need analysis showed the students were still lacked mastering critical thinking skills, especially the inference skill. The distribution of natural science material in lectures also analyzed in this phase. It included identifying the themes and expected learning outcomes. The literature study showed the limitations of the concept of critical thinking skills developed by Facione [9]. Critical thinking skills were divided into six indicators: interpretation, analysis, inference, explanation, evaluation, and self-regulation. Of those indicators, for each topic, only five indicators were developed in which each indicator included self-regulation. The instrument was a multiple-choice test with four choices followed by reasons for choosing answers and attitudes taken after explaining the reasons. The instrument to be developed, indicators corresponding to the themes, the rubrics, the specified instrument, the answer key, the answer rubric, and the work instructions were selected and constructed in the Design phase. In the Development phase, the
instrument was validated and revised using the reviews and suggestions from experts. The instrument then tested for its content validate by testing it to 112 students. It was used to determine the validity of the items and their reliability. The final stage, *Disseminate*, was carried out by publishing the results in the seminar forum or journal.

3.2. **Expert Validation**

Expert validation experts included three items for each question: topic, construct, and language. The results were shown in Table 3. It showed the twenty questions have V Aiken value in the range 0.6 - 1, so all items were valid.

| Item | a* | b* | c* |
|------|----|----|----|
| 1    | 0.85 | 0.9  | 0.9  | 0.85 | 0.9  | 0.85 | 0.85 | 0.9  |
| 2    | 0.85 | 0.9  | 0.9  | 0.85 | 0.9  | 0.85 | 0.85 | 0.85 |
| 3    | 0.95 | 1    | 0.95 | 1    | 1    | 0.95 | 1    | 1    |
| 4    | 0.85 | 0.8  | 0.85 | 0.8  | 0.85 | 0.85 | 0.85 | 0.85 |
| 5    | 0.85 | 0.95 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 |
| 6    | 0.85 | 0.9  | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.9  |
| 7    | 0.6  | 0.65 | 0.7  | 0.75 | 0.75 | 0.75 | 0.8  | 0.8  | 0.8  |
| 8    | 0.8  | 0.8  | 0.85 | 0.85 | 0.8  | 0.8  | 0.8  | 0.8  | 0.8  |
| 9    | 0.9  | 0.9  | 0.9  | 1    | 1    | 1    | 1    | 1    | 1    |
| 10   | 0.75 | 0.75 | 0.75 | 0.8  | 0.75 | 0.75 | 0.75 | 0.8  | 0.8  |
| 11   | 0.8  | 0.95 | 0.95 | 0.9  | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |
| 12   | 0.85 | 0.9  | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.9  | 0.9  |
| 13   | 0.8  | 0.9  | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 |
| 14   | 0.85 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.9  | 0.9  |
| 15   | 0.8  | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 |
| 16   | 0.75 | 0.95 | 0.9  | 0.9  | 0.95 | 0.95 | 0.95 | 0.9  | 0.9  |
| 17   | 0.85 | 0.8  | 0.75 | 0.75 | 0.8  | 0.8  | 0.8  | 0.8  | 0.85 |
| 18   | 0.9  | 0.95 | 0.9  | 0.95 | 0.95 | 0.95 | 0.95 | 0.9  | 0.9  |
| 19   | 0.9  | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.9  | 0.95 |
| 20   | 0.9  | 0.95 | 0.9  | 0.85 | 0.85 | 0.85 | 0.85 | 0.9  | 0.95 | 0.95 |

*Note: a: Topic; b: Construct; c: Language

Overall, all items have high validity for each aspect. However several items should be revised, including the length of the answer choices should be equated (for question no. 1) and the picture in the questions should be clarified (Q5, Q7).

3.3. **Quest Program Analysis**

The items were tested on 122 students in semesters 4 and 6 of physics teacher education (UPGRIS), science teacher education (UNS), and elementary teacher education (UPGRIS). The Quest
analysis obtained Reliability of items = 0.95; Reliability of topics= 0.41; MNSQ infit of 0.99 with SD 0.11; and internal consistency of 0.58. It showed the item reliability was rated in high category but for sample reliability was still relatively low. The sample may be still working at random instruments, so information is incomplete.

The analysis also presented a map of the relationship between items and participants. The right-hand side showed the item number while the left-hand side represented the distribution of participants (Figure 1). It showed no questions classified as difficult so all questions can be answered by the participants.

![Figure 1](image_url). Distribution of items based on models fit.

In addition to showing the reliability, distinguishing power, and difficulty of the questions, the Quest analysis also displayed the distribution of items based on model compatibility. The parameter used in the data was the acceptance values for MNSQ INFIT were between 0.77 - 1.30. Figure 2 showed all items were in the range of that value, so all items were declared to be fit to the model.

Training and practicing critical thinking skills cannot be done instantaneously. It required a long time. However, various instruments can be used as a means to get students accustomed to solving their daily problems [13-15]. The results of the item analysis were shown in Table 4.

Table 2 showed 4 items (8,9,15,16) or 20% were classified as bad questions, 3 items (2,10,12) or 15% classified as moderate questions, and 13 items (1,3,4,5,6,7,11,13,14,16,17,18,19,20) or 65%, classified as good quality questions. Based on the level of difficulty, the questions consisted of 2 categories: medium (50%) and easy (50%).
**Table 4.** Point biserial and b score

| Item | b  | Judgment | Pt-bis | Judgment | Item | b  | Judgment | Pt-bis | Judgment |
|------|----|----------|--------|----------|------|----|----------|--------|----------|
| 1    | 0.92 | Easy     | 0.43   | Good     | 11   | 0.69 | Medium   | 0.37   | Good     |
| 2    | 0.63 | Medium   | 0.22   | Moderate | 12   | 0.83 | Easy     | 0.21   | Moderate |
| 3    | 0.84 | Easy     | 0.36   | Good     | 13   | 0.69 | Medium   | 0.49   | Good     |
| 4    | 0.94 | Easy     | 0.41   | Good     | 14   | 0.96 | Easy     | 0.49   | Good     |
| 5    | 0.94 | Easy     | 0.31   | Good     | 15   | 0.34 | Medium   | 0.14   | Poor     |
| 6    | 0.91 | Easy     | 0.30   | Good     | 16   | 0.92 | Easy     | 0.11   | Poor     |
| 7    | 0.44 | Medium   | 0.51   | Good     | 17   | 0.70 | Medium   | 0.33   | Good     |
| 8    | 0.57 | Medium   | 0.16   | Poor     | 18   | 0.56 | Medium   | 0.46   | Good     |
| 9    | 0.90 | Easy     | 0.11   | Poor     | 19   | 0.61 | Medium   | 0.37   | Good     |
| 10   | 0.95 | Easy     | 0.28   | Moderate | 20   | 0.42 | Medium   | 0.23   | Good     |

**Figure 2.** Map distribution of items and participants

4. **Conclusion**

The results showed the items have high Aiken validity ranging from 0.6-1. Internal consistency (Alpha) reliability was 0.58. All items have MNSQ infill values of 0.7 - 1.30 which means the instrument matches the model fit.
5. Acknowledgment

The researchers would like to express their gratitude to the Rector of PGRI University of Semarang, Indonesia, through Institute for Research and Community Services for the motivation, suggestions, research infrastructure and facilities, and research grants. The gratitude is also addressed to LPDP who provided funding support by Doctoral Dissertation Funding Program under contract numbers: PRJ-21/LPDP.4/2019, and Universitas Sebelas Maret, Indonesia for the opportunity to conduct the research.

References
[1] D. P. dan Kemahasiswaan 2019 Panduan Penyusunan Kurikulum Pendidikan Tinggi di Era Industri 4.0.
[2] Ahrari S, Samah B A, Bin Hassan M S H, Wahat N W A, and Zaremohzzabieh Z 2016 Think. Ski. Creat. 22 121
[3] Ab Kadir M A 2017 Think. Ski. Creat. 23 79
[4] Durak H Y 2018 Technol. Knowl. Learn. 25 179
[5] Borja RE 2018 Afr. Educ. Res. J. 6 160
[6] Tiruneh D T, Gu X, De Cock M, and Elen J 2018 Int. J. Educ. Res. 87 1
[7] Djamas D and Tinedi F 2018 Int. J. Inf. Commun. Technol. Educ. 14 66
[8] Dowd J E, Thompson Jr R J, Schiff L A and Reynolds J A 2018 CBE-Life Sciences Education 17 ar4
[9] Facione P A 2011 Insight assess 11
[10] Kong S C 2015 Comput. Educ. 89 16
[11] Sumarni W, Supardi K I, and Widiarti N 2018 IOP Conference Series: Materials Science and Engineering 349 012066
[12] Mabruroh F and Suhandi A 2017 Journal of Physics: Conference Series 812 012056
[13] Retnawati H 2016 Validitas Reliabilitas & Karakteristik Butir (Yogyakarta: Parama Publishing)
[14] Cargas S, Williams S, and Rosenberg M 2017 Think. Ski. Creat. 26 24
[15] Mcclellan S and Mcclellan S 2016 Coll. Undergrad. Libr. 23 295
[16] Susilowati, Sajidan, and Ramli M 2018 J. Penelit. dan Eval. Pendidik. 22 49