Validating the Assessment for Measuring Indonesian Secondary School Students Performance in Ecology

A Rachmatullah¹, F Roshayanti² and M Ha*¹

¹Division of Science Education, College of Education, Kangwon National University, Chuncheon-si, Republic of Korea
²Department of Biology Education, Faculty of Mathematics, Science and Information Technology Education, Universitas PGRI Semarang, Semarang, Indonesia

*arifraach@gmail.com

Abstract. The aims of this current study are validating the American Association for the Advancement of Science (AAAS) Ecology assessment and examining the performance of Indonesian secondary school students on the assessment. A total of 611 Indonesian secondary school students (218 middle school students and 393 high school students) participated in the study. Forty-five items of AAAS assessment in the topic of Interdependence in Ecosystems were divided into two versions which every version has 21 similar items. Linking item method was used as the method to combine those two versions of assessment and further Rasch analyses were utilized to validate the instrument. Independent sample t-test was also run to compare the performance of Indonesian students and American students based on the mean of item difficulty. We found that from the total of 45 items, three items were identified as misfitting items. Later on, we also found that both Indonesian middle and high school students were significantly lower performance with very large and medium effect size compared to American students. We will discuss our findings in the regard of validation issue and the connection to Indonesian student’s science literacy.

1. Introduction

Ecology is one branch in Biological science that addresses the inter-correlation between all the living and non-living beings exist in the Earth or frankly could be called as the study about ecosystems. Besides that the idea of ecosystems is one out of four core life sciences ideas that is necessary to be delivered to children in the school levels, from elementary to high school [1]. In addition Barman & Mayer [2] in their study found out that most of Biology teachers think that ecosystems concept and other related-concepts are important topics being learned by students because it provides the inter-connection of science discipline. Thus, students could learn the integration of theories while learning about ecosystems that can establish the so called meaningful learning. There are several sub-concepts addressed when learning about ecology and ecosystems, but the most ardently-discussed topic is the topic about interdependence of ecosystems. Interdependence of ecosystem is the topic in ecology whose discussions are related to how every component in ecosystems interact and influence one another. This discussion is needed to be delivered to student because it is related to the problems that are being faced by the degree of biodiversity in this world, particularly in regard to the loss of biodiversity due to habitat destruction [3]. The loss of biodiversity may be impacted by the loss of components of ecosystems that could be assumed as the results of human activities, such as exploiting and even hunting animal in the exaggerated ways in order to fulfill their economy or even just for their...
enjoyment [4]. The loss of biodiversity is strongly related to the equilibrium of life in the Earth which directly or indirectly may influence human in the later time. By having more knowledge about this issue of interdependence, students are expected to have wiser and better attitude towards the ecosystems and environment. Given that Indonesia is known as the second most biodiverse country in the world after Brazil, thus it is important to keep the richness of flora and fauna in Indonesia by educating youth to prevent the loss of biodiversity.

In addition, having more knowledge about interdependence in ecosystems that influence environmental attitude and behavior is one of the ways becoming the scientifically literate society. Because as what mentioned by Bybee, McCrae & Laurie about becoming scientifically literate person, student needs to be able to implement what they know about particular knowledges in their daily life [5]. The issue on producing the scientifically literate society is emphasized in the newly published Indonesian Curriculum and become one of the goals of the curriculum. It responses to the internationally published results about Indonesian student performance in Trends in International Mathematics and Science Study (TIMSS) and The Programme for International Student Assessment (PISA), in the science literacy sections that Indonesian students are always placed in the bottom rank [6]. Thus, the newly implemented Curriculum 2013 attempts to increase the achievement of Indonesian students in those international assessment programs [7].

Assessment is one of the important aspects on tackling the issue of science literacy and on finding out how far Indonesian student’s science literacy is increased. Heretofore, there are few studies which developed the assessment to measure student’s performance on the concept of interdependence on ecosystems and one of them is American Association for the Advancement of Science (AAAS). AAAS has developed one assessment in the topic of interdependence of ecosystems for American students. It has been known that the assessment is valid and reliable for American students. But, there is still no study uncovering whether the instrument is also valid for students in other countries, especially for Indonesian students, given that there is still no study examining how well Indonesian students on doing the other internationally published assessment. Therefore in this current study we are focused to examine the fitness of AAAS Ecology assessment for Indonesian secondary school students, the interaction effect between gender and grade on Indonesian secondary students’ ecology performance and also the differences between Indonesian and American students on the AAAS Ecology assessment.

2. Method

2.1. Participants

The participants joining this study consisted of 611 Indonesian secondary students. The secondary school level in Indonesia is divided into two levels which are middle and high school students. Among the total sample of participants, 218 students were middle school students and 393 students were high school students. They were in their second grade, 8th and 11th grade. In terms of gender, the participants were made up by 36% male students and 64% female students. They were recruited from two private and two public schools in West Java province.

2.2. Research Instrument

The instrument administered to students was the Ecology assessment that developed by American Association for the Advancement of Science (AAAS) and publicly published in their website (http://assessment.aaas.org/topics/IE/). The Ecology topic that addressed in the assessment was about Interdependence on Ecosystems and was consisted of 45 items that divided into three key ideas of interdependence of ecosystems which are (1) food web, (2) competition in environment and (3) contributing factors on growth rate and mortality in ecosystem. The assessment was in the multiple choice form with four options in for every item.

The total of forty-five items, we believe that it would take more time and give more cognitive pressures to students if all of the items were given to them in one time. Thus, we decided applying one of the Rasch methods called linking item methods to reduce the cognitive pressures obtained by students when doing on the assessment. The linking item method is the method that partitions the
instrument into several packages whose some of the items are called as common items because those items could be found in other packages too. We divided the total of forty-five items into two packages (package A and package B). Each package was consisted of 33 items, and 21 items among those 33 items were common items/linking items. Frankly, the linking item method that utilized in our current study is depicted in Figure 1.

![Figure 1. Linking Items](image)

2.3. Data Analyses

In terms of examining the validation of AAAS Ecology assessment, first Rasch analyses through Winstep V.3.92.1 software were performed. Finding out whether the linking item methods was successfully done or not is one of the important issues to be addressed and it can be explored by performing the correlation test to item difficulties of common items in the package A and package B. The following analysis was determining the misfitting items by investigating infit and outfit mean-square (MNSQ) values. The cut-off that suggested by Linacre [8] was used to address this misfitting items. The cut-off for infit and outfit MNSQ that could produce the productive measurement is between 0.5 – 1.5. Therefore, item with MNSQ value beyond that range is called as misfitting item. Another consideration for addressing validity issue of the instrument is generalizability of the item or whether bias exists in the item or not or even whether some particular items could be well-answered by some particular groups, we used Differential Item Functioning (DIF) to find out this issue by testing the gender bias. The DIF contrast cut-off suggested by Boone et al. [9] was used in our current study. The cut-off is 0.64, which means that the value higher than 0.64 indicates the existence of item bias with medium magnitude. The validation of the assessment is not limited to construct validity, the issue of internal consistency is also important. Thus, we also reported item and person reliability with using the standard of interpretation that suggested by DeVellis [10].

The raw data obtained from the student answer is in the form of categorical data which are right and wrong answer. Data in the form of categorical data cannot be directly analysed by using statistical analyses, especially parametric test, because in order to perform parametric test, the data should be in the form of interval data. Therefore, using the raw score is not a right way to run statistical test. This issue can be tackled by performing Rasch analyses. A set of value, which one for each student, can be obtained after performing Rasch analysis. The values are called as person measure that corresponds to every student raw score but the values have been converted to interval form. We used these values to perform further statistical analysis. Examining the interaction between gender and grade was done by performing analysis of variance (ANOVA) test. Answering the difference in performance between Indonesian students and American students was done by running the Independent sample t-test. The comparison was done by comparing the mean of every item achieved by Indonesian students and for American students we used the percentage of students responding correctly reported in the AAAS website. All statistical analyses were done through SPSS V.22.

3. Results and Discussions

The thematic method will be used to report the findings of our study. The findings of our study will be directly followed by the discussions. The first finding is related to the validation of AAAS Ecology assessment and the issue of misfitting items that for further analysis will be deleted. Following to the validation findings, the finding of interaction effect between gender and grade will be reported. Later, in the last section the comparison between Indonesian students and American students will also be reported and discussed.
3.1. Instrument Validation

The first findings that responses issue addressed in the first research question are related to the validation of the instrument. Aforementioned, linking item was utilized in our current study. To find out whether the linking item is working we did correlation test for the difficulty of common items in both packages. We found that the correlation of the common items in significantly correlated (p < 0.001) with the coefficient correlation 0.971. This highly correlated indicates that the difficulty level of common items are similar in both packages, thus we can combine the data by putting missing value on the data of students who just did one particular package and so for students who just did another package.

After combining the data, Rasch analysis was run to find out the whether there are some misfitting items or not and whether some items are biased or not. Based on the findings, in terms of infit MNSQ which is related to the pattern of responses, all the values are in the range of 0.5 – 1.5, which is indicating most of student pattern responses are fit with the Rasch model. In contrast, we found some misfitting items based on outfit MNSQ. Outfit MNSQ is sensitive to outliers or unpredicted responses or in short it is indicating that high achievers did not answer correctly in the items that they are expected to correctly answer those items or vice versa for the low achievers. Based on results shown in Table 3, there are three items outside the range of 0.5 – 1.5 which are item coded A07 with outfit MNSQ 2.68, item coded B07 with outfit MNSQ 1.79 and item coded B09 with outfit MNSQ 1.94.

The second issue of validating the instrument is generalizability of the instrument. As noted above, this generalizability issue is related to the existence of biases in items. This generalizability is identified by using DIF. We attempted to identify the bias issue in the group of gender, male and female. There were three items whose DIF contrast exceeded the benchmark 0.64. Those three items are C11 (0.74), B07 (0.87) and B11 (0.99). We considered the outfit value first then DIF values as additional consideration when deciding and refining the package of the instrument. The issue of outfit value is crucial because it shows how well the item can differentiate the high achiever and low achiever [9]. The DIF issue actually could be handled by analyzing the wording of the item, when the wording of the item is still not indicating inequality of particular groups, in our analysis is gender, the item could be still included in the measurement. In addition the items that are indicated the occurrence of bias mentioned above, the DIF contrast values are still near the cut-off, thus we still considered to include the items. Therefore, we just deleted three misfitting items based on outfit MNSQ values for further analyses. Another thing related to validating instrument, that is not less important, is reliability or internal consistency. There are two values of reliability computed through Rasch analysis, item and person reliability. Those reliabilities can also be interpreted by using common reliability interpretation benchmark [11]. The item reliability before and after deleted the three misfitting items are similar, which is 0.98. This value of item reliability is categorized as the ‘excellent’ reliability [12], which indicates the difficulty level’s probability of every item would remain same if the instrument used in the different population [9]. Therefore, the instrument shows not being dependent on respondents. Similar to the finding from item reliability, the person reliability values are also similar between overall and refined version. The person reliability values are 0.83. Based on DeVellis [10], the coefficient reliability above 0.8 indicates a “very good” instrument. In addition, based on interpretation from Fisher, the value between 0.81 – 0.90 indicates as “good instrument” that can differentiate participants into three categories based on their ability [12].

3.2. Indonesian Secondary Student’s Performance

In response to the second research question addressing the effect of interaction of gender and grade on the performance, ANOVA test was performed. Based on ANOVA test we found that in the pool of middle school students, male sample (M = –0.533, SD = 0.887) has lower performance than female one (M = –0.285, SD = 0.964). Contrarily, in the pool of high school student, male sample (M = 0.628, SD = 1.323) has higher mean than female students (M = 0.467, SD = 1.082). In addition, we did not find the significant effect of gender on the performance (F[1, 561] = 0.20, p > 0.05, η² = 0.000), while we found significant effect with large effect size (F[1, 561] = 94.94, p < 0.001, η² = 0.145).
Ultimately, we also found that the interaction effect of gender and grade was statistically significant on the performance \((F[1, 561] = 4.34, p < 0.05, \eta^2 = 0.008)\).

Further interpretation of the findings, we assume that the performance on ecology assessment does not depend on gender, but it depends on educational level. It is shown in the Figure 2 that most of middle school student’s ability is below 0 (zero) point, which indicates that most of them just could correctly answer less than 50\% of total items. In contrast, there are less high school students who are below zero point zone and most of them are above the zero point zone. This indicates that many high school students could correctly answer the questions more that 50\% of total items. It aligns with the theory of learning progression whereby the higher level of education has higher level of performance. Another interesting finding is the interaction effect between gender and grade, as shown in Figure 3a, the mean of male students in middle school level was lower than female students while in the high school level the mean of male students was higher than female student. The reason of this finding is still unclear, given that we do not have any interview. Thus, further study is needed to scrutinize this finding.

![Wright Map Indonesian Middle and High School Students](image)

**Figure 2.** Wright Map Indonesian Middle and High School Students

### 3.3. Comparison with American Students

The instrument was originally used for American students, thus here we attempted to compare the performance of Indonesian secondary students to American secondary students. The comparison is based on the mean of traditional difficulty level. Higher mean indicates that students have higher performance and lower mean indicates having lower performance. Based on our analysis, we found that in the pool of middle school student, American sample has higher mean \((M = 0.585, SD = 0.128)\)
than Indonesian sample ($M = 0.424$, $SD = 0.127$) and the difference was statistically significant with very large effect size ($t = 5.784$, $p < 0.001$, $d = 1.28$). Similar to the finding from middle school level, we found American high school students ($M = 0.674$, $SD = 0.128$) also have higher mean than Indonesian high school student ($M = 0.595$, $SD = 0.174$). The difference was also statistically significant with medium effect size ($t = 2.355$, $p < 0.05$, $d = 0.52$). The results are visualized in the Figure 3b.

The AAAS Ecology assessment addresses the questions in the type of high order thinking (HOT) where the answer is not beyond the article of the question instead the answer is in the text itself [13]. The significant lower performance of Indonesian students compared to American students indicates that Indonesian students are still lack of using their high order thinking ability or it could be caused by the lack of emphasizing high order thinking in biology classes. It is particularly for middle school students, given that the effect size shows ‘very large’ differences. Besides high order thinking ability, the questions of the assessment also address system thinking, where students are asked to investigate the effect of one component if the particular component is added or removed from the system of ecosystems. The low performance of Indonesian students indicated that they are lack of thinking a system which later impacted their ability to precisely predict. These high order thinking and system thinking are the necessary ability to be scientifically literate students [14]. Therefore, in response to the low performance of Indonesian students in science literacy assessment that reported by PISA and TIMSS, in order to increase the performance we believe that the teaching activities with emphasizing the higher order thinking and system thinking ability is suggested to be used in Indonesian science classroom, especially biology classrooms. Several teaching methods and approaches that cover those kinds of ability are problem based learning and project based learning.

![Figure 3](image)

**Figure 3.** (a)The interaction effect between grade and gender on AAAS Ecology assessment, (b) the comparison between Indonesian and American secondary school students

### 4. Conclusion

Based on our current study, we found that most of questions in AAAS Ecology assessment in the topic of interdependence in ecosystems are psychometrically valid to be administered to Indonesian middle and high school students. Besides that we also found that educational level give more effect on students’ performance on this assessment than gender. We found that Indonesian secondary students were significantly lower performance than American students. These findings could be used to response the longstanding issue of Indonesian student’s science literacy, that always place in the bottom rank. We believe that more assessment and teaching and learning activities that employ and address higher order thinking and system thinking are needed to be implemented in Indonesian science and or biology instructions.

### References

[1] National Research Council 2012 *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (Washington, DC: The National Academies Press)

[2] Barman CR and Mayer DA 1994 *The American Biology Teacher* 56 160-3.

[3] Coleman FC and Williams SL 2002 *Trends in Ecology & Evolution* 17 40-4.
[4] Allesina S and Bodini 2004 *Journal of Theoretical Biology* **230** 351-8.
[5] Bybee R, McCrae B and Laurie R 2009 *Journal of Research in Science Teaching* **46** 865-83.
[6] Rachmatullah A, Octavianda RP, Ha M, Rustaman NY and Diana S 2017 *IOP Journal of Physics: Conf. Series* **812** 012009
[7] Mendikbud [Indonesian Ministry of Education and Culture] 2016 Salinan Lampiran Permendikbud No.21 th 2016 tentang Standar Isi Pendidikan Dasar dan Menengah [A copy of the additional documents of Indonesian ministry of education and culture regulation number 21 year 2016 on Content Standards for Primary and Secondary Education].
[8] Linacre JM 2002 *Rasch Measurement Transactions* **16** 878
[9] Boone WJ, Staver JR and Yale MS 2014 *Rasch Analysis in The Human Sciences* (Dordrecht: Springer)
[10] DeVellis RF 2003 *Scale Development: Theory and Applications 2nd edn* (Thousand Oaks,CA: Sage)
[11] Fisher WP Jr 2008 *Rasch Measurement Transactions* **22** 160-3
[12] Fisher WP Jr 2007 *Rasch Measurement Transactions* **21** 1095
[13] Zohar A and Dori YJ 2003 *The Journal of the Learning Sciences* **12** 145-181.
[14] Norris SP and Phillips LM 2003 *Science Education* **87** 224-40.