JPII 10 (2) (2021) 270-281

Jurnal Pendidikan IPA Indonesia

http://journal.unnes.ac.id/index.php/jpii

DETERMINATION OF GENDER DIFFERENTIAL ITEM FUNCTIONING IN TEGAL-STUDENTS' SCIENTIFIC LITERACY SKILLS WITH INTEGRATED SCIENCE (SLiSIS) TEST USING RASCH MODEL

P. Susongko*[1], Y. Arfiani[2], M. Kusuma[3]

[1,2,3]Universitas Pancasakti Tegal, Indonesia

DOI: 10.15294/jpii.v10i2.26775

Accepted: October 21st 2020. Approved: June 29th 2021. Published: June 30th 2021

ABSTRACT

The emergence of Differential Item Functioning (DIF) indicates an external bias in an item. This study aims to identify items at Scientific Literacy Skills with Integrated Science (SLiSIS) test that experience DIF based on gender. Moreover, it is analyzed the emergence of DIF, especially related to the test construct measured, and concluded on how far the validity of the SLiSIS test from the construct validity of consequential type. The study was conducted with a quantitative approach by using a survey or non-experimental methods. The samples of this study were the responses of the SLiSIS test taken from 310 eleventh-grade high school students in the science program from SMA 2 and SMA 3 Tegal. The DIF analysis technique used Wald Test with the Rasch model. From the findings, eight items contained DIF in a 95 % level of trust. In 99 % level of trust, three items contained DIF, items 1, 6, and 38 or 7%. The DIF is caused by differences in test-takers ability following the measured construct, so it is not a test bias. Thus, the emergence of DIF on SLiSIS test items does not threaten the construct validity of the consequential type.

Keywords: gender; Differential Item Functioning (DIF); SLiSIS test

INTRODUCTION

Citizen science literacy has a very significant effect on the progress of a Nation. The reason is that the scientific literacy of the community has a positive effect on the quality of economic development, democracy, and community culture (Hanushek & Woessmann, 2016; Roth & Lee, 2016; Rudolph & Horibe, 2016). Therefore, students' scientific literacy must be the main goal in science education (McFarlane, 2013). Therefore, many attempts were made to increase students' scientific literacy in science learning by developing science learning models and assessments (Ardianto & Rubini, 2016; Rusilowati et al., 2016; Ratini et al., 2018; Fakhriyah et al., 2019).

In Indonesia, science education in high school aims to: (1) build and apply information, knowledge, and technology logically, critically, creatively, and innovatively; (2) demonstrate the ability to think logically, critically, creatively, innovatively, and independently; (3) demonstrate the ability to analyze and solve complex problems; (4) demonstrate the ability to analyze natural phenomena, use the environment productively and responsibly, also master the knowledge needed for higher education (Ministry of Education and Culture of the Republic of Indonesia, 2018). The purpose is in line with the scientific literacy skills developed by the 2015 PISA (Program for International Science Student Assessment), which includes: (1) explaining phenomena scientifically; (2) evaluating and designing scientific investigations; (3) interpreting scientific data and evidence (Chiang & Tzou, 2018).

The competency standards set by the Government before 2020 were measured through the National Examination. However, there are some weaknesses in the implementation of the national examination. First, the Government does not use

*Correspondence Address
E-mail: purwosusongko@gmail.com

the National Examination results as a determinant of graduation, so there is no guarantee of compliance with competency standards for high school students who graduate. Second, not all subjects that build scientific competence are tested, and students may choose just one subject. Those weaknesses cause students who pass to be not comprehensive under the competency standards that students should master. Likewise, at the International level, the Government has never taken a survey to study achievement at the high school level, so it does not have the quality parameters of high school graduates in Indonesia.

In this regard, there needs to be a comprehensive examination that can ensure the competence of high school graduates following the predetermined competency standards. Therefore, SMA 2 and SMA 3 Tegal, Central Java, have conducted integrated science literacy tests on high school students in the twelfth grade of MIPA (Mathematics and Natural Sciences) to ensure that the graduates meet national standards and the main objectives of science education at the international level. The test is called the SLiSIS test (Scientific Literacy Skills with Integrated Science Test). SLiSIS test is a standardized test tested empirically and meets content standards, scientific literacy achievements, and measurement models.

In the aspect of content, SLiSIS Test covers Mathematics, Physics, Chemistry, and Biology competencies in an integrated manner through integrated science cases. Several studies show that science learning presented in an integrated manner significantly affects increasing student scientific literacy (Turiman et al., 2012; Tamassia & Frans, 2014). In scientific literacy achievement, SLiSIS Test refers to PISA 2015 (OECD, 2016). SLiSIS Test uses 14 testlets, and each testlet consists of 3 items. In each testlet there is one scientific news and three items that measure the achievement of scientific literacy according to PISA 2015 standards. Item validation and scoring models do not use the classical test theory that has been widely used but uses Rasch modeling. The use of classical test theory only produces scores at the ordinal level, while Rasch modeling can produce scores at the interval level to meet the measurement assumptions (Mari et al., 2012; Bond & Fox, 2015; Susongko, 2016; Rusch et al., 2017). The validity with Rasch modeling refers to the validity of Messick, where construct validity is considered a single concept consisting of several aspects (Runnels, 2012; Ravand & Firoozi, 2016). Rasch's analysis explains the construct validity which is more comprehensive than classical test theory. There are at least six aspects of construct validity: content, substantive, structural, external, generalisability, and consequential aspects (Sabah et al., 2013; Wang et al., 2014; Jong et al., 2015).

The construct validity of consequential aspects is related to the desired and undesirable consequences (e.g., bias) of the assessment and the implications derived from the score's meaning. The validity of consequential aspects focuses on the implications of the interpretation of the score as information. Evidence regarding the consequential aspects of construct validity also discusses the actual and potential consequences of testing the scores used, especially in terms of sources of invalidity such as bias and fairness (Welner, 2013). The consequential aspect of construct validity is the aspect that is highly considered in the SLiSIS test. The reason is the use of SLiSIS test results used by schools in determining the graduation status of high school students of the Science program at SMA 2 and SMA 3 Tegal. SLiSIS test is a high-stakes test, so bias can be a major issue in administering the test.

Previous studies have shown a gender bias in scientific tests. For example, on the physics test and teaching physics, there is a gender bias (Hofer, 2015; Wilson et al., 2016). Gender bias is also found in chemistry and biology tests (Grunspan et al., 2016; Rachmatullah & Ha, 2019). In addition, several studies have shown a gender bias in scientific literacy tests in the PISA and TIMMS surveys (Lisova & Kovalchuk, 2017; Cheema, 2019). Bias is the emergence of several characteristics of different items in individuals with the same abilities but from different ethnic groups, genders, cultures, or religions (Rouquette et al., 2016). In other words, an item can occur if individuals who have the same ability but come from different groups do not have the same opportunity to answer an item correctly. These conditions originate from several character items or situations that are not relevant to the test objectives. Bias is a systematic error that affects the validity of test scores (Demirtasli, 2015). Items must be tested for potential contain bias to ensure the accuracy of the decision to be based on test scores. The method of determining item bias is focused on the validity of test items between different subgroups. Measurement bias can be generated from the presence of differential item functioning (DIF) in items (Rouquette et al., 2016).

DIF describes the phenomenon of one or several items from the test "functioning" differently in the group of individuals to be compared. These individuals are distinguished by characteristics such as age, gender, ethnicity, religion, country of origin (Chiang & Tzou, 2018). Statistically, that means the function parameters that link latent variables (constructs measured) with observations (responses to items) differ in the various groups involved (Kendhammer et al., 2013). If the data are analyzed using the Rasch model, the DIF phenomenon will emerge if differences in the parameters of difficulties in the groups are compared (Millsap, 2012).

Many studies show DIF caused by gender on science tests, especially on science tests that are high-stakes tests. All tests in mathematics, physics, chemistry, biology, and measurement of critical thinking ability are not free from the presence of DIF based on gender (French et al., 2012; Steinmayr et al., 2015). In international surveys such as PISA and TIMSS, DIF testing is not only done based on gender but also conducted against the cultural background, country, and level of scientific literacy ability of students (Mesic, 2012; Lyons-Thomas et al., 2014; Choi et al., 2015; Demirtasli, 2015; Huang et al., 2016; Chiang & Tzou, 2018; Cheema, 2019). Thus it is possible that the SLiSIS test items still contain DIF based on gender.

The appearance of DIF on a test item indicates an external bias in an item (Embretson & Reise, 2013). It is one of the things that threaten the validity of the test so that it can reduce the level of confidence in the score generated by the SLiSIS test. A preliminary study of the scientific literacy test using the unit testlet analysis and involving 112 test participants showed the existence of DIF based on gender as many as two items from the 17 items given. These items are related to the theme of nuclear physics and Astronomy (Susongko et al., 2019). The SLiSIS test was applied to 310 students. Therefore, it is necessary to investigate the extent of the existence of DIF on the test.

Various methods have been proposed to identify DIF, such as Mantel-Haenzel (MH), difference of difficulties, Lord's $\chi 2$, Non-compensatory DIF (NCDIF), SIBTEST, logistic regression, and others (Cuevas & Cervantes, 2012). Some statistical tests used directly to evaluate the existence of DIF in Rasch modeling are the use of Wald test, Likelihood Ratio test (LR), the score test, and Simultaneous Item Bias Test (SIBTEST). The first three used are tests using large sample sizes, the Maximum Likelihood estimation approach, and parametric assumptions (Strobl et al., 2015). Several studies compare the effectiveness of the Wald test, MH, LR test, and SIBTEST methods and show that the Wald test method is the most effective compared to others (Hou et al., 2014). Several studies show that the Wald test method with type I errors and maximum Likelihood estimates are the most sensitive compared to other variations (Woods et al., 2013; Battauz, 2019). It shows that the Wald method is the most effective in detecting LDIF in the Rasch model approach.

This study aims to identify items on the SLiSIS test that experience DIF based on gender. When the items are known, an analysis of the occurrence of the DIF is mainly related to the type of scientific literacy skills measured and concluded by concluding the validity of the SLiSIS test from the construct validity of consequential type. It is to assess bias or whether an item requires more in-depth information and study, primarily related to whether the emergence of differences in the opportunity to answer correctly under the construct measured or not. The existence of a DIF on an item is not automatically biased. However, if the difference in opportunities to answer correctly between male and female groups is caused by characters that do not fit the measurement construct, the item is called bias (He & van de Vijver, 2012).

METHODS

The study was conducted with a quantitative approach by using survey or non-experimental methods. In the design of survey methods, some trends, behaviors, or opinions of a population by examining a population sample were described quantitatively. From this sample, the researcher generalizes or makes claims about the population (Creswell & Poth, 2016). The stages of this study were compiling the SLiSIS test, conducting empirical trials, and analyzing DIF based on student responses to the test.

The SLiSIS test is a scientific literacy test that aims to measure scientific literacy skills that refer to the achievements of scientific literacy used by PISA in the 2015 survey. PISA divides scientific literacy skills into three domains, they are: (1) explaining phenomena scientifically as well as recognizing, offering, and evaluating explanations for various natural and technological phenomena; (2) interpret scientific data and evidence as well as analyze and evaluate data, claims, and arguments in various representations and draw scientific conclusions; (3) evaluating and designing scientific investigations as well as describing and evaluating scientific investigations and making generalizations from explanations

(OECD, 2016). Specifically, items that measure the skills of the Evaluate and design scientific inquiry in the SLiSIS test are focused on the skills of making generalizations from scientific explanations or investigations.

The SLiSIS test material is a brief description relating to integrated science themes or scientific news. For each scientific reading, there are three multiple-choice items with five alternative answers. Scientific reading is taken from various sources such as www.ScienceNews.org, www.sciencenewsforstudents.org, www.readwork.org, and some of the integrated science exams on college entrance selection in Indonesia. Each item sequentially measures students' ability to explain phenomena scientifically (first item), interpret data and scientific evidence (second item), also evaluate and design scientific investigations (third item). SLiSIS test consists of 14 testlets, with each testlet consisting of three items, so that the number of items is 42. Scoring each item is considered to be independent and dichotomous (1 or 0).

The samples of this study were the responses to the SLiSIS Test conducted on 310 students of senior high school grade XII of Science program from SMA 2 and SMA 3 Tegal city. The SLiSIS test for SMA 2 Tegal was held on February 11, 2020, at 08.00-10.00 WIB. Meanwhile, for SMA 3 Tegal city, the test was held on February 21, 2020, from 7.30-9.30 WIB. From 310 students, there are 102 male students and 208 female students with ages between 17 to 19 years. All students come from the area of Tegal city and surrounding areas. The distribution of students who are subject to this study can be seen in Table 1.

Table 1. Distribution of Research Subjects based on Gender

No	School Name	Class	Number of Male	Number of Female	Total
1	SMA N 2 Tegal	XII MIPA 1	10	16	26
2		XII MIPA 2	5	24	29
3		XII MIPA 3	9	24	33
4		XII MIPA 4	9	22	31
Subtotal			33	86	119
5	SMA N 3 Tegal	XII MIPA 1	9	22	31
6		XII MIPA 2	12	19	31
7		XII MIPA 3	12	21	33
8		XII MIPA 4	12	20	32
9		XII MIPA 5	12	19	31
10		XII MIPA 6	12	21	33
Subtotal			69	122	191
Total			102	208	310

The research procedure began with estimating the difficulty level of items with Rasch modeling involving all responses or only involving responses from male students or female students. The Rasch model analysis used software version R 3.5.0 through package version 0.15-6 (Mair et al., 2019). The basic Rasch modeling uses the following formula:

$$P_i(\theta) = \frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}}$$

, with:

$P_i(\theta)$: the opportunity for someone with the ability to answer the *i* item correctly.

b_i : difficulty level parameters for the *i* problem

i : 1, 2, 3,... 42

e : natural logarithm

θ : parameters of the participants' ability (Bond & Fox 2015)

The Wald test analysis was then performed to determine whether the differences in the difficulty level of items in the two groups by gender were significant. The W test was used to test the significance of the effect of the independent variable (X_i) partially on the dependent variable (Y) in the logistic regression model carried out by the Wald Test. The Wald value in the W (Wald) test uses the formula:

$$W_i = \left[\frac{\widehat{\beta_i}}{Se\widehat{\beta_i}}\right]^2 \quad ; \quad SE = \sqrt{\frac{p(1-p)}{n}}$$

The hypothesis used for the w test is:

H_0 : $\beta_i = 0$ (There is no significant effect of gender (X_i) on the level of item difficulty (Y))

H_1 : $\beta_i \neq 0$ (There is a significant influence between the gender variables (X_i) on the level of item difficulty (Y))

For *i* = 1, 2, ... , p

Criteria for decision making is:
H_0 is refused if $W_i > \chi^2_{\alpha,1}$
H_0 is accepted if $W_i > \chi^2_{\alpha.1}$; $\alpha = 0{,}05\ or\ 0.01$
(Woods et al., 2013)

Mapple software version 13 was used to draw the characteristic curve of items detected by DIF. Previously, an equation was made to equalize the scale of item difficulty of male participant responses and the scale of item difficulty of male participants responses using the linear regression method.

RESULTS AND DISCUSSION

The result of the study began by describing the results of item analysis with Rasch modeling involving all responses of SLiSIS test takers and the test participants' responses of male or female only. Moreover, Wald test analysis results on the student responses to the SLISIS test will also be included in the presentation. The data from the analysis can be seen in Table 2.

Table 2. Analysis Results on Level of Difficulty and Wald Test on SliSIS Test Items

No	Item level of Difficulty	Item Level of Difficulty (Male Respondent)	Item Level of Difficulty (Female Respondent)	Z Statistics (Wald Test)	P Value	Explanation (p=95 %)	Explanation (p=99 %)
1	-0,064	-0,563	0,162	-2,812	0,005	DIF	DIF
2	-2,414	-2,512	-2,374	-0,302	0,763	NON DIF	NON DIF
3	-1,122	-0,76	-1,309	1,955	0,051	NON DIF	NON DIF
4	-1,908	-2,114	-1,827	-0,740	0,460	NON DIF	NON DIF
5	-0,762	-0,973	-0,671	-1,088	0,276	NON DIF	NON DIF
6	0,772	0,262	1,056	-3,102	0,002	DIF	DIF
7	-0,078	0,054	-0,141	0,785	0,432	NON DIF	NON DIF
8	0,004	0,054	-0,019	0,295	0,768	NON DIF	NON DIF
9	0,291	-0,115	0,491	-2,420	0,015	DIF	NON DIF
10	1,004	1	1,006	-0,025	0,980	NON DIF	NON DIF
11	2,103	1,907	2,225	-0,922	0,356	NON DIF	NON DIF
12	-0,160	-0,202	-0,141	-0,243	0,808	NON DIF	NON DIF
13	0,669	0,471	0,773	-1,195	0,232	NON DIF	NON DIF
14	-0,683	-0,918	-0,582	-1,226	0,220	NON DIF	NON DIF
15	-0,091	-0,516	0,102	-2,407	0,016	DIF	NON DIF
16	-2,137	-1,903	-2,256	0,915	0,360	NON DIF	NON DIF
17	-1,635	-1,482	-1,709	0,691	0,490	NON DIF	NON DIF
18	0,168	0,054	0,223	-0,685	0,494	NON DIF	NON DIF
19	0,304	0,429	0,243	0,750	0,453	NON DIF	NON DIF
20	0,742	0,817	0,706	0,433	0,665	NON DIF	NON DIF
21	-0,667	-0,76	-0,626	-0,498	0,619	NON DIF	NON DIF
22	-0,078	0,346	-0,284	2,540	0,011	DIF	NON DIF
23	0,100	0,262	0,021	0,977	0,328	NON DIF	NON DIF
24	0,113	0,262	0,041	0,896	0,371	NON DIF	NON DIF
25	0,100	0,137	0,082	0,226	0,882	NON DIF	NON DIF
26	0,387	0,728	0,223	2,011	0,044	DIF	NON DIF
27	-2,034	-1,636	-2,256	0,697	0,090	NON DIF	NON DIF
28	-1,067	-1,088	-1,059	-0,101	0,919	NON DIF	NON DIF
29	1,311	1,297	1,319	-0,079	0,937	NON DIF	NON DIF
30	0,359	0,514	0,284	0,923	0,356	NON DIF	NON DIF
31	-1,492	-1,148	-1,672	1,696	0,090	NON DIF	NON DIF
32	2,473	2,323	2,566	-0,613	0,540	NON DIF	NON DIF
33	2,046	2,423	1,875	1,475	0,140	NON DIF	NON DIF
34	0,209	-0,158	0,387	-2,177	0,029	DIF	NON DIF
35	1,187	1,194	1,184	0,039	0,969	NON DIF	NON DIF
36	1,119	0,953	1,21	-0,960	0,337	NON DIF	NON DIF
37	0,526	0,599	0,491	0,430	0,667	NON DIF	NON DIF
38	0,669	1,144	0,449	2,646	0,008	DIF	DIF
39	0,222	0,221	0,233	-0,010	0,992	NON DIF	NON DIF
40	-1,318	-1,088	-1,432	1,156	0,248	NON DIF	NON DIF
41	0,787	0,514	0,934	-1,651	0,099	NON DIF	NON DIF
42	0,045	-0,03	0,082	-0,452	0,651	NON DIF	NON DIF

Table 2 shows eight items out of 42 items detected containing DIF when using the Wald test with a 95% level of trust. It can be seen that the eight items have a P value of less than 0.05. Information on the items detected by DIF can be seen in Table 3.

Table 3. List of SLiSIS Test Items Detected by DIF at 95 % Level of Trust

No	Item	Scientific Competencies	The Beneficiary
1	**1**	Explain phenomena	Male
2	**6**	Evaluate and design scientific enquiry	Male
3	9	Evaluate and design scientific enquiry	Male
4	15	Evaluate and design scientific enquiry	Male
5	22	Explain phenomena	Female
6	26	Interpret data and evidance	Female
7	34	Explain phenomena	Male
8	**38**	Interpret data and evidance	Female

Table 3 shows that out of the 14 items that measure Evaluate and Design Scientific Inquiry skills, there are three items number 6, 9, and 15 that experience DIF and consistently benefit the male test participants. Likewise, the 14 items that measure the ability to interpret data and evidence have two items: numbers 26 and 38, which experience DIF and consistently benefit the female test participants. For items that measure the explain phenomena, three items, number 1, 22, and 34, experience DIF. However, these items are not consistently seen from the groups that are benefited. For example, number 1 and number 34 benefit the male students while number 22 benefits the female students. ICC description for DIF detected items is in Figure 2, Figure 3, and Figure 4. Figure 1 explains the difference in the Item Characteristics Curve (ICC) for items detected by DIF (number 2) and those not detected by DIF (number 22).

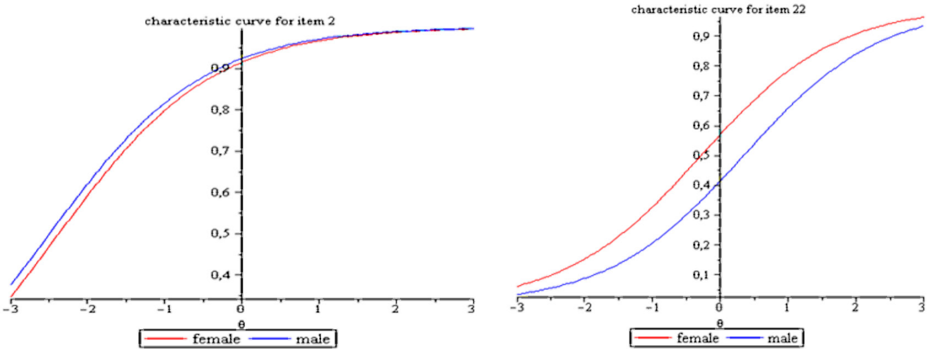


Figure 1. ICC for Items Number 2 and Number 22

In Figure 1, it can be seen that for item number 2, where DIF is not detected, the ICC for male and female test-takers coincides so that it can be concluded that the opportunity to answer correctly between groups of male and female at all levels of ability is the same. Otherwise, for DIF detected items, in item number 22, it is seen that the opportunity of the female group (in red) in answering item number 22 is higher than the chance of the male group (in blue). Figure 1 explains the difference in the Item Characteristics Curve (ICC) for items detected by DIF (number 1) and those not detected by DIF (number 34). Moreover, Figure 2 shows that in the two items that measure explain phenomena, the chance of the male group answering correctly (in blue) is higher than the chance of the female group (in red) for all ability levels. However, item number 22 is different from the two items. As shown in Figure 1, in item no 22, , the chance of the male group answering correctly is lower than the female group.

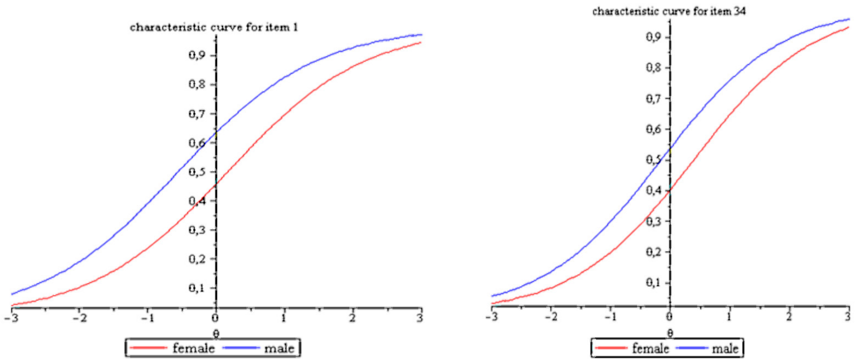


Figure 2. ICC for Item Number 1 and Number 34

Figure 3 shows that the two items that measure evaluate and design scientific inquiry show that at all levels of ability, the chance of the female group answering correctly (in red) is higher than those for the male group (in blue).

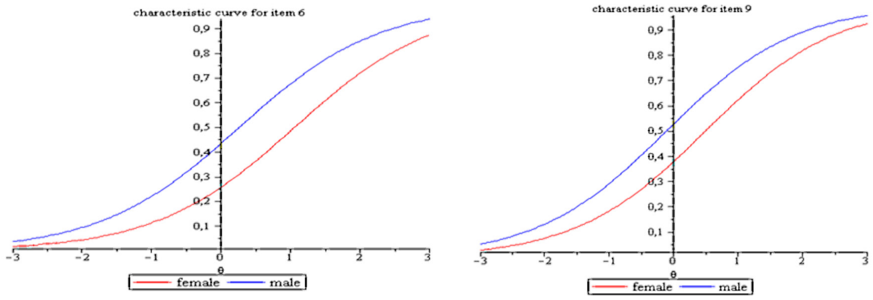


Figure 3. ICC for Item Number 6 and Number 9

Whereas Figure 4 shows that the two items that measure data interpretation and evidence at all levels of ability, the chance of the female group answering correctly (in red) is higher than those of the male group (in blue).

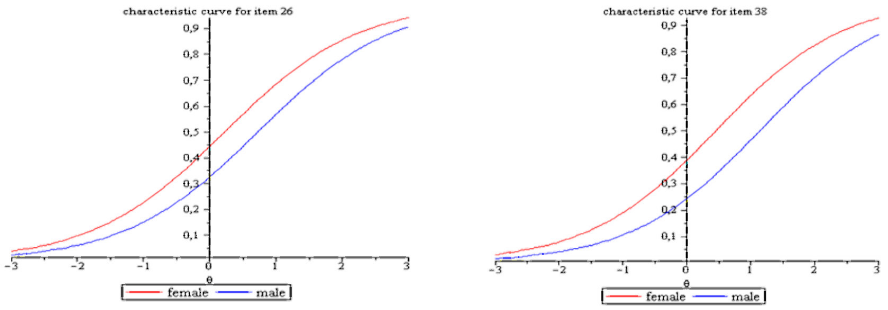


Figure 4. ICC for Item Number 26 and Number 38

When observed from eight items that contain DIF, five items benefit the male while three items benefit the female. Items that measure Evaluation and design scientific inquiry consistently benefit the male students, while items that measure the interpretation of data and evidence consistently benefit the female. The result is following some previous studies that the male students benefit a lot from the performance of science (Ganley et al., 2014; Disenhaus, 2015; Reilly et al., 2015; Lisova & Kovalchuk, 2017; Wang & Degol, 2017; Balart & Oosterveen, 2019).

The following are examples of two items that measure Evaluate and design scientific inquiry skills (numbers 6 and 9) and contain DIF.

6. Based on this research, if X = = the possibility of stopping growth or shrinking tumors in cancer patients with high fiber diets and Y = the possibility of stopping growth or shrinking tumors in cancer patients on low fiber diets, the relationship of X and Y is as follows:

A.Y=5X
B. Y=X/5
C. X=Y/5
D. Y=X
E. Y>X

9. Earth's sea surface temperature has risen by about half a degree Celsius from 1930 to 2010. Based on this data, the temperature rise in 2130 is estimated at ...

A. 0,5 °
B. I,0°
C. 1,5°
D. 2,0°
E. 2,5°

To answer both items required a strong mathematical reasoning ability. For example, at number 6, students must state the relationship between variables in mathematical equations, while at number 9, students are required to make generalizations from existing data and then make predictions. These abilities are needed to answer 14 SLiSIS test items which measure Evaluate and design scientific inquiry skills. The results of this study are following several previous DIF studies which benefit the male in measuring mathematical reasoning abilities (Coletta et al., 2012; Reilly, 2012; Stoet & Geary, 2012; Taylor & Lee, 2012; Ong et Al., 2015; Yildirim, 2019).

The following are examples of two items that measure data interpretation and evidence skills (numbers 26 and 38) and contain DIF.

26. Scientists have succeeded in giving false memories to the mouse that was electrocuted at specific locations. Which evidence from the text supports this conclusion?

A. Scientists stimulate the mouse brain areas that are activated in the first location.
B. The mouse is allowed to explore the first location calmly.
C. The mouse receives shocks in the second location.
D. The mouse is afraid of locations where they are not shocked.
E. The mouse is afraid of the second location.

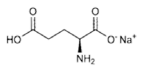
38. If the temperature variation to the height where the fragrant root plant grows is considered linear, then at an altitude of 1000 m, the place temperature is ...

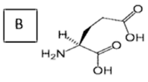
A. 22,33° C
B. 17 ° C
C. 25° C
D. 20,33° C
E. 18,67 ° C

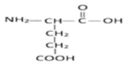
These items can be answered quickly if the test-takers pay attention to the reading in the form of scientific news given according to the item. From the accuracy in reading, then the test-takers can provide the correct interpretation. Both of these items contain DIF and are beneficial for female students. The results of this study are consistent with some previous studies where there are DIFs that benefit female in the measurement of reading ability and textual interpretation (Taylor & Lee, 2012; Hyde, 2014; Voyer& Voyer, 2014; Balart & Oosterveen, 2019).

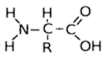
The following are examples of items that measure the skill of explaining scientifically and contain DIF phenomena. Item 1 and item 34 benefit the male students, while item 22 benefits the female students.

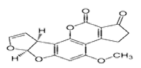
1. The Chemical formula from MSG is:

A


B


C


D


E


22. From these readings, photosynthesis is basically

A. One of the big ideas of philosophy.
B. The beginning of life on Earth.
C. One of humanity's best inventions.
D. One of the most famous scientific innovations.
E. One of the keys to all processes on Earth.

34. In the human body, ammonia occurs as a result of the breakdown of:

A. Amino acids
B. Fat
C. Carbohydrates
D. Vitamins
E. Gastric acid

When sifted through, to answer items number 1 and 34, students need the ability to think critically, not the ability to recall, while item 22 requires the interpretation ability of reading or reading ability. For example, when asked about the chemical formula of Monosodium Glutamate (MSG), students who can think critically will choose a chemical formula in sodium (Na). From the five chemical formulas, only option A contains Na, so it is easy to determine that the correct answer is option A. Similarly, item 34 is asked about the presence of ammonia in the human body. Again, if students can think critically, then it is easy to find the answer because without seeing the molecular formula, it can be ascertained that only ammonia acid has a similarity in naming with ammonia. Therefore students can answer that ammonia is a breakdown of amino acids without remembering the chemical formula of amino acids and ammonia. The ability to conclude based on the data provided is one indicator of critical thinking ability (Fisher, 2011; McPeck, 2016). The results of this study are following several previous studies, which found that there are differences in the ability to think critically based on gender (Aliakbari & Sadeghdaghighi, 2011; French et al., 2012; Harish, 2013; Preiss et al., 2013). Item 22 benefits the female students because they measure ability and accuracy in reading and interpreting. The condition corresponds as happened with items 26 and item 38.

The emergence of items that experience DIF is not necessarily a weakness of a measurement instrument. The result is at least obtained from methodological reasons as well as the reasons for the test construction. Some DIF methodology studies show that the more sample sizes used, the more items detected by DIF (Zwick, 2012). Likewise, it is found that the use of samples that are not equivalent to the presence of DIF can be undetected (Rahmawati, 2019). The DIF methodology study also shows that the non-uniform DIF and Crossing DIF are not the real DIF (Ong et al., 2015; Rouquette et al., 2016; Gómez-Benito et al., 2018). Based on consideration of these methodological aspects, it is necessary to be careful in determining the level of trust when assigning the DIF status of an item. The level of trust used should be as high as possible so that the errors for rejecting correct items are minimal. For example, when using a 99% level of trust, only three items from the SLiSIS Test experienced DIF: items 1, 6, and 38 or about 7% of all items used in the SLiSIS Test.

From the constructed test, it can be seen that mathematical reasoning, verbal reasoning, and critical thinking become part of the construct measured in the SLISIS test so that the emergence of DIF does not mean there is a test bias. Bias occurs when differences in scores on items or indicators of a particular construct do not correspond to differences in the nature or abilities of the underlying or construct measured by the test (He & van de Vijver, 2012). The score difference that emerges on the items detected biased the SLiSIS test based on abilities that become indicators of the construct measured in the SLiSIS test. Thus the consequence validity of the SLiSIS test is not affected by containing DIF based on gender in some of its items.

The main finding in this study is that there is a tendency for male students to benefit more in working on items that measure Evaluation and scientific inquiry design, while female students are more likely to benefit in working on items that measure interpretation of data and evidence. The DIF knowledge can be reflective material in learning science. It is needed to strengthen the mathematics logic ability for female students and verbal logic for male students.

CONCLUSION

Using a 95% level of trust, eight items in the SLiSIS test contain DIF based on gender. They are items 1, 6, 9, 15, 22, 26, 34, and 38 or by 19 %. While in 99% level of trust, there are three items with DIF, items 1, 6, and 38 or by 7%. The emergence of DIF on SLiSIS test items is due to differences in test takers' ability under the measured construct, so it is not a test bias. Thus the appearance of DIF on SLiSIS test items does not threaten the construct validity of consequential type. If gender differences cause the emergence of DIF, the differences are not following the construct being measured. Such a condition is called test bias and will threaten the validity of the test. In science learning for male students, it is necessary to strengthen the ability to interpret data and evidence, while for female students, it is necessary to strengthen the ability to evaluate and scientific inquiry design. In addition, in school, mathematics learning must be strengthened in mathematical logic, while language learning must be strengthened in verbal logic.

ACKNOWLEDGMENTS

Thank you to the Ministry of Education and Culture of the Republic of Indonesia for funding this research through the higher education applied research grant projects of 2019 and 2020 no 321/K/A-5/LPPM-UPS/V/2020. Hopefully, the results of this study can strengthen the competence of Science educators in Indonesia when conducting an assessment of the Indonesian students' scientific literacy competence.

REFERENCES

Aliakbari, M., & Sadeghdaghighi, A. (2011). Investigation of the relationship between gender, field of study, and critical thinking skill: the case of Iranian students. In *Proceedings of The 16th Conference of Pan-Pacific Association of Applied Linguistics*, Hongkong, 8th – 10th August 2011.

Ardianto, D., & Rubini, B. (2016). Comparison of students' scientific literacy in integrated science learning through model of guided discovery and problem based learning. *Jurnal Pendidikan IPA Indonesia*, *5*(1), 31-37.

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications*, *10*(1), 1-11.

Battauz, M. (2019). On Wald tests for differential item functioning detection. *Statistical Methods & Applications*, *28*(1), 103-118.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.*)*. Routledge.

Cheema, J. R. (2019). Cross-country gender DIF in PISA science literacy items. *European Journal of Developmental Psychology*, *16*(2), 152-166.

Chiang, P. M., & Tzou, H. I. (2018). The application of differential person functioning on the science literacy of Taiwan PISA 2015. *Humanities & Social Sciences Reviews*, *6*(1), 08-13.

Choi, Y. J., Alexeev, N., & Cohen, A. S. (2015). Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the TIMSS 2007 mathematics test. *International Journal of Testing*, *15*(3), 239-253.

Coletta, V. P., Phillips, J. A., & Steinert, J. (2012) FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects. In *AIP Conference Proceedings 1413*, Los Angeles: 09 Januari 2012.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Cuevas, M., & Cervantes, V. H. (2012). Differential item functioning detection with logistic regression. *Mathématiques et sciences humaines. Mathematics and Social Sciences*, (199), 45-59.

Demirtasli, Ç. (2015). A study on detecting of differential item functioning of PISA 2006 science literacy items in Turkish and American samples. *Eurasian Journal of Educational Research*, *58*, 41-60.

Disenhaus, N. (2015). *Boys, writing, and the literacy gender gap: what we know, what we think we know* (Dissertation). Burlington: University of Vermont

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Fakhriyah, F., Masfuah, S., & Mardapi, D. (2019). Developing scientific literacy-based teaching materials to improve students' computational thinking skills. *Jurnal Pendidikan IPA Indonesia*, *8*(4), 482-491.

Fisher, A. (2011). *Critical thinking: An introduction* (2th ed.). Cambridge University Press.

French, B. F., Hand, B., Therrien, W. J., & Vazquez, J. A. V. (2012). Detection of sex differential item functioning in the Cornell critical thinking test. *European Journal of Psychological Assessment*, *28*(3), 201-207.

Ganley, C. M., Vasilyeva, M., & Dulaney, A. (2014). Spatial ability mediates the gender Difference in middle school students' science performance. *Child Development*, *85*(4), 1419-1432.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104-109.

Grunspan, D. Z., Eddy, S. L., Brownell, S. E., Wiggins, B. L., Crowe, A. J., & Goodreau, S. M. (2016). Males under-estimate academic performance of their female peers in undergraduate biology classrooms. *PloS one*, *11*(2), 1-16.

Hanushek, E. A., & Woessmann, L. (2016). Knowledge capital, growth, and the East Asian miracle. *Science*, *351*(6271), 344-345.

Harish, G. C. (2013). Critical thinking skills among ninth standard students in relation to gender, intelligence and study habits. *International Journal of Education and Psychological Research (IJEPR)*, *2*(3), 13-20.

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, *2*(2), 2307-0919.

Hofer, S. I. (2015). Studying gender bias in physics grading: The role of teaching experience and country. *International Journal of Science Education*, *37*(17), 2879-2905.

Hou, L., la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98-125.

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology*, *36*(2), 378-390.

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, *65*, 373-398

Jong, C., Hodges, T. E., Royal, K. D., & Welder, R. M. (2015). Instruments to measure elementary preservice teachers' conceptions: An application of the Rasch rating scale model. *Educational Research Quarterly*, *39*(1), 21-48

Kendhammer, L., Holme, T., & Murphy, K. (2013). Identifying differential performance in general chemistry: Differential item functioning analysis of ACS general chemistry trial tests. *Journal of Chemical Education*, *90*(7), 846-853.

Lisova, T. V., & Kovalchuk, Y. O. (2017) Gender Differences in Mathematics and Science Achievement of Eighth Grade Students in Ukraine on TIMSS 2011. *European Educational Research Conference*, Copenhagen: 21-25 August 2017

Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender Differential Item Functioning in Mathematics in Four International Jurisdictions. *Education & Science/Egitim ve Bilim*, *39*(172).20-32

Mair, P., Hatzinger, R., Maier, M. J., Rusch, T., & Mair, M. P. (2019). Package 'eRm'. *Version 0.14-0*.

Mari, L., Carbone, P. and Petri, D. (2012). Measurement fundamentals: A pragmatic view. IEEE *Transactions on Instrumentation and Measurement, 61*(8), 2107-2114

McFarlane, D. A. (2013). Understanding the challenges of science education in the 21st century: New opportunities for scientific literacy. *International Letters of Social and Humanistic Sciences*, (04), 35-44.

McPeck, J. E. (2016). *Critical thinking and education* (1th ed.). Routledge.

Mesic, V. (2012). Identifying country-specific cultures of physics education: A differential item functioning approach. *International Journal of Science Education*, *34*(16), 2483-2500.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance* (1th ed.). Routledge.

Ministry of Education and Culture of the Republic of Indonesia.(2018). Peraturan Menteri Pendidikan Dan Kebudayaan Republik Indonesia Nomor 37 Tahun 2018.

OECD .(2016). PISA 2015 Assessment and analytical framework: science, reading, mathematic and financial literacy, PISA, OECD Publishing, Paris. retrieved from http://dx.doi.org/10.1787/9789264255425-en

Ong, Y. M., Williams, J., & Lamprianou, I. (2015). Exploring crossing differential item functioning by gender in mathematics assessment. *International Journal of Testing*, *15*(4), 337-355.

Preiss, D. D., Castillo, J. C., Flotts, P., & San Martín, E. (2013). Assessment of argumentative writing and critical thinking in higher education: Educational correlates and gender differences. *Learning and Individual Differences*, *28*, 193-203.

Rachmatullah, A., & Ha, M. (2019). Examining high-school students' overconfidence bias in biology exam: a focus on the effects of country and gender. *International Journal of Science Education*, *41*(5), 652-673..

Rahmawati, R. (2019). Efek sub-sampel pada deteksi differential item functioning (DIF) dengan metode regresi logistik. *JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)*, *1*(1).21-30

Ratini, R., Muchtar, H., Suparman, M. A., Tamuri, A. H., & Susanto, E. (2018). The influence of learning models and learning reliance on students' scientific lteracy. *Jurnal Pendidikan IPA Indonesia*, *7*(4), 458-466.

Ravand, H., & Firoozi, T. (2016). Examining construct validity of the master's UEE using the Rasch model and the six aspects of the Messick's framework. *International Journal of Language Testing*, *6*(1), 1-18.

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PloS one*, *7*(7), 1-16

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology*, *107*(3), 645–662.

Roth, W. M., & Lee, S. (2016). Scientific literacy as collective praxis. *Public Understanding of Science*, *11*(1), 33-56.

Rouquette, A., Hardouin, J. B., & Coste, J. (2016). Differential Item Functioning (DIF) and Subsequent Bias in Group Comparisons using a Composite Measurement Scale: a Simulation Study. *J Appl Meas*, *17*, 312-334.

Rudolph, J. L., & Horibe, S. (2016). What do we mean by science education for civic engagement?. *Journal of Research in Science Teaching*, *53*(6), 805-820.

Runnels, J. (2012). Using the Rasch model to validate a multiple choice English achievement test. *International Journal of Language Studies*, *6*(4), 141-155.

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information & Management*, *54*(2), 189-203.

Rusilowati, A., Kurniawati, L., Nugroho, S. E., & Widiyatmoko, A. (2016). Developing an Instrument of Scientific Literacy Assessment on the Cycle Theme. *International Journal of Environmental and Science Education*, *11*(12), 5718-5727.

Sabah, S., Hammouri, H., & Akour, M. (2013). Validation of a scale of attitudes toward science across countries using rasch model: Findings from TIMSS. *Journal of Baltic Science Education*, *12*(5), 692-702.

Steinmayr, R., Bergold, S., Margraf-Stiksrud, J., & Freund, P. A. (2015). Gender differences on general knowledge tests: Are they due to Differential Item Functioning?. *Intelligence*, *50 (*May–June 2015), 164-174.

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement?. *Review of General psychology*, *16*(1), 93-102.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289-316.

Susongko, P. (2016). Validation of science achievement test with the rasch model. *Jurnal Pendidikan IPA Indonesia*, *5*(2), 268-277.

Susongko, P., Widiatmo, H., Kusuma, M., & Afiani, Y. (2019). Development of integrated science-based science literacy skills instruments using the Rasch model. *Unnes Science Education Journal*, *8*(3).

Tamassia, L., & Frans, R. (2014). Does integrated science education improve scientific literacy?. *Journal of the European Teacher Education Network*, *9*, 131-141

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, *25*(3), 246-280.

Turiman, P., Omar, J., Daud, A. M., & Osman, K. (2012). Fostering the 21st century skills through scientific literacy and science process skills. *Procedia-Social and Behavioral Sciences*, *59*, 110-116.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174.

Wang, C. C., Ho, H. C., Cheng, C. L., & Cheng, Y. Y. (2014). Application of the Rasch model to the measurement of creativity: The creative achievement questionnaire. *Creativity Research Journal*, *26*(1), 62-71.

Wang, M. T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology review*, *29*(1), 119-140.

Welner, K. G. (2013). Consequential validity and the transformation of tests from measurement tools to policy tools. *Teachers College Record*, *115*(9), 1-6.

Wilson, K., Low, D., Verdon, M., & Verdon, A. (2016). Differences in gender performance on competitive physics selection tests. *Physical Review Physics Education Research*, *12*(2), 1-16.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532-547.

Yildirim, O. (2019). Detecting gender differences in PISA 2012 mathematics test with differential item functioning. *International Education Studies*, *12*(8), 59-71.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), 1-30.