Birlesik Dünya Arastırma
BD CENTER
Innovasyon ve Yayıncılık Merkezi

Cypriot Journal of Educational Sciences

Volume 17, Issue 12, (2022) 4530-4543

CJES
Cypriot Journal of Educational Sciences

www.cjes.eu

Differential item functioning detection in fundamental physics test

Duden Saepuzaman, Universitas Negeri Yogyakarta, Graduate School, Jl. Colombo No.1, Yogyakarta, 5528, Indonesia / Universitas Pendidikan Indonesia, Physics Education Department, Jl. Dr. Setiabudhi No. 299, Bandung, 40154, Indonesia https://orcid.org/0000-0002-7810-2328

Edi Istiyono*, Universitas Negeri Yogyakarta, Jl. Colombo No.1, Yogyakarta, 5528, Indonesia https://orcid.org/0000-0001-6034-142X

Haryanto Haryanto, Universitas Negeri Yogyakarta, Jl. Colombo No.1, Yogyakarta, 5528, Indonesia https://orcid.org/0000-0003-3322-904X

Suggested Citation:

Duden Saepuzaman, Edi Istiyono & Haryanto (2022). Differential item functioning detection in fundamental physics test. *Cypriot Journal of Educational Science*. 17(12), 4530-4543.
https://doi.org/10.18844/cjes.v17i12.7642

Received from September 13, 2022; revised from November 19, 2022; accepted from December 26, 2022

Abstract

The study's objective was to identify the differential item functioning (DIF) or bias using item response theory. This research is exploratory research with a quantitative approach. In this study, the data used is a fundamental physics test, specifically motion one and two dimensions, dynamics and work and energy concept, with 25 items in the two-tier multiple choices in form. Two hundred and fifty-four prospective physics teachers at West Java and Banten provinces, Indonesia, consisting of 103 males (reference group) and 151 females (focal group). The data analysis includes two stages: conformity of the logistic model parameters and item characteristics, followed by DIF detection. The model suitability test shows the appropriate parameter is 2PL. DIF detection has three methods: simple area indices, Raju and the Lord method. Research results show two things. First, the model fit test shows that the data best fit the 2-parameter or 2-PL logistic parameter model, and all items have good characteristics. Second, the most significant item containing uniform DIF was number 5, followed by nonuniform DIF items 9 and 25. These three items were biased because the chances of answering correctly for male and female students differed. Significantly, this means that this item favors one group. This study illustrates the importance of constructing instrument items so that all test takers, both male and female, have an equal chance to answer the question correctly.

Keywords: Differential item functioning, physics test, simple area indices, Raju method, Lord method;

* ADDRESS FOR CORRESPONDENCE: Istiyono, Edi, Universitas Negeri Yogyakarta, Jl. Colombo No.1, Yogyakarta, 5528, Indonesia
E-mail address: edi_istiyono@uny.ac.id / Tel.: +62-813-2572-0501

1. Introduction

A valid and consistent (reliable) measuring instrument is essential in making a measuring instrument (Arifin et al., 2020; Babalola et al., 2021; Olumorin et al., 2021; Omolafe, 2021; Saepuzaman et al., 2021b, 2021c). Using measuring instruments that meet these two criteria will obtain results following what is measured without other factors. Another important measurement aspect is ensuring that the instruments used are not biased. A test bias is a condition on a test instrument that is impacted by variables other than the tested one (Retnawati, 2014). The term bias in a test and measurement is bad, has an ethnic meaning, suppresses, or is overly zealous about measuring the object (Osterlind, 2011). Bias is an unfair condition in a test that is inconsistent, infected by factors other than the aspects to be assessed, and inaccuracy in the use of the test; this shows that the blast in the test from the measurement has unsupportive meanings, consistent and valid nature of the test.

The new name reflects the objective and method of detecting bias items with different functions for different test takers. In assessing learning outcomes, bias items should be avoided because they can benefit or harm certain groups (Hernawati et al., 2021; Rahmawati et al., 2021; Retnawati, 2013). Many differential item functioning (DIF) studies have been carried out in various fields, such as language measurement. In developing language proficiency measurement instruments, the DIF is usually analysed to reduce mistrust about the international fairness of outcomes judgment. Translating a test from one language to another does not always result in two psychometrically equal tests (Bartram et al., 2018; Gökçe et al., 2021; Lissitz & Samuelsen, 2007). The DIF studies usually focus on the detection of DIF items only one method (Arim & Ercikan, 2014; Ercikan et al., 2014), comparability of students' scores from different language qualifications (Aryadoust et al., 2021; Ercikan et al., 2015; Hauger & Sireci, 2008) construct comparability examination (Ercikan & Koh, 2005), and investigate the sources that lead to DIF in international appraisal projects (Ercikan, 2002).

Research related to DIF is often carried out in order to obtain a quality instrument that is fair and does not favor certain attributes. DIF analysis has been carried out to test the quality of the Self-stigma instrument to test the psychometric properties of the Self-Stigma Scale-Short (SSS-S) version using Rasch modeling (Fan et al., 2022). The results show that among the three subdomains of the SSS-S, cognitive items seem to be the most strongly supported, and behavioral items are the least supported, or there is a bias in items that measure behavior, so it is necessary to make improvements to these items. Another study by Schauberger and Mair (2020) conducted a study on the topic of a regularisation approach for the detection of DIF in generalised partial credit models. This study uses a regularisation approach based on the lasso principle to detect uniform DIF. This Model is estimated using a probability-penalised approach that automatically detects the DIF effect and simultaneously provides true predictive properties for the detected DIF effect of different covariates. The approach was assessed meaningfully from several simulation studies. An application is presented using data from an inventory of children's depression. Researchers generally use one method in determining bias. This study tries to present various methods in DIF analysis. It becomes important as a DIF analysis because using various methods makes it possible to directly check which items are suspected of being biased so that the analysis obtained becomes more accurate and provides meaningful input on instrument improvement.

Many methods detect DIF (Davidson et al., 2021). This study carried out DIF detection using three methods with the item response theory (IRT) approach. The first stage of analysis is to test the assumptions of local dimensionality and independence before finally finding the Model's fit (Retnawati, 2014). Unidimensional, meaning that every test item measures only one competence. Only if the exam comprises only one dominating component that gauges a subject's performance may unidimensional assumptions be demonstrated. In practice, the unidimensional assumption can be proven through factor analysis. In unidimensional IRT, the relationship between three item parameters, such as item difficulty index (b), item difference/discriminant index (a), and pseudo-guessing index (c) (pseudo-guessing) and one ability (θ) expressed in the equation chances of getting it right. The three-parameter logistic Model (3PL) can be stated in Equation (1) (Almaleki & Alomrany, 2021)

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

The bi parameter points to the ability scale in the item characteristic curve (ICC) when the chance of answering the test taker is 50%. The parameter a_i an index of the distinguishing power of item i. The characteristic curve is proportional to the tangent (slope) direction at the point θ = b. Items with a large, distinguishing power have a very upward curve, while items with a small distinguishing power have a very gentle curve. This parameter describes the probability that a low-ability participant answers correctly to an item. The quasi-guess index in the three-parameter logistic Model allows subjects with low abilities to have the chance to answer the questions accurately. According to the normal distribution origin, the participant's ability score (θ) is usually between –3 and +3.

The two parameters logistic (2PL) model and the one-parameter Model (1PL) are the three parameters logistic model (3PL) cases. When the pseudo-guessing index is equal to 0 (c = 0), the Model becomes 2PL. Likewise, in the 2PL, when the item discriminating power index is 1, this Model becomes a 1PL, better known as the Rasch model. Thus, the 2PL model and 1PL are respectively expressed in Equations (2) and (3) (Almaleki & Alomrany, 2021).

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1+e^{a_i(\theta - b_i)}} \tag{2}$$

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}} \tag{3}$$

The IRT of the 1PL, 2PL, and 3PL models need to choose whether the data to be analysed follows one of the three models. At least two events can be carried out (Retnawati, 2014). The two methods are statistical methods and graphical methods. Several methods for identifying DIF in a test kit include the Simple Area Indices method, the Raju Area Index method, the Lord's Chi-Square method, and the Logistic Regression method.

1.1. Simple area indices

As previously mentioned, the concept of item bias or DIF is defined as the between two groups and there is a disparity in their chances of answering properly, usually called the Focal group and the Reference Group (Angoff, 1993, Cuhadar et al., 2021). In unidimensional IRT, DIF is expressed as the difference in correctly answering an item among the focal and the reference groups. Since the DIF measure is expressed as 'how much difference is there' between the two groups, the characteristic curve is marked with the shaded area shown in Figure 1. The area is named the marked area (SIGNED-AREA), calculated mathematically by the integration method. Because the DIF size is related to the size of a simple area, then to Camilli and Shepard, this method is called Simple Area Indices (Camilli, 2018).

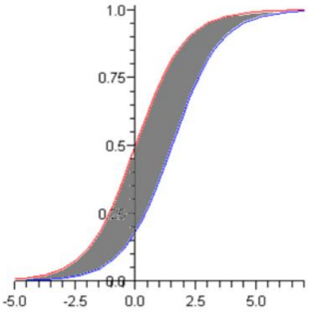


Figure 1. Characteristic Curves of Two Non-Intersecting Groups of Items

In Figure 1, the characteristic item curves do not intersect each other. Because the area is an integration of the probability of correctly answering the reference minus the focal group, the item favors the reference group if it is positive. Conversely, if it is negative, the item benefits the vocal group. The shaded area is presented in Equation (4).

$$SIGNED - AREA = \int [P_R(\theta) - P_F(\theta)]\, d\theta \tag{4}$$

In the DIF analysis of an item, it could be that the characteristic item curves of the two groups intersect each other. The positive and negative DIF measures cancel each other out if this happens, as illustrated in Figure 2. In this case, the area size can be calculated using the UNSIGNED-AREA equation, which is integral to the squared difference between the odds of answering the reference group correctly and the vocal group, as presented in Equation (5).

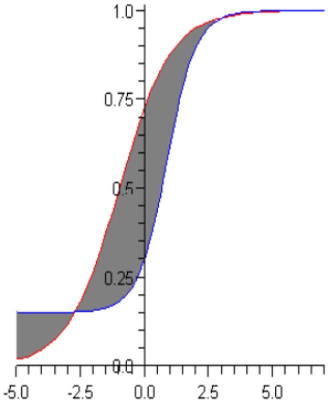


Figure 2. Characteristic Curves of the Two Intersecting Groups

$$UNSIGNED - AREA = \sqrt{\int [P_R(\theta) - P_F(\theta)]^2\, d\theta} \tag{5}$$

1.2. Raju area index method

Raju developed the Raju area index method in 1988 and modified it in 1990. One strategy for comparing item characteristic functions in this area is to compare the characteristic curve (ICC) itself rather than its parameters (Hambleton et al., 1991). The area between the ICCs will be 0 if the item parameters are scaled and the same ICC. The item has DIF if the area between the ICCs is not zero. According to Hambleton et al. (1991) (Azad et al., 2021; Hambleton et al., 1991), in general, the size of the area can be found with the Equation (6)

$$A_i = \sum_{\theta=r}^{s} |P_{i1}(\theta) - P_{i2}(\theta)| \Delta\theta \qquad (6)$$

where Δθ is the interval's size and is taken as small as possible, for example, 0.01. Meanwhile, the r and s values indicate the range of capabilities for the entire calculated area, and this range is usually taken at will and chosen by the user. The price of this range ranges from 3 standard deviations (SD) below the lower group's mean ability to 3 SD above the upper group's mean ability. Raju, quoted by Hambleton et al. (1999) (Azad et al., 2021)., derived a formula that calculates the area formed by ICCs focal and reference groups for three parameters stated in Equation (7)

$$A = (1 - c) \left| \frac{2(a_2 - a_1)}{Da_1 a_2} \right| \ln\left[1 + e^{Da_1 a_1 (b_2 - b_1)/(a_2 - a_1)}\right] - (b_2 - b_1) \qquad (7)$$

1.3. Lord method

Lord's method uses the chi-squared to detect DIF using Equation (8)

$$({\chi_i}^2) = {v_i}' \Sigma^{-1} v_i \qquad (8)$$

where v_i is the vector of the difference in parameter estimation of the i-th item between the reference group and the focal group, Σ^(-1) is the variance-covariance matrix used for item parameter estimation (Camilli et al., n.d.; Stark & Chernyshenko, 2002; Terwee et al., 2021). In practice, identification of DIF using the Lord method will be carried out with the assistance of the R program. an item is identified as DIF if it has a statistically significant chi-square DIF at p < 0.05 (Le Roux et al., 2020). According to Lopez (Uysal et al., 2019), the DIF Lord method can simultaneously test for differences in one or more parameters across reference groups and focus groups. This method's advantages are that it is easy to adapt to any parametric model. Its critical values are easily obtained for different df and significance levels, and the sensitive index to both DIFs is consistent and inconsistent. This study detects DIF in a test device to assess works and energy concepts learning outcomes.

2. Methods

The DIF load on the physics test is detected using a quantitative approach in an exploratory study. The subjects were 254 prospective physics teachers in West Java and Banten provinces, Indonesia, consisting of 103 males (reference group) and 151 females (focal group). In this study, the grouping is based on gender. This grouping ignores ethnicity by assuming the ethnicity of the population used is relatively the same because it is in the same relative area (Hoffmann, 2021). The fundamental physics test instrument consists of 25 items in multiple-choice and five choices. The physics material tested includes motion one and two dimensions, dynamics and work and energy concepts. The test kits used previously were validated. Analysis of prospective physics teachers' response data includes two main stages: model fit and DIF detection. The fit of the IRT model begins with a dimensional test using factor analysis, which is then analysed using chi-square statistics with BILOG MG software. The second stage is DIF detection. DIF detection in this study was carried out using three methods: the Simple Area Indices, the Raju Area Index or Raju statistics, and the Lord method. DIF analysis using the Simple Area Indices method was analysed for each parameter using BILOG MG 3.0. In contrast, the probability area was conducted using Wolfram Alpha computational intelligence (available on the page: https://www.wolframalpha.com/). The Lord method was processed using R software for analysis using the Raju Area Index method or Raju statistics.

3. Result and discussion

3.1. Unidimensional and model fit test

Before the fit model test, the first thing to do is test the dimensional test, unidimensional or multidimensional. The term 'unidimensional' refers to the fact that each item assesses only one competence (Khoeruroh & Retnawati, 2020; Retnawati, 2014). The dimensionality evaluation in this study was verified using Statistical Package for the Social Sciences factor analysis. First, a feasibility test analysis, such as the Kaiser Meyer Olkin- Measure Sampling Adequacy (KMO-MSA) and the Barlett tests, was performed. The KMO-MSA test is used to determine the sample's adequacy, whereas the Barlett test assesses the data's homogeneity. Factor analysis can proceed if the KMO-MSA value is greater than 0.5 and Barlett's significant test is less than 0.05 (Hair et al., 2019). According to the analysis, the value of KMO-MSA is 0.938, and the value of the significant Bartlett test is 0.000 (Saepuzaman et al., 2021a; Sepuzaman et al., 2021). It indicates that the sample selected meets the sample adequacy standards and that the data is homogeneous enough to conduct factor analysis.

Another finding is the eigenvalues with more than one factor. Based on these eigenvalues, physics test instruments have four components or factors and contribute to 36.851% of the total variance (Saepuzaman et al., 2021a; Shrestha, 2021; Yustiandi & Saepuzaman, 2021). These eigenvalues can then be presented in the scree plot in Figure 3.

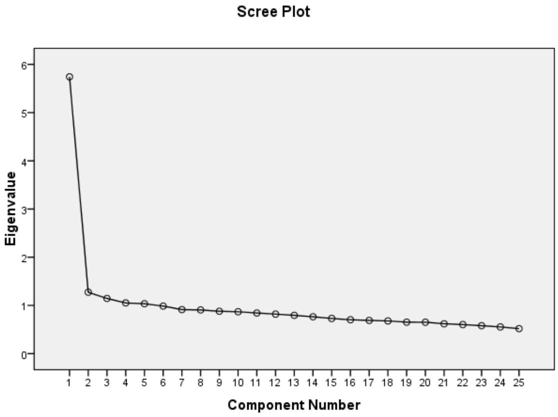


Figure 3. Analysis factor scree plot

The scree plot from the factor analysis in Figure 3 reveals a fairly sharp decline between factors 1 and 2, and the eigenvalues then start to skew at the third factor, virtually forming a right-angled angle. In the physics test, there is just one dominant element or factor. Local independence is another test. This premise of local independence is met if a participant's response to one item has no bearing on their response to another (Volk et al., 2021). According to DeMars (2021), local independence can be established by demonstrating unidimensional assumptions.

The next step is to test the fit of the logistic parameter model. The model fit test was performed in this study using the chi-square statistical method (Nurhasanah et al., 2021). The results of the analysis of the fit of the logistic parameter model are presented in Table 1.

Table 1. The fitness of items with 1 PL, 2 PL, and 3PL

Parameter logistic	Item	Sum
1PL	5, 7,8,9,11,17, 20, 23,25	9
2PL	1,3,4,5,6,7,8,9,11,13, 15, 16, 17, 20, 21, 22, 25	17
3PL	3,5,6,7,8,9,11,13, 14,15, 18,19,20,21,22, 25	16

Table 1 shows that the data fit more when analysed using the 2PL logistic parameter model (IRT 2PL). It means that the opportunity for students (male and female) to answer correctly is a function of discriminant power (a), difficulty index (b), and ability (θ) parameters.

3.2. DIF detection

From the 25 Physics test questions tested, only 17 followed the logistic Model. Furthermore, the 17 test items will be detected in the load of the differentiating item functioning. Furthermore, only items matching the Model are analysed for their DIF load. One attempt to estimate the existence of DIF can be by looking at the item characteristic function or ICC of the item parameters of the two groups. This method is known as the Simple Area Indices method. If the item parameters and the ICC focal and reference group are identical, the area between the ICC will be 0. But, if the area is more than zero, the item may contain DIF (Akbay, 2021). After knowing the Model's fit, the first step of this method is estimating parameters, discriminant index (a), and difficulty level (b) parameters; of focal and reference groups. With the assistance of BILOG MG 3.0, the parameters for the two groups are presented in Table 2.

Table 2. Item parameters in the male and female groups

Item	Male (Reference)		Female (Focal)	
	a	*b*	*a*	*b*
1	1.109	−1.518	1.151	−1.535
3	0.289	1.595	0.299	1.705
4	1.032	−0.837	1.209	−0.891
5	0.810	0.025	0.870	−0.502
6	1.060	−0.333	1.089	−0.316
7	0.407	−1.021	0.372	−1.171
8	0.707	0.036	0.719	−0.089
9	0.977	−0.919	0.692	−1.347
11	0.692	−0.482	0.679	−0.517
13	0.933	0.592	0.833	0.719
15	0.993	−0.598	0.940	−0.618
16	0.917	−2.501	1.032	−2.432
17	0.605	−1.495	0.567	−1.576
20	0.625	0.789	0.691	0.611
21	1.025	−0.098	0.983	−0.141
22	1.067	0.186	1.143	0.189
25	0.549	0.000	0.735	0.209

Table 2 shows that all items have good discriminant (a) and difficulty (b) levels. It refers to the criteria for the item being said to be good if it has a difficulty level between −2 and +2 and the discriminant index between 0 and 2 (Hambleton et al., 1991; Hambleton & Swaminathan, 1985; Otaya et al., 2020).

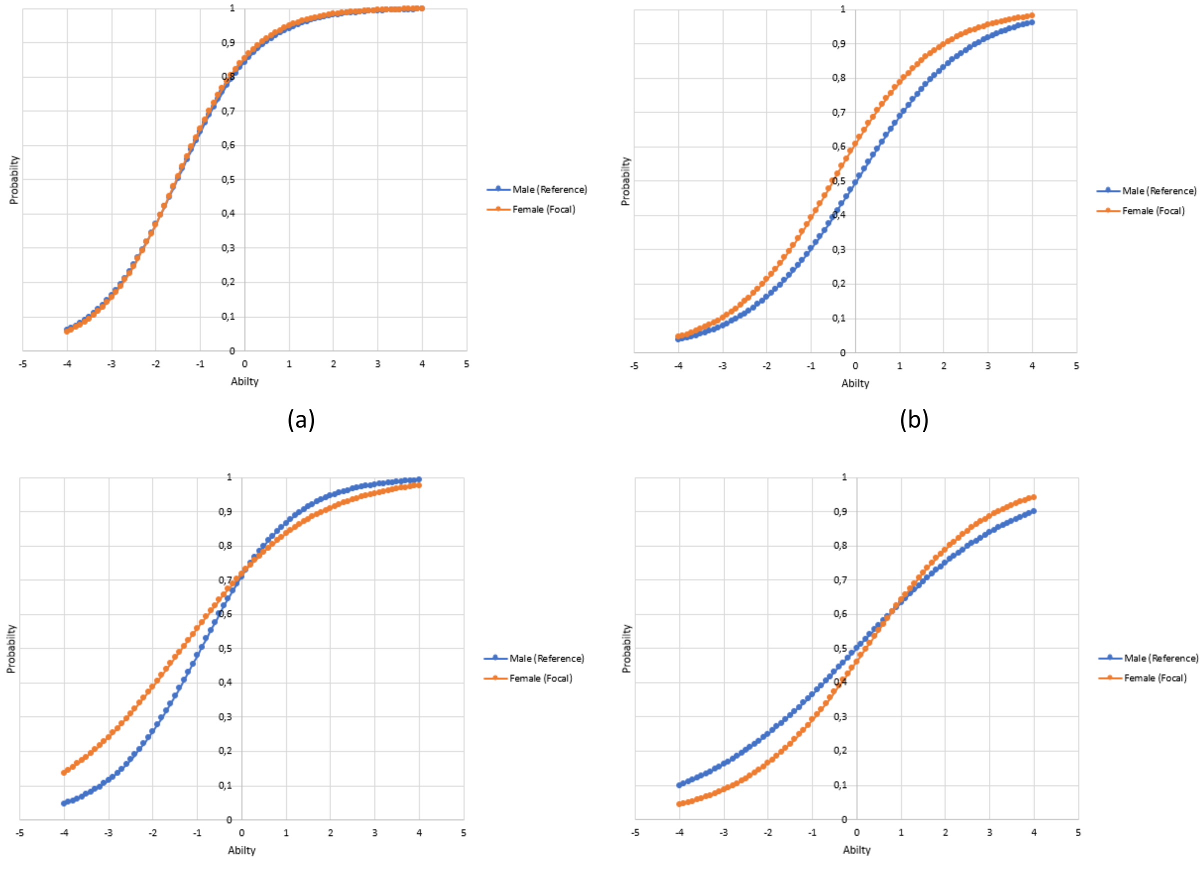
The size of the area formed by the two curves can be calculated with the help of Wolfram Alpha computational intelligence (available on the page: https://www.wolframalpha.com/). The size of the area formed by these differences in opportunities is shown in Table 3. Table 3 shows the largest area or difference in opportunities between the two groups, Signed-Area and Unsigned-Area, owned by items 5, 9 and 25. The area values of these three points are relatively larger than the areas for other items. According to this method, the largest detected DIF is owned by items 5, 9 and 25.

Table 3. The largest area or difference in opportunities between the male and female groups

Item	The intersection of the two curves	Signed-area	Unsigned-area
1	Intersect	−0.0227765	0.0173579
3	Do not intersect	0.0786000	0.0294566
4	Intersect	−0.0669627	0.0722334
5	do not intersect	−0.4942340	0.1992943
6	Intersect	0.0155101	0.0139175
7	Intersect	−0.0520621	0.0466819
8	Do not intersect	−0.1114780	0.0435617
9	Intersect	−0.2903660	0.2169696
11	Intersect	−0.0275997	0.0148031
13	Intersect	0.1037230	0.0715214
15	Intersect	−0.0137387	0.0263472
16	Intersect	−0.0001076	0.0478334
17	Intersect	−0.0262216	0.0349282
20	Intersect	−0.1287410	0.0742515
21	Intersect	−0.0407619	0.0259039
22	Intersect	0.0042753	0.0301235
25	Intersect	0.1879411	0.1629337

DIF between these groups can illustrate from the ICC by identifying the difference in area between the curves. An example for items 1, 5, 9 and 25 ICC is presented in Figure 4. Figure 4 shows differences in the chances of the two male and female groups answering correctly for items 1, 5, 9 and 25. In item 1, the difference in an area formed between the two curves is very small, indicated by the two curves being slightly close together. It shows that the opportunities for males and females in answering item 1 are relatively the same, or there is no indication of DIF. The conditions differ when looking at the ICC curves for items 5, 9 and 25. Figure 4 (b), (c) and (d) show that there is a reasonably large area formed between the curves of the two groups. This condition indicates that males' and females' chances of answering questions on item 1 are significantly different, meaning that items 5, 9 and 25 indicate the presence of DIF. Another finding for item 5 is that female (female) has a greater chance of correctly answering item 5 than male (reference) for the distribution of all student abilities. In item 9, women (female) have a greater chance of correctly answering item 9 than males (reference) for students with abilities −4 to 0.

In contrast, for students with abilities 0 to 4, women (females) can answer true smaller than males (reference). The opposite is in item 25. Students with abilities −4 to 0.7 male (reference) have a greater chance of answering item 9 than women (female). Meanwhile, students with abilities from 0.7 to 4 male (reference) have a smaller chance of correctly answering item 9 than women (female). These findings further confirm that items 5, 9 and 25 indicated DIF.



(a)

(b)

(c)

(d)

Figure 4. ICC for focal and reference group for (a) Item 1, (b) Item 5, (c) Item 9 and (d) Item 25

Another method was used for further analysis of DIF indications on these items. One of the DIF checks can be by comparing the results obtained with other methods. In this study, the DIF analysis of unidimensional IRT is applied to the Simple Area Indices, namely the Raju Area Index method or Raju statistics and the Lord method. The Raju Area Index method and the Lord method using R Program analysis. In general, the R output for these methods is presented in Table 4.

Table 4. DIF detection with raju and lord methods

Item	Raju				Lord	
	Signed area		Unsigned area			
	Stat.	*p*-value	Stat.	*p*-value	Stat.	*p*-value
b1	0.7178	0.4729	0.7080	0.4789	0.5214	0.7705
b3	0.2792	0.7801	−0.5638	0.5729	0.7961	0.6716
b4	0.3722	0.7097	1.3906	0.1644	2.5962	0.2730
b5	−3.1312	0.0017 **	3.3467	0.0008 ***	14.4943	0.0007 ***
b6	0.5199	0.6031	0.9178	0.3587	0.905	0.6361
b7	−0.1287	0.8976	−0.5478	0.5838	0.5014	0.7783
b8	−0.9111	0.3623	1.0681	0.2855	1.4731	0.4788
b9	−1.7300	0.0836 .	−1.6582	0.0973 .	3.0364	0.2191
b11	0.1794	0.8576	0.5697	0.5689	0.3462	0.8411

Item	Raju				Lord	
	Signed area		Unsigned area			
	Stat.	*p*-value	Stat.	*p*-value	Stat.	*p*-value
b13	0.4160	0.6774	−0.4160	0.6774	0.2039	0.9031
b15	0.4326	0.6653	0.4546	0.6494	0.2244	0.8939
b16	0.7015	0.4830	0.7148	0.4748	0.6196	0.7336
b17	−0.3946	0.6932	1.0518	0.2929	1.4775	0.4777
b20	−1.5108	0.1308	1.5164	0.1294	2.6967	0.2597
b21	−0.2487	0.8036	0.4964	0.6196	0.2649	0.8760
b22	−0.1528	0.8785	1.1685	0.2426	1.5143	0.4690
b25	1.2716	0.2035	1.7124	0.0868 .	4.7344	0.0937 .
DIF item	b5		b5		b5	
potensional DIF	b9		b9, b25		b25	
Treshold	−1.96 and 1.96 (significance level: 0.05)				5.9915 (significance level: 0.05)	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Table 4 shows that the items detected that significantly contained DIF with the Raju and Lord analysis was item 5. But looking at the statistical numbers and p-value above, several items other than item 5 have the potential for DIF, namely those with a stat, and the p-value is close to the threshold (threshold). These items are points 9 and 25. And the potential for this item to contain DIF is also strengthened by the previous Simple Area Indices method, which shows that the extent of the difference in this item's probability is greater than the area of other items. Based on DIF detection using these three methods, the results showed that the 17 items analysed showed that item number 5 contained DIF significantly. Meanwhile, items close to the DIF (potential DIF) threshold value are items number 9 and point 25.

If an item-identified DIF is given to a child, the results will be less accurate in measuring students' abilities. It happens because the DIF allows students with the same competence to demonstrate systematic disparities in specific groups (Kolmos et al., 2020). A bias in the physics assessment is in line with several studies. Andriani et al. (2019) detected bias in physics questions on the national standard school exam (USBN) using Rasch modeling with the research results that out of the 40 questions analysed, one question was detected with DIF or bias. In line with this, Li and Singh (2021) show that female students' self-efficacy and interest in physics are lower than male students at the beginning of the lesson, but the gender gap in this motivational construct becomes larger at the end of the lesson. It is due to an unfair and non-inclusive learning environment in which female students' self-efficacy and interest in physics are lower than male students at the beginning of the lesson, and the gender gap in this motivational construct becomes larger at the end of the lesson. This study can still be developed by looking at other factors that influence the response of test participants besides ability and gender, such as ethnicity (Steele, 2018), demographics and school background. In addition, this research is still limited in the number of items, and it is hoped that further research will have more items that can measure indicators of competency achievement. This is important in efforts to package the instrument into more items to choose from. In order to obtain quality instrument constituent items.

4. Conclusion

It can be concluded based on the findings that the fundamental physics test instrument is unidimensional and local independent. Regarding the logistic parameter model test, the Model's fit showed that the data matched the logistical Model 2 parameters or 2 PL with the number of items that matched 17 items and good item characteristics. DIF detection using Simple Area Indices, the Raju method, and the Lord method means relatively the same results. The three items detected by DIF mean that it benefits certain students if categorised by gender, male and female. This study mainly investigates the significance of DIF utilising the unidimensional IRT for male and female students. It is still necessary to develop further research that considers other sources of bias besides measuring the main factors or aspects such as ethnicity (Steele, 2018), demographics and school background.

Acknowledgements

We would like to thank the Government of Indonesia for funding this research through the Directorate of Resources, Directorate General of Education, Ministry of Education, Research and Technology, according to the 2021 Funding and Research Contract Number: 040/E5/PG.04.00.PT/2022.

References

Adams, et al. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado learning attitudes about science survey. Physical Review Special Topics - Physics Education Research, 2(1), 010101. https://doi.org/10.1103/PhysRevSTPER.2.010101

Akbay, L. (2021). Impact of retrofitting and item ordering on DIF. Journal of Measurement and Evaluation in Education and Psychology, 12(2), 212–225. https://doi.org/10.21031/epod.886920

Almaleki, D. A., & Alomrany, A. G. (2021). The effect of methods of estimating the ability on the accuracy and items parameters according to 3PL model. International Journal of Computer Science & Network Security, 21(7), 93–102.

Andriani, N., Suhendi, E., & Samsudin, A. (2019). Analisis butir dan deteksi bias soal fisika pada Ujian Sekolah Berstandar Nasional (USBN) dengan menggunakan pemodelan rasch untuk standarisasi penilaian. Seminar Nasional Fisika, 5(1), 167–172. http://proceedings.upi.edu/index.php/sinafi/article/view/583

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. Lawrence Erlbaum Associates, Inc (Vol. 453, pp. 3–23). https://psycnet.apa.org/record/1993-97193-001

Arifin, S., Retnawati, H., & Putranta, H. (2020). Indonesian air force physical tester reliability in assessing one-minute push-up, pull-up, and sit-up tests. Sport Mont, 18(2), 89–93. https://doi.org/10.26773/smj.200614

Arim, R. G., & Ercikan, K. (2014). Comparability between the American and Turkish versions of the Timss mathematics test results. Egitim ve Bilim, 39(172), 33–48.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. Language Testing, 38(1), 6–40. https://doi.org/10.1177/0265532220927487

Azad, A. K., Raju, V., & Islam, M. E. (2021). Evaluating factors affecting employee job performance in Bangladeshi Service Sector. PalArch's Journal of Archaeology of Egypt/Egyptology, 18(15), 787–800.

Babalola, E. O., Boor, C. H. M., Aladesusi, G. A., & Shomoye, M. A. (2021). Development and validation of digital photo series for the teaching of BT in Ilorin, Nigeria. Indonesian Journal of Educational Research and Technology, 1(3), 105–116.

Bartram, D., Berberoglu, G., Grégoire, J., Hambleton, R., Muniz, J., & van de Vijver, F. (2018). ITC guidelines for translating and adapting tests (second edition). International Journal of Testing, 18(2), 101–134. https://doi.org/10.1080/15305058.2017.1398166

Camilli, G. (2018). IRT scoring and test blueprint fidelity. Applied Psychological Measurement, 42(5), 393–400. https://doi.org/10.1177/0146621618754897

Camilli, G., Shepard, L. A., & Shepard, L. (n.d.). Methods for identifying biased test items (Vol. 4).

Cuhadar, I., Yang, Y., & Paek, I. (2021). Consequences of ignoring guessing effects on measurement invariance analysis. Applied Psychological Measurement, 45(4), 283–296. https://doi.org/10.1177/01466216211013915

Davidson, M. J., Wortzman, B., Ko, A. J., & Li, M. (2021). Investigating item bias in a CS1 exam with differential item functioning. SIGCSE 2021 - Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (pp. 1142–1148). https://doi.org/10.1145/3408877.3432397

DeMars, C. E. (2021). Violation of conditional independence in the many-facets rasch model. Applied Measurement in Education, 34(2), 122–138. https://doi.org/10.1080/08957347.2021.1890743

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. International Journal of Testing, 2(3), 199–215. https://doi.org/10.1207/s15327574ijt023&4_2

Ercikan, K., Chen, M. Y., Lyons-Thomas, J., Goodrich, S., Sandilands, D., Roth, W. M., & Simon, M. (2015). Reading proficiency and comparability of mathematics and science scores for students from english and non-english backgrounds: An international perspective. International Journal of Testing, 15(2), 153–175. https://doi.org/10.1080/15305058.2014.957382

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French Versions of TIMSS. International Journal of Testing, 5(1), 23–35. https://doi.org/10.1207/s15327574ijt0501_3

Ercikan, K., Roth, W. M., Simon, M., Sandilands, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. Applied Measurement in Education, 27(4), 273–285. https://doi.org/10.1080/08957347.2014.944306

Fan, C. W., Chang, K. C., Lee, K. Y., Yang, W. C., Pakpour, A. H., Potenza, M. N., & Lin, C. Y. (2022). Rasch modeling and differential item functioning of the self-stigma scale-short version among people with three different psychiatric disorders. International Journal of Environmental Research and Public Health, 19(14), 8843.

Gökçe, S., Berberoğlu, G., Wells, C. S., & Sireci, S. G. (2021). Linguistic distance and translation differential item functioning on trends in international mathematics and science study mathematics assessment items. Journal of Psychoeducational Assessment, 39(6), 728–745. https://doi.org/10.1177/07342829211010537

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning EMEA.

Hambleton, R. K., Shavelson, R. J., Webb, N. M., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory (Vol. 2). Sage.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: principles and applications. Kluwer-Nijhoff.

Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. International Journal of Testing, 8(3), 237–250. https://doi.org/10.1080/15305050802262183

Hernawati, D., Nandiyanto, A. B. D., & Muhammad, N. (2021). The use of learning videos in order to increase student motivation and learning outcomes during the COVID-19 pandemic. ASEAN Journal of Science and Engineering Education, 1(2), 77–80. https://ejournal.upi.edu/index.php/AJSEE/article/view/33370

Hoffmann, A. L. (2021). Terms of inclusion: Data, discourse, violence. New Media & Society, 23(12), 3539–3556.

Khoeruroh, U., & Retnawati, H. (2020). Comparison sensitivity of the differential item function (DIF) detection method. Journal of Physics: Conference Series, 1511(1), 12042. https://doi.org/10.1088/1742-6596/1511/1/012042

Kolmos, A., Holgaard, J. E., & Clausen, N. R. (2020). Progression of student self-assessed learning outcomes in systemic PBL. European Journal of Engineering Education, 46(1), 1–23. https://doi.org/10.1080/03043797.2020.1789070

Le Roux, E., Corboz, J., Scott, N., Sandilands, M., Lele, U. B., Bezzolato, E., & Jewkes, R. (2020). Engaging with faith groups to prevent VAWG in conflict-affected communities: Results from two community surveys in the DRC. BMC International Health and Human Rights, 20(1), 1–20. https://doi.org/10.1186/s12914-020-00246-8

Li, Y., & Singh, C. (2021). Effect of gender, self-efficacy, and interest on perception of the learning environment and outcomes in calculus-based introductory physics courses. Physical Review Physics Education Research, 17(1), 10143. https://doi.org/10.1103/PhysRevPhysEducRes.17.010143

Lissitz, R. W., & Samuelsen, K. (2007). Further clarification regarding validity and education. Educational Researcher, 36(8), 482–484. https://doi.org/10.3102/0013189x07311612

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the mantel-haenszel with multiple ability estimates to detect differential item functioning. Journal of Educational Measurement, 32(2), 131–144. https://doi.org/10.1111/j.1745-3984.1995.tb00459.x

Nurhasanah, Rusyana, A., & Fitriana, A. R. (2021). Binary logistic regression for identification of high school student interest in Banda Aceh city in continuing study at Universitas Syiah Kuala. Journal of Physics: Conference Series, 1882(1), 12034. https://doi.org/10.1088/1742-6596/1882/1/012034

Olumorin, C. O., Babalola, E. O., Aladesusi, G. A., Issa, A. I., & Omolafe, E. V. (2021). Experts' validation of the developed 3-dimensional automated Model of the human heart to teach a biology concept in Ilorin, Nigeria. Indonesian Journal of Multidiciplinary Research, 1(2), 299–308.

Omolafe, E. V. (2021). Primary educators experts' validation of the developed mathematics mobile application to enhance the teaching of mathematics in Nigeria primary schools. ASEAN Journal of Science and Engineering Education, 2(1), 157–166. https://ejournal.upi.edu/index.php/AJSEE/article/view/38505

Osterlind, S. (2011). Test item bias. In Test item bias. https://doi.org/10.4135/9781412986090

Otaya, L. G., Kartowagiran, B., Retnawati, H., & Mustakim, S. S. (2020). Estimating the ability of pre-service and in-service Teacher Profession Education (TPE) participants using Item Response Theory. REID (Research and Evaluation in Education), 6(2), 160.

Rahmawati, F., Achdiani, Y., & Maharani, S. (2021). Improving students' learning outcomes using 5e learning cycle model. ASEAN Journal of Science and Engineering Education, 1(2), 97–100.

Retnawati, H. (2013). Pendeteksian keberfungsian butir pembeda dengan indeks volume sederhana berdasarkan teori respons butir multidimensi. Jurnal Penelitian Dan Evaluasi Pendidikan, 17(2), 275–286. https://doi.org/10.21831/pep.v17i2.1700

Retnawati, H. (2014). Teori respons butir dan penerapannya [Item response theory and its application]. Nuha Medika.

Saepuzaman, D., Haryanto, H., Istiyono, E., Retnawati, H., & Yustiandi, Y. (2021a). Analysis of items parameters on work and energy subtest using item response theory. Jurnal Pendidikan MIPA, 22(1), 1–9.

Saepuzaman, D., Istiyono, E., & Retnawati, H. (2021b). Analisis estimasi kemampuan siswa dengan pendekatan item response theory penskoran dikotomus dan politomus. Karst: Jurnal Pendidikan Fisika Dan Terapannya, 4(1), 8–13.

Saepuzaman, D., Retnawati, H., & Istiyono, E. (2021c). Can innovative learning affect student HOTS achievements? A meta-analysis study. Pegem Journal of Education and Instruction, 11(4), 290–305.

Schauberger, G., & Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. Behavior Research Methods, 52(1), 279–294.

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. American Journal of Applied Mathematics and Statistics, 9(1), 4–11.

Solheim, O. J. (2011). The impact of reading self-efficacy and task value on reading comprehension scores in different item formats. Reading Psychology, 32(1), 1–27. https://doi.org/10.1080/02702710903256601

Stark, S., & Chernyshenko, O. (2002). Detection of differential item/test functioning (DIF/DTF).

Steele, C. (2018). Stereotype threat and African-American Student Achievement. In Inequality in the 21st century (pp. 315–318). https://doi.org/10.4324/9780429499821-55

Terwee, C. B., Crins, M. H. P., Roorda, L. D., Cook, K. F., Cella, D., Smits, N., & Schalet, B. D. (2021). International application of PROMIS computerized adaptive tests: US versus country-specific item parameters can be consequential for individual patient scores. Journal of Clinical Epidemiology, 134, 1–13. https://doi.org/10.1016/j.jclinepi.2021.01.011

Uysal, İ., Ertuna, L., Ertaş, F. G., & KeleciOğlu, H. (2019). Performances based on ability estimation of the methods of detecting differential item functioning: A simulation study*. Journal of Measurement and Evaluation in Education and Psychology, 10(2), 133–148. https://doi.org/10.21031/epod.534312

Volk, C., Rosenstiel, S., Demetriou, Y., Sudeck, G., Thiel, A., Wagner, W., & Höner, O. (2021). Health-related fitness knowledge in adolescence: Evaluation of a new test considering different psychometric approaches (CTT and IRT). German Journal of Exercise and Sport Research, 1–13. https://doi.org/10.1007/s12662-021-00735-5

Walsh, M., Hickey, C., & Duffy, J. (1999). Influence of item content and stereotype situation on gender differences in mathematical problem solving. Sex Roles, 41(3–4), 219–240. https://doi.org/10.1023/A:1018854212358

Xie, B., Davidson, M. J., Franke, B., McLeod, E., Li, M., & Ko, A. J. (2021). Domain experts' interpretations of assessment bias in a scaled, online computer science curriculum. L@S 2021 - Proceedings of the 8th ACM Conference on Learning @ Scale (pp. 77–89). https://doi.org/10.1145/3430895.3460141

Yustiandi, Y., & Saepuzaman, D. (2021). Analysis of model fit and item parameter of work and energy test using item response theory. Gravity : Jurnal Ilmiah Penelitian Dan Pembelajaran Fisika, 7(2). https://doi.org/10.30870/gravity.v7i2.10563