International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 06, 2020
ISSN: 1475-7192

Differential Item Functioning (DIF) Using Rasch Model in Diagnostic Test Instrument

[1]Mohd Zaharuddin Omar,[*2]Zamri Mahamod

ABSTRACT -- *Detecting differential item is one way of determining the reliability of an instrument. The researcher used Rasch model (Differential Item Functioning Analysis – DIF) to identify the items which are gender-biased in the Diagnostic Test of Lower Secondary Bahasa Melayu System (DTLSBMS) instrument. Gender-biased detection is done on all items since the study found the existence of different academic performance based on gender and the analyses of public examination results like UPSR, PMR and SPM. Female students' performance was better than male students. The items in the instrument should be fair to all test takers. This instrument has 289 items and six constructs which were verified by seven experts. The samples were 935 Form One students from ten districts in Pahang, Malaysia. Alpha Cronbach index values (KR20) for the whole items were 0.98 while the Alpha Cronbach index values for each construct was between 0.71 and 0.95. The result of the study found that four items of Morphology construct and two items of syntax had critical DIF t-value (more than ±2.0 logits) whereby the DIF contrast value was still at ±0.5. In short, UDSBMR instrument has good reliability and fairness as a diagnostic test tool.*

Keywords-- *Reliability, Differential Item Functioning – DIF, Gender-Biased, DIF Critical t Value, DIF Contrast, Fairness.*

I. INTRODUCTION

Differential aspect shows an item's or a test's unfairness in assessing students' performance. The assessment procedure should be fair to all students and does not discriminate students based on their race, gender, geography and their being disabled. To study the differential aspect of a test, seven questions should be covered before a test can be considered `fair'. (1) Are the items challenging (too difficult) or offensive the students? (2) Are there any items or tasks which favour a certain race, gender or an origin? (3) Are the diverse groups represented fairly in the assessment? (4) Are all students given the same opportunity in getting the knowledge and skills to be assessed? (5) Are there any items or tasks which require knowledge and skills not related with the measurable results? (6) Are there any answers to the questions and items which depend on the untaught knowledge? But the students relied on a different source. (7) Are the language and format used in the assessment too common to the students? (Azizi, 2010). In this study, the researcher focused on item (2) which is identifying the existence of gender-biased items.

Detecting the differential aspect was done since the academic achievement between the male and female students was imbalanced in many countries (Rodriguez, 2003). In Malaysia, female students' academic performance was better than male's (Hanita Mohd Yusoff & Norzaini Azman, 2018; Zalizan Mohd Jelas et al., 2012). The study on the 2001 – 2005 public examinations' results (UPSR, PMR and SPM) was carried out in 2005.

[1] *Education, IPG Campus Dato' Razali Ismail, Terengganu, Malaysia.*
[2]* *National University of Malaysia, Bangi, Selangor, Malaysia, d-zam@ukm.edu.my.*

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

The same study was also done on the 1996 – 2000 public examinations' results. The studies showed female students performed better in most subjects like Science, English, Mathematics and Bahasa Melayu. This was depicted in Table 2 which is SPM National Average Grade (*Gred Purata Nasional* - GPN) in 2006, 2011 and 2017 and they proved that female students' performance was better than male (*Lembaga Peperiksaan Malaysia*, KPM, 2018). Therefore, it is crucial for this study to detect the existence of differential items to maintain the quality and reliability of the instrument so that the items prepared did not favour any specific gender or group. Differential happens when one group of respondents scored higher compared to another group of respondents though the students' abilities were the same or almost the same. Bon and Fox (2007) suggested three guidelines to detect differential item functioning using Rasch Model: (i) t ±2.0 (t≥ + 2.0 ≤ -2.0); (ii) DIF contrast ±0.5 (≥+0.5≤-0.5); and (iii) $p<0.05$ (significant).

Table 1: National Average Grade (GPN) Based on Gender (2006, 2011 and 2017)

Year	*Daily Schools* *National Average Grade (GPN) SPM*	
	Male	*Female*
2006	6.16	5.50*
2011	5.58	4.72*
2017	5.41	4.58*

*Lower value is better

Based on a library study and the analysis of students' performance in UPSR and PT3, language system was identified as the aspect of language which required attention and monitoring. Therefore, the researcher decided to choose the language system as a diagnostic test content to be developed. Hopefully, this instrument can help students and teachers to identify any existing weaknesses before the students start learning Bahasa Melayu at the lower secondary level especially in form One. Lower Secondary Bahasa Melayu Diagnostic Test (UDSBMR – *Ujian Diagnostik Bahasa Melayu Menengah Rendah*) consists of six sets of instruments constructed to identify the weaknesses and difficulties in the language system of Form One Bahasa Melayu subject. This instrument was developed based on the Curriculum and Assessment Standard Document (DSKP – *Dokumen Standard Kurikulum dan Pentaksiran*) for Primary School Bahasa Melayu (Level 2) and Curriculum and Assessment Standard Document for Form One Bahasa Melayu (Lower Secondary) which consists of morphology, word formation, syntax, spelling, vocabulary and proverbs.

II. METHODOLOGY

This study aims to develop a UDSBMR instrument which fulfills the requirements of a test which has acceptable validity and reliability. The items are fair to all samples taking the test. In relation to this objective, the following are the research questions and hypotheses.

1. Does every construct in the UDSBMR instrument have acceptable reliability values?

   Ho1. No construct in the UDSBMR instrument has unacceptable reliability values.

   2. Are there any items in the UDSBMR instruments gender-biased?

Ho2. No item in each construct is gender-biased.

The researcher used Rasch model to analyse the instrument's reliability. Alpha Cronbach (KR20) reliability analysis, respondents' and items' reliability were used in this study. The values set on the reliability index value was 0.8 – 1.0 as very good, 0.6 – 0.8 as less acceptable and below 0.6 as unacceptable (Bond & Fox, 2007; Fisher, 2007) set 0.94 – 1.0 as excellent, 0.91 – 0.94 as very good, 0.81 – 0.90 as good, 0.67 – 0.80 as moderate and 0.67 and below as weak.

Differential Item Functioning (DIF) was done using Rasch Model via Winstept Application. Winstept depicts a graph showing different items' difficulty levels in two studied groups. The analysis was done automatically with a table of the two groups' measurements (gender) with *two tailed t-test*. Significant DIF was measured based on *t*-value at $p<0.05$ (Confidence 95%) and the critical *t*- value level set at +2.0/-2.0 ligits to all DIF analysis which is important not to have differential item (Bon & Fox, 2007). The *t*- test analysis aims to determine gender-biased items. However, if the critical *t* -value was outside ±2.0 logit, the logit index value for DIF contrast had to be re-evaluated. If the logit index value was at ±0.5 (-0.5 to 0.5), the item concerned was not biased (Lai & Eton, 2002; Bon & Fox, 2007). If code 1 is for male and code 2 is for female, the negative value (-) shows the item concerned was easily agreed upon by male whereby the positive value (+) shows an item was easily agreed upon by female. GDIF contrast index was used to show the gap among all items while comparing the male and female groups. The size of GDIF which exceeds 0.5 shows the existence of GDIF. The negative index of GDIF contrast means the items are easier to be verified while the positive index was harder to be verified by either male or female.

1) Study Samples

The researcher decided to choose Pahang as the study location since the state had average performance records in UPSR compared to other states specifically in Bahasa Melayu Writing and Comprehension in the past five years. Pahang also has students with different mother tongues like Bahasa Melayu, Chinese, Tamil and native languages. Three main dialects used by the residents are Pahang, Kelantan and Terengganu dialects. The information of the study samples is shown in Table 2.

Table 2: Number of Study Samples

Category	*District*	*Schools*	*No. of Samples*	*Total*
Urban	Jerantut	SMK Jerantut	86	450
	Betong	SMK Ketari	100	
	Maran	SMK Maran	95	
	Kuantan	SMK Pelindung	88	
	Raub	SMK Mahmud	83	
Sub urban	Temerloh	SMK Kuala Krau	106	485
	Bera	SMK Mengkarak	104	
	Rompin	SMK Sungai Puteri	91	
	Pekan	SMK Paloh Hinai	89	
	Lipis	SMK Merapoh	97	

*SMK – Sekolah Menengah Kebangsaan

The schools were randomly chosen based on the categories. The school names were written on small pieces of paper which were rolled and put into two containers marked urban and sub-urban. The total of study samples were 935 students from 10 schools representing 10 districts in Pahang. The samples include low, average and high achievers in UPSR Bahasa Melayu subject. To use Rasch Model, at least 200 samples needed for each study group (Wright & Stone, 2004). If the samples exceeded 300 students the findings of the analysis would get high reliability (Schulz, 1990). Equal or nearly equal number of male and female students is not needed in DIF study (Linacre, 2010). The findings of DIF analysis are not always repetitive even though the samples used have similar characteristics. So, the findings would only explain the items' achievement based on the responses given by the samples.

2) Study Tool

UDBMMR is the main study tool in this study. UDBMMR contains six sets of test based on six main constructs which had 289 items. The constructs are morphology (111 items), word formation (50 items), syntax (50 items), vocabulary (15 items), proverbs (28 items) and spelling (35 items). The item building covers important aspects of language system. The items built as matching, filling in the blanks, giving short answers and constructing sentence.

III. RESULTS AND ANALYSIS

Instrument's Reliability. The findings of the study concern the first study issue which is to answer the first study null hypothesis (Ho1) claiming that all constructs in the UDSBMR instrument have Alpha Cronbach reliability values, acceptable respondents' and items' reliability. The findings of the study are shown in Table 3.

Table 3: Alpha Cronbach Reliability, Respondent and Item

Language Skills Constructs	*No. of Items*	*Respondent Reliability*	*Item Reliability*	*Alpha Cronbach (KR20) Reliability*
Morphology	111	0.94	0.97	0.95
Word Formation	50	0.86	0.97	0.88
Syntax	50	0.85	0.98	0.88
Vocabulary	15	0.61	0.95	0.71
Proverbs	28	0.76	0.98	0.80
Spelling	35	0.76	0.96	0.80
UDSBMR	289	0.97	0.97	0.98

The findings of the study show the Alpha Cronbach (KR20) reliability index for all items is 0.98, while the reliability indexes for each construct are morphology (0.95), syntax (0.88), word formation (0.88), proverbs (0.80) and spelling (0.80). Only vocabulary construct is at a moderate level with the value of 0.71 (Fisher, 2007). In short, all constructs have very good reliability index values except for vocabulary. Nonetheless, all constructs have acceptable reliability values (Bond & Fox, 2007).

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

On the whole, the respondent reliability index is very good with the value of 0.97. Meanwhile, the reliability indexes for each construct are between 0.61 and 0.94 with three constructs are in the very good category (0.8 – 1.0) and they are syntax (0.85), word formation (0.86) and morphology (0.94). The other three constructs are in `less acceptable' level and the constructs are proverbs (0.76), spelling (0.76) and vocabulary (0.61). No construct is at the level of `unacceptable' < 0.6 (Bond & Fox, 2007).

Generally, the item reliability index is very good with 0.97 (Bond & Fox, 2007). The reliability indexes for all constructs are between 0.95 and 0.98. The items' reliability for morphology is 0.97, word formation (0.97), syntax (0.98), vocabulary (0.95), proverbs (0.98) and spelling (0.96). The findings show the UDSBMR instrument has very good (Bond & Fox, 2007) and excellent (Fisher, 2007) items' reliability indexes.

To summarize, UDSBMR instrument has good and acceptable Alpha Cronbach reliability indexes, respondents' reliability and items' reliability. Only respondents' credibilities for proverbs, vocabulary and spelling are at moderate and less acceptable level (Bond & Fox, 2007). However, the researcher decided to accept this value since it is still above 0.6 which is less acceptable but not rejected.

IV. DISCUSSION

The findings of the study concern the second study issue which is to answer the second study hypothesis (Ho2) which claims that no item in every construct is biased towards a certain gender in the UDSBMR instrument. The findings of the study used the Differential Item Functioning (DIF) analysis as shown in Table 4 to Table 9 and Figure 1 to Figure 6.

1) Morphology Construct

Differential Item Functioning (DIF) analysis for Morphology is shown in Table 4 and Figure 1.

Table 4: DIF Item for Morphology Construct

Person Class	*Dif. Measure*	*Person Class*	*Dif. Measure*	*Dif. Contrast*	*t*	*Name*
1	-0.73	2	-0.41	-0.32	2.06	A16
2	-0.41	1	-0.73	0.32	2.06	A16
1	-0.49	2	-0.19	-0.3	2.03	AA25
2	-0.19	1	-0.49	0.3	2.03	AA25
1	0.86	2	0.56	0.3	2.09	AA48
2	0.56	1	0.86	-0.3	2.09	AA48
1	0.52	2	0.22	0.3	2.06	AA55
2	0.22	1	0.52	-0.3	2.06	AA55

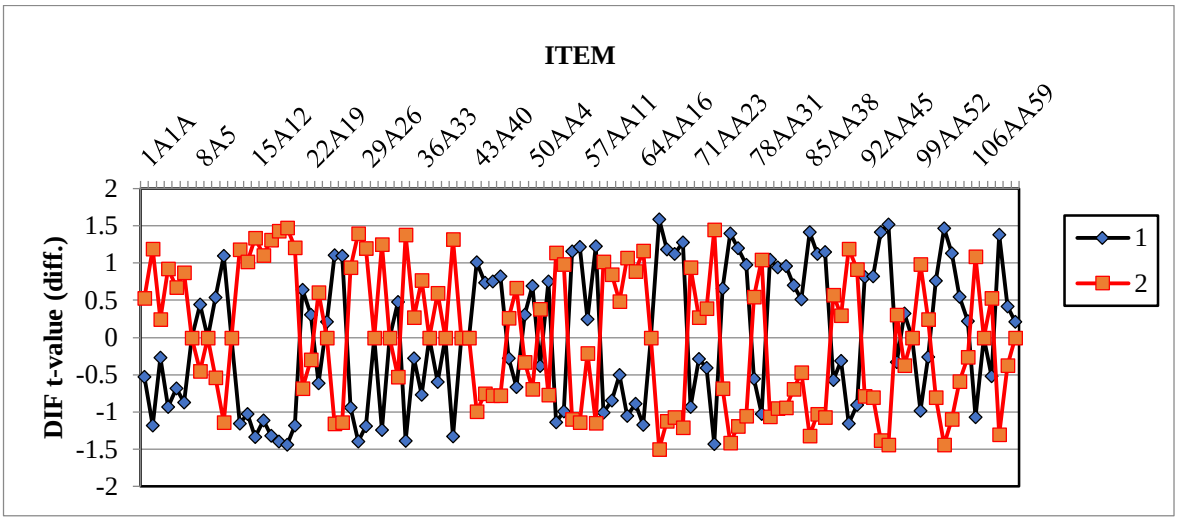

Figure 1: DIF *t*-Value Plot for Morphology Construct

Differential Item Functioning (DIF) analysis was done on 111 items in Morphology construct. Four items identified to have critical *t*-value over ±2.0 were A16 (2.06), AA25 (2.03), AA48 (2.09) and AA55 (2.06). However, after having checked the DIF Contrast, the value was found to be at ±0.5 logits which are A16 (-0.32), AA25 (-0.3), AA48 (0.3) and AA55 (0.3). The significance here is no gender biased item. Table 4 shows the critical *t*- value for morphology construct items was more than ±2.0 logits but it did not exceed the DIF Contrast value which was ±0.5 logits. The findings show there is no gender-biased item in Morphology construct.

2) Word Formation Construct

Differential Item Functioning (DIF) analysis for word formation construct is shown in Table 5 and Table 2.

Table 5: DIF Item Word Formation Construct

Person Class	*Dif. Measure*	*Person Class*	*Dif. Measure*	*Dif. Contrast*	*t*	*Name*
1	-0.27	2	0.04	-0.3	-2.03	B20
2	0.04	1	-0.27	0.3	-2.03	B20
1	-0.6	2	-0.27	-0.33	-2.09	B36
2	-0.27	1	-0.6	0.33	-2.09	B36

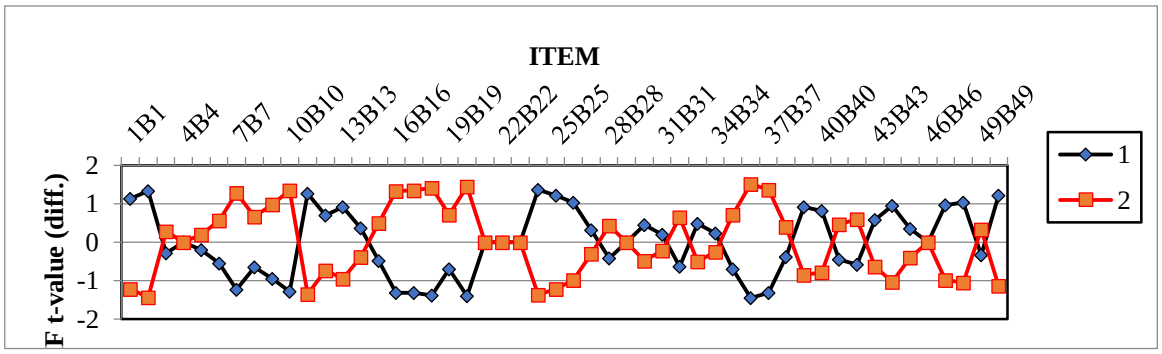

Figure 2: Plot DIF t-Value Word Formation Construct

The Differential Item Functioning (DIF) analysis was carried out on 50 items in Word Formation construct. The analysis show two items have critical *t*- value which are more than ±2.0 logits – B20 (-2.03) and B36 (-2.09). However, after checking the DIF contrast, the values are at ±0.5 logits and they are B20 (0.3) and B36 (-0.33). This is significant since there is no differential item. Table 5 shows critical *t*- value and DIF Contrast of Word Formation construct items which were identified to have more than ±2.0 logits but not higher than ±0.5 logits for DIF contrast. The findings show there is no gender-biased item in Word Formation construct.

3) Syntax

Differential Item Functioning (DIF) Analysis for syntax is shown in Table 6 and Figure 3.

Table 6: DIF Item for Syntax Construct According to Gender

Person Class	*Dif. Measure*	*Person Class*	*Dif. Measure*	*Dif. Contrast*	*t*	*Name*
1	-0.53	2	-0.26	-0.27	1.74	C6
2	-0.26	1	-0.53	0.27	1.74	C6
1	-0.48	2	-0.75	0.27	1.64	C42
2	-0.75	1	-0.48	-0.27	1.64	C42

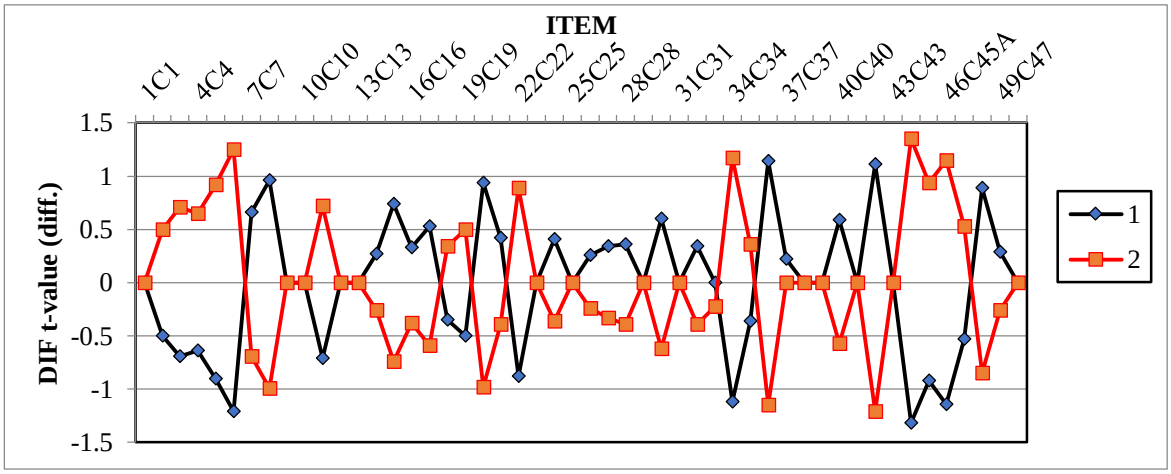

Figure 3: DIF t-Value Plot for Syntax Construct

Differential Item Functioning (DIF) analysis was done on 50 items in the Syntax construct. The findings show critical *t*-values are between 1.64 logits (the lowest) and 1.74 logits (the highest), while DIF Contrast values are between -0.27 logits (the lowest) and 0.27 logits (the highest). All items are identified to have the critical *t*-value not exceeding ±2.0 logits and DIF Contrast below the value of ±0.5 logits which is significant as no differential item. Table 6 shows critical *t*-value and DIF Contrast of the highest Syntax construct items. The findings show no item in this construct is gender-biased.

4) Vocabulary

Differential Item Functioning (DIF) analysis for vocabulary is shown in Table 7 and Figure 4.

Table 7: DIF Item for Vocabulary Construct According to Gender

Person Class	*Dif. Measure*	*Person Class*	*Dif. Measure*	*Dif. Contrast*	*t*	*Name*
1	0.03	2	-0.06	0.09	0.59	D1
1	-0.48	2	-0.4	-0.08	-0.49	D2
1	-0.02	2	-0.02	0	0	D3
1	-0.45	2	-0.51	0.07	0.41	D4
1	0.09	2	-0.19	0.28	1.79	D5
1	0.2	2	0.07	0.14	0.9	D6
1	0.36	2	0.17	0.19	1.23	D7
1	-0.14	2	-0.19	0.04	0.27	D8
1	0.35	2	0.4	-0.05	-0.34	D9A
1	0.58	2	0.64	-0.07	-0.45	D9B
1	0.01	2	0.01	0	0	D10
1	-0.37	2	-0.26	-0.11	-0.67	D11
1	-0.54	2	-0.44	-0.1	-0.62	D12
1	-0.15	2	-0.07	-0.08	-0.54	D13
1	0.54	2	0.81	-0.26	1.78	D14

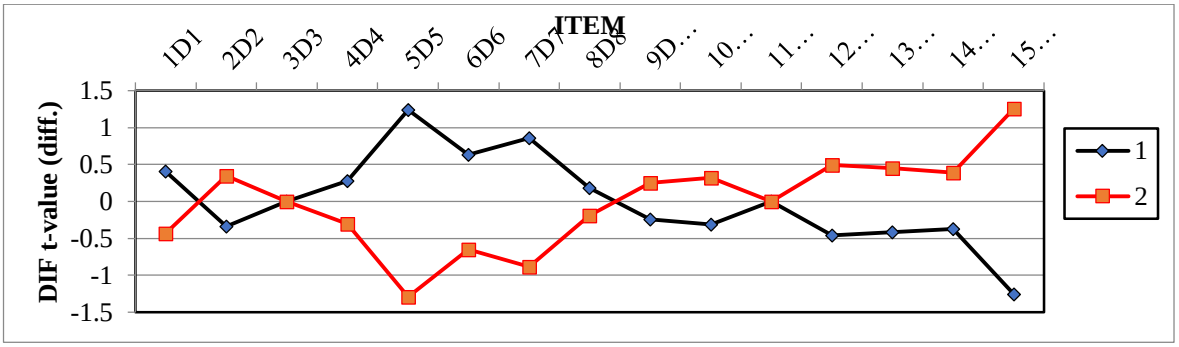

Figure 4: DIF t-Value Plot for Vocabulary Construct

Differential Item Functioning (DIF) analysis was done on 15 items in Vocabulary construct. The findings of the analysis show the critical t-values between –0.67 logits (the lowest) and 1.78 logits (the highest) do not exceed ±2.0 logits. Meanwhile, the DIF Contrast values do not exceed ±0.5 logits. All items found to have critical *t*-value of not more than ±2.0 logits and DIF Contrast not exceeding the value of ±0.5 logits and this is significant as there is no differential item. Table 7 shows the critical *t*-value and DIF Contrast for Vocabulary construct items. The findings show no item is this construct is considered gender-biased.

5) Proverbs

Differential Item Functioning Analysis for Proverbs portrayed in Table 8 and Figure 5.

Table 8: DIF Item for Proverbs Construct according to gender

Person Class	*Dif. Measure*	*Person Class*	*Dif. Measure*	*Dif. Contrast*	*t*	*Name*
1	-0.94	2	-0.68	-0.27	-1.64	D15

2	-0.68	1	-0.94	0.27	1.64	D15
1	-0.85	2	-0.6	-0.25	-1.59	D16
2	-0.6	1	-0.85	0.25	1.59	D16
1	0.13	2	-0.2	0.33	2.26	D26
2	-0.2	1	0.13	-0.33	-2.26	D26
1	-0.66	2	-0.95	0.29	1.76	D32
2	-0.95	1	-0.66	-0.29	-1.76	D32
1	-0.31	2	-0.54	0.22	1.46	D35
2	-0.54	1	-0.31	-0.22	-1.46	D35

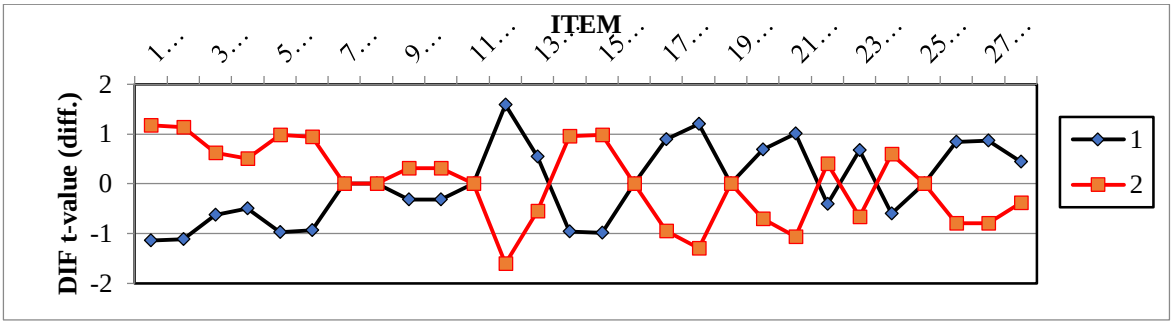

Figure 5: DIF t-Value Plot for Proverbs Construct

Differential Item Functioning (DIF) was done on 28 items in the Proverb construct. The findings show only one item which has critical *t*-value more than ±2.0 logits which is item D26 with the value of -2.26 logits. However, after checking the DIF Contrast, the value was found at ±0.5 logits which is 0.33 logits (significant no differential item). Table 8 shows all items of Proverb construct have critical *t*-value at ±2.0 logits and the DIF Contrast is at the value of ±0.5 logits except for item D26. The findings show no item in Proverb construct is gender-biased.

6) Spelling

Differential Item Functioning (DIF) analysis for Spelling was shown in Table 9 and Figure 6.

Table 9: DIF Item Spelling Construct

Person Class	*Dif. Measure*	*Person Class*	*Dif. Measure*	*Dif. Contrast*	*t*	*Name*
1	-0.37	2	-0.66	0.29	1.85	E4
2	-0.66	1	-0.37	-0.29	1.85	E4
1	-0.21	2	-0.43	0.22	1.46	E9
2	-0.43	1	-0.21	-0.22	1.46	E9
1	-0.36	2	-0.64	0.28	1.75	E10
2	-0.64	1	-0.36	-0.28	1.75	E10
1	-0.33	2	-0.09	-0.23	1.58	E13
2	-0.09	1	-0.33	0.23	1.58	E13
1	0.05	2	0.26	-0.21	1.51	E23
2	0.26	1	0.05	0.21	1.51	E23
1	0.05	2	0.32	-0.27	1.93	E24

2	0.32	1	0.05	0.27	1.93	E24
1	1.11	2	0.89	0.21	1.51	E32
2	0.89	1	1.11	-0.21	1.51	E32

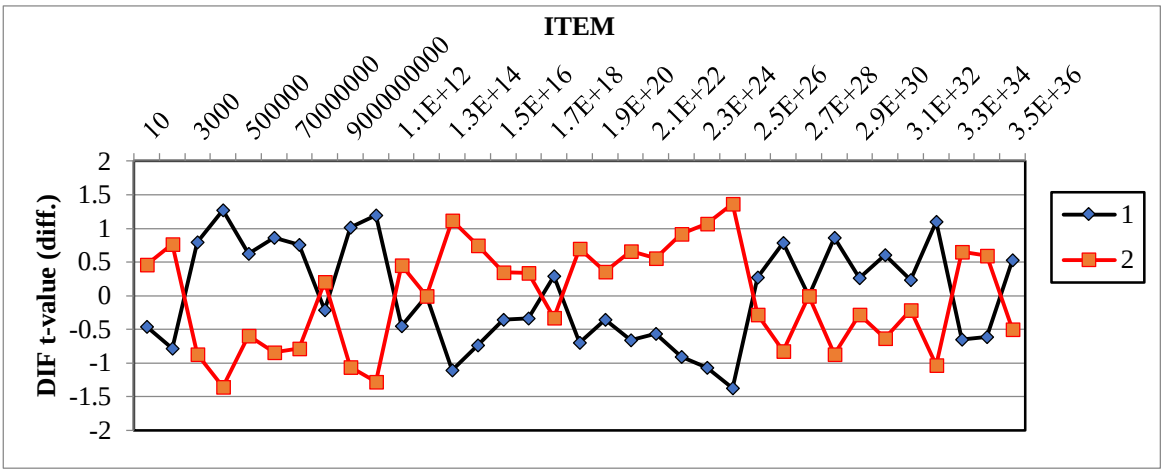

Figure 6: DIF t-Value Plot for Spelling Construct

Differential Item Functioning (DIF) analysis was held on 35 items in Spelling construct. The findings show the critical t-values are between 1.46 logits (the lowest) and 1.93 logits (the highest) which are not higher than ±2.0 logits. Meanwhile, the DIF Contrast values are between -0.28 logits (the lowest) and 0.28 logits (the highest) which are not more than ±0.5 logits. All items found to have critical t-value not more than ±2.0 logits and not less than ±0.5 logits DIF Contrast which is significant no differential item. Table 9 shows critical t-value and DIF Contrast for Spelling construct items. The findings show no item in Spelling construct is gender-biased.

V. CONCLUSION

To summarise, all items in all constructs are not gender-biased. Besides, this instrument has Alpha Cronbach (KR20) reliability indexes, good and acceptable respondents' and items' reliability. Though the respondents' reliability for Proverb, Vocabulary and Spelling are at the level of moderate and less acceptable, the researcher decided to accept the values since the values are still 0.6 which is less acceptable but not rejected. This proves that the items in the instrument are fair and do not favour a particular gender. The findings are good and in contrast with other findings claiming that female students have better performance in most subjects including Bahasa Melayu. This is due to the researcher's mastery and strictness in item building so as to avoid doubts and confusion among the respondents. Besides, the researcher made sure the respondents were those who only had good and moderate performance. Weak and excellent students were not allowed to take the test. This is because their weak reading skills could affect the study's findings in detecting students' weaknesses in mastering the language system. On the other hand, excellent students could most probably answer nearly all questions. Consequently, identifying the causes of their mistakes in mastering the language system would be a failure. These two groups are not eligible to take this diagnostic test to fulfill the main objective of the study which is to produce more accurate findings. Nonetheless, all students were encouraged to take the test to check their own mastery.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

This study benefits lower secondary school students and teachers. Students, educators and parents need the test results. Students can use the study instruments as a self-test or their teachers can help to detect language aspects that they have not yet mastered (Azizi, 2010; Siti Rahayah, 2008). This instrument helps in giving the information on the aspects of language system which students have not mastered based on the answers provided. So, students could focus on and improve their language weaknesses using correct learning strategies with the help of their teachers, parents and friends. Building this test instrument could help solve their problems of time constraint and lack of skills faced by teachers in preparing high quality test tool. In short, this study is important and useful to education specifically Bahasa Melayu subject. Good mastery of Bahasa Melayu will improve the unity of multi-racial community and also the development of the country. The Ministry of Education's vision to develop individuals who could contribute towards the peace, harmony and prosperity of a nation could be realized.

REFERENCES

1. Abdul Ghafar, M. N. (2011). Pembinaan dan Analisis Ujian Bilik Darjah. Johor: Universiti Teknologi Malaysia Press.
2. Ariffin, S. R. (2008). Inovasi Dalam Pengukuran dan Penilaian Pendidikan. Selangor: Universiti Kebangsaan Malaysia.
3. Ahmad, A. (2014). Pentaksiran Pendidikan. Kuala Lumpur: Dewan Bahasa dan Pustaka.
4. Bachman, L. F., Lyle, F., & Palmer, A. S. (1996). Language testing in practice: Designing and developing useful language tests. Oxford University Press.
5. Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences. East Sussex: Psychology Press.
6. Brown, H. D., & Abeywickrama, P. (2010). Principles of language assessment. In Language Assessment: Principles and Classroom Practices. Abeywickrama, P. & Brown, H. D. (Eds.), New York: Pearson Longman pp. 25-51.
7. Ebel, R. L., & Frisbie, D. A. (1991). Essentials of Educational Measurement. New Jersey: Prentice Hall.
8. Fisher Jr, W. P. (2007). Rasch measurement transaction. Transaction of the Rasch Measurement SIG American Educational Research Association, 21(1), 1095.
9. Fisher, W. P. (2007). Rating scale instrument quality criteria. Rasch Measurement Transactions, 21(1), 1095.
10. Gronlund, N. E. (1993). How to make achievement tests and assessments. Boston: Allyn & Bacon.
11. Gronlund, N. E., & Linn, R. L. (1990). Measurement and Evaluation in Teaching. New York: McMillan Publishing Company.
12. Henning, G. (1987). A guide to language testing: Development, evaluation, research. Massachusetts: Newberry House Publishers.
13. Kementerian Pelajaran Malaysia. (2006). Manual Prosedur Pengendalian Ujian Diagnostik Pengajaran-Pembelajaran Sains dan Matematik Dalam Bahasa Inggeris (PPSMI) 2006. Putrajaya: Lembaga Peperiksaan Malaysia.
14. Linacre, J. M. (1997). KR-20/Cronbach Alpha or Rasch person reliability: Which tells the "truth". Rasch Measurement Transactions, 11(3), 580-581.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

15. Linacre, J. M. (1999). Understanding Rasch measurement: Estimation methods for Rasch measures. Journal of Outcome Measurement, 3, 382-405.
16. Linacre, J. M., Stone, M. H., William, J., Fisher, P., & Tesio, L. (2002). Rasch Measurement. Rasch Measurement Transactions, 16, .
17. Linacre, J. M. (2004). Estimation methods for Rasch measures. Introduction to Rasch Measurement, 25-48.
18. Linacre, J. M. (2004). Test validity, and Rasch measurement: Construct, content, etc. Rasch Measurement Transactions, 18(1), 970-971.
19. Linacre, J. M. (2010). Predicting responses from Rasch measures. Journal of Applied Measurement, 11(1), 1-10.
20. Linacre, J. M. (2010). When to stop removing items and persons in Rasch misfit analysis. Rasch Measurement Transactions, 23(4), 1241.
21. Lembaga Peperiksaan Malaysia. (2018). Laporan Peperiksaan SPM 2017. Putrajaya: Lembaga Peperiksaan Malaysia.
22. Nordin, A.B. & Abu Bakar, B. (2008). Pentaksiran dalam Bilik Darjah. Selangor: Longman.
23. Noll, V. H., Scannell, D. P., & Craig, R. C. (1979). Introduction to educational measurement. Massachusetts: Houghton Mifflin Harcourt (HMH).
24. Neukrug, E. S., & Fawcett, R. C. (2014). Essentials of testing and assessment: A practical guide for counselors, social workers, and psychologists. Tennessee: Nelson Education.
25. Wright, B. D., & Stone, M. H. (2004). Making measures. Chicago: Phaneron Press.