Evaluating the Quality of Exam Questions: A Multidimensional Item Response

Faiz Zulkifli, Rozaimah Zainal Abidin, Zulkifley Mohamed

Abstract*— The purpose of this research is to propose a new approach for evaluating the quality of the exam questions. Exam results were obtained from students taking the statistics and probability course in Universiti Teknologi MARA (UiTM). The number of exam questions is set by 10 questions with 30 items that have varying degrees of difficulty. A total of 214 students' results have been extracted from the iCGPA system. "Multidimensional Item Response Analysis (MIRA)" was applied for the 1PL (Rasch), 2PL and 3PL models to evaluate the quality of the exam questions. The models were estimated using MH-RM algorithm in the R package. Model fitting comparison is based on the log-likelihood, SE, AIC and BIC statistics. The statistic and Zh statistic were calculated to identify the item misfit and person misfit. Through model fittings, all three models give the value of all acceptable and almost identical statistic. 5 items are considered as misfit by the 1PL model. For the 2PL and 3PL models, 5 items are categorized as misfit. The reduction in the number of misfit items can be attributed to the addition of information to the IRA model. On the other hand, the analysis of person fit provides different misfit percentages between the IRA models. This is probably because most students can answer all the questions very well. In conclusion, the quality of exam questions for statistics and probability courses needs to be improved by increasing the degree of difficulty of the questions that incorporate higher-order thinking skill.

Keywords: Exam quality, item misfit, multidimensional item response analysis, person misfit.

I. INTRODUCTION

Assessment process has different objectives, such as a candidate classification for educational scholarship or school placement purposes, determining livelihood assistance to those who qualify. Assessment of information can help identify the person or the knowledge of a candidate. Therefore, it is essential for an assessment to be reliable.

In the field of education, the assessment used should reflect the level of a majority of students and may have different objectives from the assessment used in the classroom. It has a strong influence on education policies and curricula at various levels of education globally. Therefore, investigating the variables involved in the construction, application, and achievement of this examination is critical [1].

Literature is often referred to as a performance test that deserves expert assessment [2]. This assessment class is used in various fields, for example, in a speech competition where the speaker conducts some screening tests with the presence of experts in evaluating the quality of the presentation based on determined set of criteria. Similarly, with the test with open answer items, where an individual can provide written response, a performance expert is required to correct the answer. For uniformity, an assessment with an open answer item will be referred to in this work as an open evaluation.

In Malaysia, open evaluations are widely used in each level of educational examinations. Other evaluations with open items are not very common in Malaysia. For example, the Programme for International Student Assessment (PISA) uses open items to grade language proficiency and Mathematics [3].

Open evaluations have been deemed to be difficult since the 1920s, when preliminary studies assessed aspects of limited writing and fluency. However, only in 1950 that teachers and researchers intensified the search of methods that could produce validity and reliability for this assessment [1]. One of the methods that can evaluate the validity and reliability of exam questions is Item Response Analysis (IRA). In [2] had used Rasch model in evaluating final exam questions of Probability and Statistics course. Rasch analysis is the most widely used method by researchers compared to other models in IRA. Commonly used reliability statistic in Rasch are Kuder-Richardson 20 (KR20), KR21 and Spearman-Brown reliability coefficients [4], [5].

The main objective of this study is to introduce a Multidimensional Item Response Analysis (MIRA) in evaluating the quality of the exam questions. In addition to the 1PL (Rasch) model, this study will show the results of the 2PL and 3PL models for IRAs which are rarely employed by researchers. Quality question items are measured through model fitting, item misfit, and person misfit. The data selected in this study were taken from the results of the examination for the statistics and probability course. This course is selected because the type of question item used in the examination is open evaluation.

II. METHODOLOGY

Material

This study utilizes data from statistical and probability student's exam results. The exam questions are structured where students need to write their answers without an answer option. This course is a mandatory subject for the diploma

Revised Version Manuscript Received on September 16, 2019.

Faiz Zulkifli, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, Tapah Road, Perak, Malaysia.

Rozaimah Zainal Abidin, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, Tapah Road, Perak, Malaysia.

Zulkifley Mohamed, Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, Malaysia.

programme in statistics and mathematics at Universiti Teknologi MARA (UiTM). Student performance is evaluated through two types of assessments: continuous and final assessment. The final assessment is the final examination taken at the end of the semester. The preparation of final exam papers is based on the test specification table (TST) that needs to be followed by the question paper makers. TST must be followed to ensure that the examination questions are in accordance with the established standards. Based on TST, the number of final exam questions is 10 questions with 30 items that have varying degrees of difficulty. The question will examine students' understanding of five topics for statistics and probability subject.

The sample data used in this study were taken from three UiTM campuses namely Kelantan, Perak and Pahang. A total of 214 students' results using Cluster Sampling procedure have been extracted from the iCGPA system which contains marks for each question item. In [6] mentioned the sample size did not affect the measurement precision significantly for the 30-item test. The scores obtained need to be standardized first so that the full score for each item is the same. The quality of the exam questions is evaluated based on the model fitting, item misfit, and person misfit of the three types of IRA models, namely, 1PL (Rasch), 2PL and 3PL.

Methods

Item response analysis

Item response analysis (IRA) is widely used in relation to educational analysis for making predictions, estimations, or conclusions of an individual's ability. It is monotonicity where the logistic IRA models assume a monotonically increasing probability form concerning the trait level. [7], listed three types of logistic model of item response analysis that may be used to estimate the ability of an individual. The three models are single parameter model, dual parameters model, and triple parameters model. The three models differ in the number of parameters to calculate and describe the item characteristics. The single parameter model (1PL) is also known as Rasch model which only utilises item difficulty level as a parameter for calculating a person's ability. The dual parameters model (2PL) uses the difficulty level and the discrimination of item. The triple parameters model (3PL) uses the item difficulty level, discrimination, and the guessing the correct answer as parameters.

Item response theory with one parameter logistic model (1PL):

$$P(X=1|\theta,b)=\frac{e^{(\theta-b)}}{1+e^{(\theta-b)}} \quad (1)$$

In this model, θ represents examinee trait level. The parameter location shown in ability scale is named as item difficulty level (b). It is defined as the amount of the latent trait needed to have 0.5 probability of endorsing the item.

Item response theory with two parameters logistic model (2PL):

$$P(X=1|\theta,a,b)=\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}} \quad (2)$$

In this model, item response function varies both in their discrimination and difficulty (i.e.; location) parameters. An item discrimination (a) indicates how well the item can discriminate between a person with the latent trait ability.

Item response theory with three parameters logistic model (3PL):

$$P(X=1|\theta,a,b,c)=c+(1-c)\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}} \quad (3)$$

In this model, the inclusion guessing (c) suggests that respondents who are very low on the trait may choose the correct answer. In other words, respondent with low trait levels may still have a small probability of endorsing an item.

For [8] multidimensional ordinal response model, it is assumed that there are k_j unique categories for item j. Here we define that the boundaries lead to the conditional probability for the response $x_{ij}=k$ for 1PL, 2PL and 3PL to be

$$\Phi(x_j=k)=P(x_j\geq k)-P(x_j\geq k+1)$$

where $P(x_j)$ is obtained from (1), (2), and (3).

The IRA models used in this study involved high-dimensionality problems. The parameter estimation of the models needs to employ stochastics estimation methods. The *R* package can be used to estimate the parameters of multidimensional confirmatory item factor analysis methods. One of the methods that is implemented in the *mirt* package [9] is the Metropolis-Hastings Robbins-Monro (MH-RM) method for exploratory [10] and confirmatory [11] polytomous models.

The $s-\lambda^2$ statistic for MIRA model

Item-fit analysis has been proposed as a method of locating extraneous dimensions affecting the responses to test items. The selection of the model is important since it is important to identify the most effective test model that maintains the integrity of the observed data. In addition, item-fit is able to identify item misfit. Several alternative item fit indices have been proposed such as χ^2, Q statistic, G^2, $S-\chi^2$, scaling corrected χ^{2*}, and adjusted χ^2 degrees of freedom ratios $(\chi^2/df s)$. All the indices specified are calculated by examinees by their level of ability. The frequency of observed response options is compared with the expected frequency based on the estimated IRA model. Therefore, in this study, the $S-\chi^2$ statistic proposed by [12] are used to determine the item misfit.

A Chi-square fit statistic comparing the observed and expected responses is computed as;

$$S-\lambda^2=\sum_{K=1}^{K}\frac{N_k\left(O_{ik}-E_{ik}\right)^2}{E_{ik}\left(1-E_{ik}\right)}$$

where

N_k is the number of counts in score category k.

O_{ik} directly counted from the item response matrix.

E_{ik} is the expected proportion of correct responses to item i under total score k.

The Z_h statistic for MIRA model

Person fit is a broad set of statistical methods used to identify the congruence of examinees' response patterns with the specified response model [13]. When a pattern is incongruent, it is assumed that a person randomly selects a response to items get the end of the question faster which would contribute person misfit since guessing is one of the mechanisms of selecting items. Several specific measurement disturbances have been associated with item response aberrancy. For example, in [14] identified that misfit may occur due to different reasons such as cheating, guessing or taking a different approach to answering questions.

Person-fit statistic can be used to identify response patterns associated with specific educational deficits or to extract deviant response patterns to create a more nearly unidimensional data matrix. Besides that, person fit has the potential value as response consistency/validity indices in personality measurement [15].

The Z_h statistic is a standardised person fit value of lz, which is generalised to categorical data [16]. The categorical lz statistic can be defined as:

$$lz = P\left(Y\left|\theta_i\right.\right) = \sum_{i=1}^{n}\sum_{j=1}^{A+1}\delta_j\left(v_i\right)\log Pg\left(\theta\right)$$

where,

Y_i is the item responses, θ_i contains the latent trait estimates, n is sample size, A is the number of j categories in item i. $\delta_j\left(v_i\right)$ is used to ensuring that only the probabilities of the chosen responses are summed. Thus, $\delta_j\left(v_i\right)$ is set equal to 1 when $j = k$, and it is set to 0 when $j \neq k$. Finally, the standardized form of lz can be defined as Z_h with the following transformation:

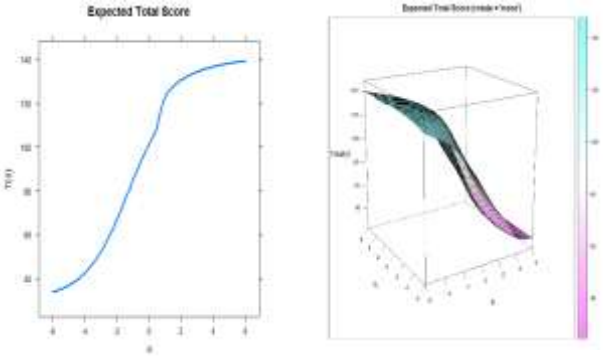
$$Z_h = \left[lz\left(\theta\right) - E\left(lz\left(\theta\right)\right) / SD\left(lz\left(\theta\right)\right)\right]$$

where $E\left(lz\left(\theta\right)\right)$ is the mean lz value for the sample and $SD\left(lz\left(\theta\right)\right)$ is the standard deviation of the observed lz value.

Typically, the distribution of Z_h is non-normal. Due to the non-normality of Z_h distribution, a hard cut-off of -1.69 should not be used as conventional Z distribution for making inferences. Many researchers create a cut off ± 2.0 to serve as the starting point in identifying outlier responses. Similar to the parameter estimations of the models, item fit, and person fit were estimated in the R programming with multidimensional IRA package.

III. RESULTS AND DISCUSSION

The analysis of this study was initiated by reviewing the assumption of IRA model to the data sample. Using *plot ()* with '*itemscore*' type, the plot of the expected total score against model parameters is shown in Fig. 1. Plot (a) is for the 1PL model while plot (b) is a 2PL model. *Plot ()* function can only show plot for 2 and 3 dimensions. Both plots show an upward trend exponentially. This demonstrates the monotonicity assumption for the IRA model was successfully followed.



parameters

Model fitting

The final exam for a statistical course typically has structured formats and the scores given can be categorized as an ordinal scale. The categories have more than two options or are known as polytomous. The estimation of multidimensional item factor analysis uses the MH-RM algorithm. Through *mirt ()* with '*MHRM*' method, model fitting can be obtained. The AIC and BIC tests are used for model comparison purposes against the same data. Model fitting results of three IRA models are summarized in Table 1. All the models are significant at the 5% significance of level. The estimation of MHRM takes longer if the IRA model has many factors in which the higher iterations are shown by the 2PL and 3PL models than the 1PL model.

Based on Table 1, the log-likelihood, SE, AIC, and BIC values become smaller when the number of parameters for the IRA model is added. However, the differences shown are not significantly different. This shows that all the three models are appropriate to evaluate the quality of the final exam questions. The 2PL model is generated through the addition of guessing parameters into the 1PL model. The 3PL model has an additional discrimination parameter to the 2PL model. A detailed analysis of difficulty, guessing and discrimination estimates is not discussed in this study. However, the existence of these parameters can contribute useful information for the IRA model.

Table 1: Model fitting results of IRA model

	Model		
	1PL	**2PL**	**3PL**
MHRM Iteration	796	858	1234
Log-likelihood	-5907.471	-5770.159	-5682.56
SE	0.107	0.105	0.105
AIC	12653.11	11880.32	11761.12
BIC	12124.75	12452.53	12427.58

Item misfit

The items for the exam questions are high in quality when the items can evaluate and categorize the ability of students into low, moderate, or excellent level. Hence, questions supplied to students should have varying degrees of

difficulty. To identify item misfit, an analysis of the plot of expected item scoring function and the $s-\chi^2$ statistic was implemented. The expected item scoring function was plotted against the ability parameter for the 1PL model shown in Fig. 2. The figure is generated by the function of *plot ()* with '*itemscore*' type. However, the same diagram cannot be obtained for the 2PL and 3PL models. Roughly, items that can be categorized as misfit items are 13, 14, 15 and 3. This is because some of these items cannot evaluate the overall student's abilities. In addition, the expected item scoring patent for Item 3 shows a flat trend, leading to a low score. This means that students are expected to not be able to answer the Item 3 despite having high abilities. Other than the items that were categorized as misfit, the expected item scoring patent shows monotonic trends which demonstrates that the items can test the ability of most students.

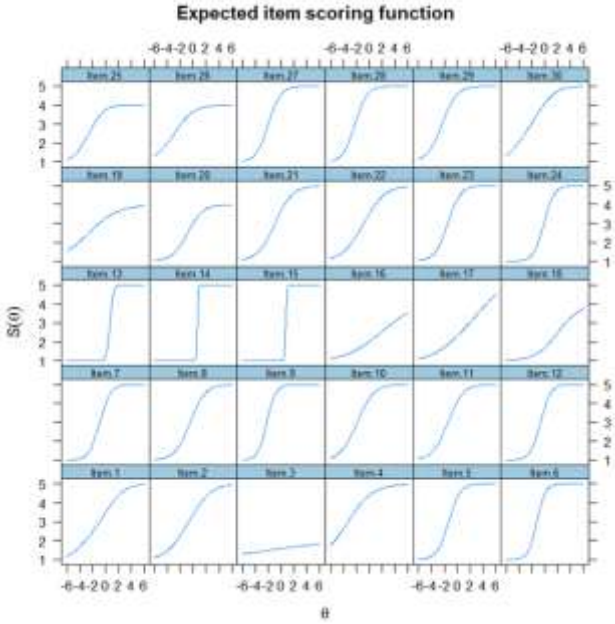


Fig. 2: Plot the expected item scoring vs ability parameter

The item misfit analysis using the plot only refers to the patent of the expected item scoring function without considering the overall performance of the student. This study will continue to analyze the item misfit based on the RMSE and P-value values for $s-\chi^2$ statistic on the 1PL, 2PL and 3PL models. These values are shown in Table 2 by itemfit () with 'S_X2' type. Items with high RMSE values with the p-value less than 0.05 (5% significance level) are categorized as misfit. There are 8 items for the 1PL model with p-value less than 0.05. Meanwhile, 5 items are categorized as misfit in the 2PL and 3PL models. The reduction in the number of misfit items compared to the 1PL model can be attributed to the addition of information to the IRA model fitting by considering the possibility of the students answering correctly and appropriating the guessing. Items categorized as misfit in all three models which are Items 8, 17, 21, 23 and 24 need to be reviewed in terms of question content and difficulty level. Such items shall be removed or rephrased before being part of the question bank.

Table 2: RMSE and p-value $s-\chi^2$ statistic

Item	1PL		2PL		3PL	
	RMSE	P-value	RMSE	P-value	RMSE	P-value
1	0.0000	0.6546	0.0000	0.5983	0.0000	0.5198
2	0.0253	0.2017	0.0220	0.2599	0.0098	0.4308
3	0.0149	0.3880	0.0112	0.4235	0.0056	0.4583
4	0.0000	0.4947	0.0178	0.3775	0.0241	0.3198
5	0.0032	0.4606	0.0000	0.5755	0.0000	0.4773
6	0.0000	0.5830	0.0145	0.4022	0.0436	0.1171
7	0.0156	0.3787	0.0257	0.2478	0.0337	0.1543
8	0.0495	0.0086	0.0548	0.0037	0.0571	0.0023
9	0.0374	0.2468	0.0517	0.1517	0.0712	0.0806
10	0.0000	0.6871	0.0000	0.6738	0.0000	0.5118
11	0.0000	0.6016	0.0000	0.5849	0.0000	0.5179
12	0.0000	0.9077	0.0000	0.8772	0.0000	0.8540
13	0.0842	0.0002	0.0000	0.5875	0.0479	0.0655
14	0.1539	0.0000	0.0000	0.4693	0.0418	0.1444
15	0.2042	0.0000	0.0244	0.3173	0.0472	0.1113
16	0.0342	0.0964	0.0344	0.0919	0.0362	0.0778
17	0.0442	0.0318	0.0458	0.0243	0.0465	0.0212
18	0.0150	0.3924	0.0028	0.4631	0.0121	0.4184
19	0.0338	0.1653	0.0242	0.2897	0.0333	0.1747
20	0.0000	0.6719	0.0000	0.6551	0.0000	0.6067
21	0.0475	0.0335	0.0496	0.0276	0.0530	0.0172
22	0.0000	0.5605	0.0000	0.5587	0.0000	0.5804
23	0.0488	0.0367	0.0488	0.0414	0.0554	0.0177
24	0.0599	0.0009	0.0589	0.0016	0.0610	0.0010
25	0.0000	0.6241	0.0107	0.4283	0.0187	0.3691
26	0.0000	0.6284	0.0000	0.5758	0.0000	0.5413
27	0.0000	0.5320	0.0000	0.5726	0.0252	0.3111
28	0.0170	0.3845	0.0390	0.1601	0.0407	0.1493
29	0.0415	0.1309	0.0485	0.0788	0.0355	0.2026
30	0.0111	0.4211	0.0228	0.2683	0.0124	0.4082

Person misfit

In ensuring that the model is well fitting, people who are not liked by the model need to be identified to increase the confidence in the analysis of model fitting. The individual who is said to be unsuitable for the model may also be categorized as outliers. The person misfit analysis in this study will use Zh statistic by personfit () function. The value of Zh for a part of person is shown in Table 3. Person who has a lower Zh value than -2 is said to be the highest misfit score while Zh value of more than 2 indicates the person is over fitting. A person is considered well-fitting when the Zh value approaches zero. Based on the analysis, the 1PL model complements 6.07% of the person misfit and 4.21% over fitting. The 2PL model recorded no person over fitting but had 32.24% misfit. A total of 27.1% of the 3PL models are in the category of misfit and only one person is over fitting.

Based on the analysis of the misfit person, only the 1PL model provides the lowest number of misfits. The high percentage of misfit shown by the 2PL and 3PL models can cause the model's fitting disproportionate result. However, the score given by the person misfit needs to be further examined before the refusal of the 2PL and 3PL models is made. Outlier performance on student academic performance occurs when a student obtains a high score for all question items or scores too low. Based on an observation of the

person misfit score, it is found that most students obtained a high score on all items of the question, implying students can answer all the questions very well. The result of this observation received support by [2] which states that the students feel the final examination questions given are too easy in which the difficulty of the question does not match the level of student's ability

Table 3: Zh statistic value of IRA model

1PL		2PL		3PL	
Person	Zh	Person	Zh	Person	Zh
MISFITTING		MISS FITTING		MISS FITTING	
54	-7.0325	153	-10.7064	54	-14.8823
15	-6.0582	178	-8.8278	68	-13.7516
108	-4.6789	17	-6.8667	99	-13.3064
153	-3.4674	54	-6.6674	30	-11.9326
68	-3.2827	12	-6.2615	155	-11.8167
27	-3.2633	181	-6.2465	161	-11.3715
30	-3.1842	28	-6.1182	36	-11.3274
87	-2.8053	197	-6.035	85	-11.1726
165	-2.6417	63	-6.0244	88	-10.4858
156	-2.4413	65	-5.9186	106	-9.6339
100	-2.4211	36	-5.8543	86	-9.5776
58	-2.3145	8	-5.6654	33	-8.9026
8	-2.1426	108	-5.4567	96	-7.7821
WELL FITTING		45	-5.4114	15	-7.5249
.	.	212	-5.4018	87	-7.4798
.	.	15	-5.3347	119	-7.3334
.	.	191	-5.1348	41	-7.1296
184	-0.384	87	-5.0665	12	-6.3576
173	-0.3596	128	-4.9913	111	-6.3299
206	-0.3542	185	-4.8783	58	-6.0606
211	-0.3442	20	-4.477	108	-6.0524
45	-0.3436	56	-4.4212	153	-5.7127
75	-0.2818	172	-4.4192	135	-5.6884
150	-0.2794	31	-4.2911	64	-5.3582
204	-0.2515	114	-4.1343	105	-5.213
84	-0.2104	170	-4.1325	17	-5.1527
127	-0.1978	111	-4.1215	144	-5.1512
176	-0.1511	183	-4.0715	212	-4.9552
18	-0.1479	171	-4.0476	140	-4.7774
131	-0.1408	27	-3.9975	28	-4.4167
177	-0.1202	41	-3.9306	157	-4.365
123	-0.1065	148	-3.9258	72	-4.1217
37	-0.0655	100	-3.8634	25	-4.1073
183	-0.0435	162	-3.8321	165	-3.9888
16	-0.0265	165	-3.8294	89	-3.904
70	-0.0039	33	-3.7701	194	-3.8202
56	0.0146	105	-3.6417	137	-3.6968
188	0.0175	82	-3.6034	124	-3.2723
193	0.0469	50	-3.5085	38	-3.2388
83	0.059	145	-3.4595	197	-3.1536
78	0.0795	7	-3.428	181	-3.0146
116	0.1113	70	-3.2958	141	-2.9352
112	0.1168	209	-3.2577	125	-2.8189
154	0.1181	194	-3.221	74	-2.8183
179	0.1344	154	-3.152	120	-2.7197
117	0.2299	155	-3.0479	156	-2.6392
62	0.2354	175	-3.0404	27	-2.6297
47	0.2379	2	-3.0252	77	-2.3944
107	0.2481	144	-2.726	45	-2.3563
160	0.2716	30	-2.6847	20	-2.3216

2	0.2717	173	-2.6405	56	-2.3049
22	0.2768	187	-2.6241	185	-2.2516
48	0.2986	205	-2.6169	123	-2.2291
94	0.2987	152	-2.5412	31	-2.2259
134	0.3255	157	-2.5182	118	-2.2113
67	0.3422	177	-2.4723	136	-2.1737
180	0.3573	26	-2.4668	170	-2.1607
139	0.3577	52	-2.4203	2	-2.0659
46	0.392	130	-2.4182	WELL FITTING	
53	0.3941	210	-2.4135	.	.
98	0.4065	3	-2.3845	.	.
35	0.4372	124	-2.3763	.	.
55	0.4388	156	-2.3643	13	-0.1068
190	0.4443	77	-2.3057	142	-0.0926
152	0.4514	68	-2.2923	103	-0.0805
104	0.4545	57	-2.1742	139	-0.0703
.	.	19	-2.0588	175	-0.0236
.	.	123	-2.0196	158	-0.0016
.	.	206	-2.0176	90	0.0072
OVER FITTING		WELL FITTING		121	0.0191
91	2.0046	.	.	188	0.054
43	2.0104	.	.	104	0.0551
44	2.0267	.	.	202	0.0714
9	2.0409	80	-0.0196	39	0.1165
41	2.0557	39	-0.0194	.	.
161	2.0634	167	0.0326	.	.
99	2.2469	.	.	.	.
86	2.5808	.	.	OVER FITTING	
141	2.6347	.	.	179	2.0893

IV. CONCLUSION

Based on the analysis results, the MIRA model for 1PL, 2PL and 3PL can be used to evaluate the quality of exam papers. The quality of the examination papers for statistics and probability subjects should be improved so that all students are tested for their ability to answer questions with varying degrees of difficulty. The question maker should not only take into consideration the weaker student's ability but also challenge excellent students' ability. This leads to stunted students' thinking skills because they are not sharpened to solve questions that require high-level thinking skills. In addition to relying on TST, the exam questions banks need to undergo thorough validation by experts in the field. It is intended to ensure non-quality questions are not used again in the examination. Further studies can be done on big data or simulation data so that they can illustrate the study population. The use of machine learning can speed up modeling and produce a more accurate analysis.

V. ACKNOWLEDGMENT

Researchers would like to express their sincerest gratitude towards UiTM and Universiti Pendidikan Sultan Idris (UPSI) for providing the opportunity to conduct this study.

REFERENCES

1. N. Behizadeh and G. Jr. Engelhard, "Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States," Assessing Writing, 16(3), 2011, pp. 189-211.
2. F. Zulkifli, R. Z. Abidin, N. F. M. Razi, N. H. Mohammad, R. Ahmad, and A. Z. Azmi, "Evaluating quality and reliability of final exam questions for probability and statistics course using Rasch model," International Journal of Engineering and Technology (UAE), 7(4), 2018, pp. 32-36.
3. J. Jerrim, M. B. Oliver, and S. Sims, "The relationship between inquiry-based teaching and students' achievement: New evidence from a longitudinal PISA study in England," Learning and Instruction, 61, 2019, pp. 35-44.
4. N. Lohgheswary, Z. M. Nopiah, and E. Zakaria, "Evaluating the reliability of pre-test differential equations questions using Rasch measurement model," Journal of Engineering Science and Technology, 11, 2016, pp. 31-39.
5. H. Othman, N. A. Ismail, I. Asshaari, F. M. Hamzah, and Z. M. Nopiah, "Application of Rasch measurement model for reliability measurement instrument in vector calculus course," Journal of Engineering Science and Technology, 10, 2015, pp. 77-83.
6. M. Şahin and Y. Yildirim, "The examination of item difficulty distribution, test length and sample size in different ability distribution," Journal of Measurement and Evaluation in Education and Psychology, 9, 2018, pp. 277-294.
7. A. Birnbaum, "Some latent trait models and their use in inferring an examinee's ability," Statistical Theories of Mental Test Scores, 1968, pp. 397-479.
8. F. Samejima, "Estimation of ability using a pattern of graded responses," Psychometrika Monograph, 17, 1969, pp. 1–100.
9. R. P. Chalmers, "mirt: A multidimensional item response theory package for the R environment," Journal of Statistical Software, 48(6), 2012, pp. 1-29.
10. L. Cai, "High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm," Psychometrika, 75(1), 2010, pp. 33-57.
11. L. Cai, "Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis," Journal of Educational and Behavioral Statistics, 35(3), 2010, pp. 307-335.
12. M. Orlando, and D. Thissen, "Likelihood-based item-fit indices for dichotomous item response theory models," Applied Psychological Measurement, 24, 2000, pp. 50-64.
13. J. M. Felt, R. Castaneda, J. Tiemensma, and S. Depaoli, "Using person fit statistics to detect outliers in survey research," Frontiers in Psychology, 8, 2017, pp. 863.

15. B. D. Wright, "Misunderstanding the Rasch model," Journal of Educational Measurement, 14(3), 1977, pp. 219-225.
16. K. K. Tatsuoka, and M. M. Tatsuoka, "Detection of aberrant response patterns and their effect on dimensionality," Journal of Educational Statistics, 7(3), 1982, pp. 215-231.
17. F. Drasgow, M. V. Levine, and E. A. William, "Appropriateness measurement with polychotomous item response models and standardized indices," Br. J. Math. Stat. Psychol., 38, 1985, pp. 67–86.