EDUCATIONAL RESOURCES
Volume 9 Issue 3 2017
DOI: 10.21315/eimj2017.9.3.8
ARTICLE INFO
Submitted: 03-04-2017
Accepted: 02-06-2017
Online: 30-09-2017

Item Response Theory for Medical Educationists

Wan Nor Arifin[1], Muhamad Saiful Bahri Yusoff[2]

[1]*Unit of Biostatistics and Research Methodology,*
[2]*Department of Medical Education,*
School of Medical Sciences, Universiti Sains Malaysia, MALAYSIA

To cite this article: Arifin WN, Yusoff MSB. Item response theory for medical educationists. Education in Medicine Journal. 2017;9(3):69–81. https://doi.org/10.21315/eimj2017.9.3.8

To link to this article: https://doi.org/10.21315/eimj2017.9.3.8

ABSTRACT

Item analysis (IA) is commonly used to describe difficulty and discrimination indices of multiple true-false (MTF) questions. However, item analysis is basically a plain descriptive analysis with limited statistical value. Item response theory (IRT) can provide a better insight into the difficulty and discriminating ability of questions in a test. IRT consists of a collection of statistical models that allows evaluation of test items (questions) and test takers (examinees) at the same time. Specifically, this article focuses on two-parameter logistic IRT (2-PL IRT) model that is concerned with estimation of difficulty and discrimination parameters. This article shows how 2-PL IRT analysis is performed in R software environment, guides the interpretation of the IRT results and compares the results to IA on a sample of MTF questions.

Keywords: *Item analysis, Item response theory, Difficulty, Discrimination*

CORRESPONDING AUTHOR Dr. Wan Nor Arifin, Unit of Biostatistics and Research Methodology, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia | E-mail: wnarifin@usm.my

INTRODUCTION

Education testing and assessment are among the most important elements in any educational program (1). In the past few decades, many medical education programs and licensing authorities at undergraduate and postgraduate levels have devoted their efforts at ensuring validity of assessments and competency of trainees (2–4).

Each assessment method has its advantages and disadvantages. The best assessment method should fulfill five criteria: validity, reliability, acceptability, feasibility and educational impacts on learning and practice (5). Validity can be defined as the degree to which an assessment measures the characteristics it is meant to measure (6–8). Validity can be proved by gathering sources of evidence to support validity in forms of content, response process, internal structure, relations to other variables and consequences (7). Interested readers may refer to paper written by Cook and Beckman (7) for detailed descriptions of each source of validity evidence.

Epstein (2) recommended four actions to improve the validity of an assessment, which are:

1. Clear expectation of an assessment,
2. Clear learning outcomes to be measured,
3. Familiar with the advantages and disadvantages of an assessment tool, and

4. Continuous evaluation and monitoring of assessment quality to avoid the unwanted effects.

These actions ensure the assessment items are relevant, understandable and discriminating between overall best and overall worst candidates (9).

One of the assessment quality measures commonly performed by medical educators is item analysis (IA), which allows assessment of the effectiveness of individual test items (10). IA typically relies on the classical test theory (CTT) with two major statistics based on difficulty and discrimination indices that depend on students' score (11).

Item response theory (IRT) can provide a better insight into the difficulty and discriminative ability of questions or test items in a test. IRT allows evaluation of test items and test takers at the same time (8). IRT is basically a collection of statistical models that allows analysis of categorical responses (dichotomous, polytomous) (8, 12). Of importance, two-parameter logistic IRT (2-PL IRT) model is concerned with estimation of difficulty and discrimination parameters of the items (8), also known as location (*b*) and slope (*a*) parameters (13). IRT analysis also allows evaluation of the items by looking into item characteristic curve and item information, and the test as a whole by evaluating test information and test characteristic curve (8, 14). Additionally, IRT model fit (by item fit and goodness-of-fit for two-way margins) and basic unidimensionality assumption can be evaluated (8, 12, 14).

The main aims of this article are to show how 2-PL IRT analysis can be performed in R software environment, guide the interpretation of the IRT results and compare the results to IA on a sample of multiple true-false (MTF) questions. Readers are encouraged to read more about IRT in (8, 13, 14). For the purpose of demonstrating the IA and IRT analysis, **mtf.csv** data set is used. The data set consists of two MTF questions, with five independent statements each. The correct response is coded as 1, and incorrect response is coded as 0. The readers may download the data set from the URL provided at the end of this article.

ITEM ANALYSIS

Difficulty index (*P*) of an item (an MTF question) is the proportion of students who answered the item correctly (9, 10, 15). The optimal range is 0.2 to 0.8; a low index may mean that students are attempting the item but are getting it wrong and a too high index may mean that regardless of poor or good students able to get it correct (9). The formula for *P* is given by:

$$P = \frac{R}{T} \qquad (16)$$

where

R = number of correct responses

T = total number of responses

Discrimination index (*D*) is a measure, of how the 'overall best' students are doing versus the 'overall worst' students on a particular item (10, 17). The *D* of an item is the degree to which the item discriminates between those who scored high and those who scored low on a test (15). Specifically, it is the difference in the proportion of students who answered the item correctly in the upper group (P_U, top 27% performers of a test), and the proportion of students who answered the item correctly in the lower group (P_L, bottom 27% performers of a test) (16). The value of *D* ranges between 1 (all of the top 27% versus none of the bottom 27% answered correctly) to –1 (all of bottom 27% versus none of the top 27% answered correctly) (10, 17). The formula for *D* is given by:

$$D = P_U - P_L \qquad (16)$$

where

$$P_U = \frac{R_U}{T_U}$$

R_U = number of correct responses in the upper group

T_U = total number of responses in the upper group

and

$$P_L = \frac{R_L}{T_L}$$

R_L = number of correct responses in the lower group

T_L = total number of responses in the lower group

The calculation can be easily performed in any spreadsheet program. The readers may download **mtf_IA.xls** file from the URL provided in the Notes section of this article to see how it is performed in the spreadsheet.

Both the difficulty and discrimination indices are used to decide on the quality of items in a test. A test should not consist of too difficult or easy items, and should be able to discriminate high scorers from low scorers. In that respect, the items can be evaluated based on a number of cutoff values as presented in Table 1.

Table 1: Classification of difficulty and discrimination values by analysis approaches

Parameters	IA	2-PL IRT
Difficulty	Range (9): <0.20: Difficult 0.20 to 0.80: Optimum >0.80: Easy	Range (14, 18): <–2.0: Easy –2.0 to 2.0: Average >2.0: Hard In practice, the values typically range from –3 to +3 (14, 18)
Discrimination	Range (17): <0.20: Poor 0.20 to 0.40: Acceptable >0.40: Very good	Range (14): 0: None 0.01 to 0.34: Very low 0.35 to 0.64: Low 0.65 to 1.34: Moderate 1.35 to 1.69: High >1.70: Very high + infinity: Perfect 0.8 to 2.5: Good (18)

IA = item analysis, 2-PL IRT = 2-parameter logistic model of item response theory.

IRT ANALYSIS

R is a multi-platform and free software environment for statistical computing (19). It is recommended to use RStudio (20), which is a free user interface for R. It provides an integrated working environment to working with R codes, viewing results and graphics. The authors used RStudio to prepare the R codes for this article. In this article, the analysis is demonstrated using three R packages: **psych** (21), **ltm** (22) and **irtoys** (23) packages. The authors used the packages because of they are relatively easy to use. However, the readers may also consider learning IRT analysis using **mirt** (24), which offers more advanced options (e.g. multidimensional IRT, more model fit assessment indices). The readers are required to learn basic skills on R and RStudio on their own before trying out the R codes (i.e. installation of R and RStudio, basic R commands and RStudio interface). The full R codes with relevant in-line comments are provided in **irt_mtf.R**. The file can be downloaded from the URL

provided at the end of this article. The codes are simplified in five steps.

Step 1. Install and load the required libraries (packages). **psych**, **ltm** and **irtoys** packages are required.

```
install.packages(c("psych", "ltm", "irtoys"))
```

Load **psych**, **ltm** and **irtoys** packages.

```
library("psych")
```

```
library("ltm")
```

```
library("irtoys")
```

Step 2. Read data set **mtf.csv** into data.mtf data frame, then preview the data set.

```
data.mtf = read.csv("mtf.csv", header = TRUE)
```

View the first six responses and list the variable names,

```
head(data.mtf)
```

	Q1A	Q1B	Q1C	Q1D	Q1E	Q2A	Q2B	Q2C	Q2D	Q2E
1	1	0	0	0	0	0	1	1	0	0
2	1	0	0	0	1	0	0	1	1	1
3	0	1	0	0	1	1	0	1	1	0
4	1	1	0	1	1	0	1	0	1	1
5	1	1	1	0	1	1	1	1	1	0
6	0	1	1	1	1	0	1	1	1	1

```
names(data.mtf)
```

```
[1] "Q1A" "Q1B" "Q1C" "Q1D" "Q1E" "Q2A" "Q2B" "Q2C" "Q2D" "Q2E"
```

Check the "dimension" of the data set, i.e. the number of rows and columns. In our data set, there are 10 variables and 160 responses,

```
dim(data.mtf)
```

```
[1] 160  10
```

Step 3. Obtain percentage of correct responses by items. The results are similar to the values of *P* obtained by IA (column "1").

```
response.frequencies(data.mtf)
```

	0	1	miss
Q1A	0.30625	0.69375	0
Q1B	0.25625	0.74375	0
Q1C	0.37500	0.62500	0
Q1D	0.40625	0.59375	0
Q1E	0.16250	0.83750	0
Q2A	0.25000	0.75000	0
Q2B	0.26875	0.73125	0
Q2C	0.34375	0.65625	0
Q2D	0.47500	0.52500	0
Q2E	0.48125	0.51875	0

Step 4. In this step, the 2-PL IRT analysis is performed mainly using **ltm** package.

Perform the analysis on **data.mtf** using **ltm()** command, then save the results in **irt.mtf** data frame,

```
irt.mtf = ltm(data.mtf ~ z1, IRT.param = TRUE)
```

Obtain the difficulty and discrimination parameter estimates,

```
coef(irt.mtf)
```

	Dffclt	Dscrmn
Q1A	-1.34813444	0.6637809
Q1B	-4.20384339	0.2572201
Q1C	-0.40398852	2.0871272
Q1D	-0.53216018	0.8114138
Q1E	-3.96868564	0.4283655
Q2A	-2.64619595	0.4320399
Q2B	-2.05447180	0.5154989
Q2C	-1.06267344	0.6670853
Q2D	-0.13935513	0.8074819
Q2E	-0.09428061	0.9122159

The results can be interpreted according to Table 1.

Next, plot item characteristic curves (ICCs) of the test, which are also known as item response functions (IRFs),

```
plot(irt.mtf, type = "ICC", legend = TRUE)
```

An item characteristic curve shows the relationship between the ability level (θ) and the probability of responding correctly. In

our case, a curve is plotted based on 2-PL IRT model equation for specific difficulty and discrimination values of an item. Note that the difficulty is the location (*b*) where 50% of respondent with *b* ability level will answer the item correctly (8, 14). For example, 50% of respondent with –0.40 ability level will answer Q1c correctly (try to draw a line from 0.5 on *y*-axis to the ICC for Q1c and look for the corresponding value on *x*-axis). Also note that, the slope at *b* ability level is steep when the discrimination (*a*) value is large.

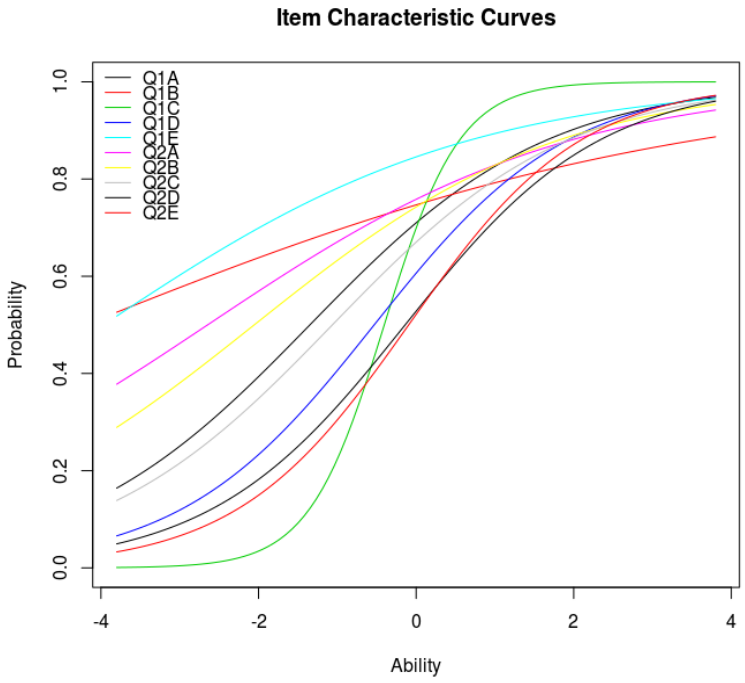

Figure 1: Item characteristic curves of the test.

Plot item information curves (IICs) of the test, which are also known as item information functions (IIFs),

```
plot(irt.mtf, type = "IIC", legend
= TRUE)
```

Plot test information function (TIF). Note the items = 0 command parameter, that instructs the plot() to give TIF instead of IIC,

```
plot(irt.mtf, items = 0, type =
"IIC")
```

Figures 2 (items) and 3 (test) are concerned with amount of information at a particular ability level. For example, inspection of Figure 2, Q1c shows that maximum item information is obtained at –0.40 ability level, which corresponds to the item difficulty. This indicates that the item is most precise at estimating –0.40 ability level, and becomes less precise as we move to the right or left of this ability level. It is observed in Figure 2 that the information is close to zero at –3 and +2 ability levels, which means that the item lacks precision in measuring ability beyond these two ability levels. This interpretation also applies to Figure 3 that shows the information curve for the test (note that the test is most precise at around –0.5 ability level).

Then, estimate the amount of information that can be obtained by the test between –3 and +3 range of ability,

```
information(irt.mtf, c(-3,3))
```

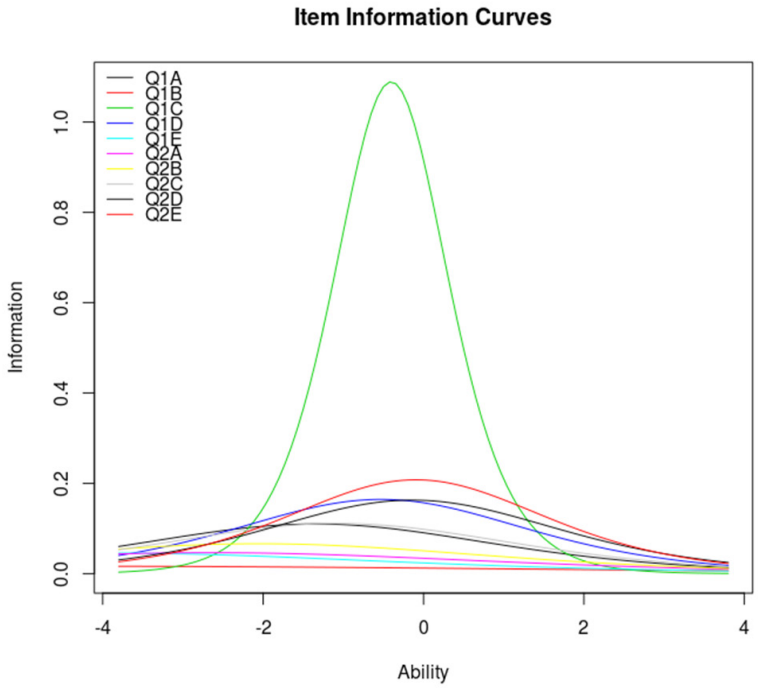

Figure 2: Item information curves of the test.

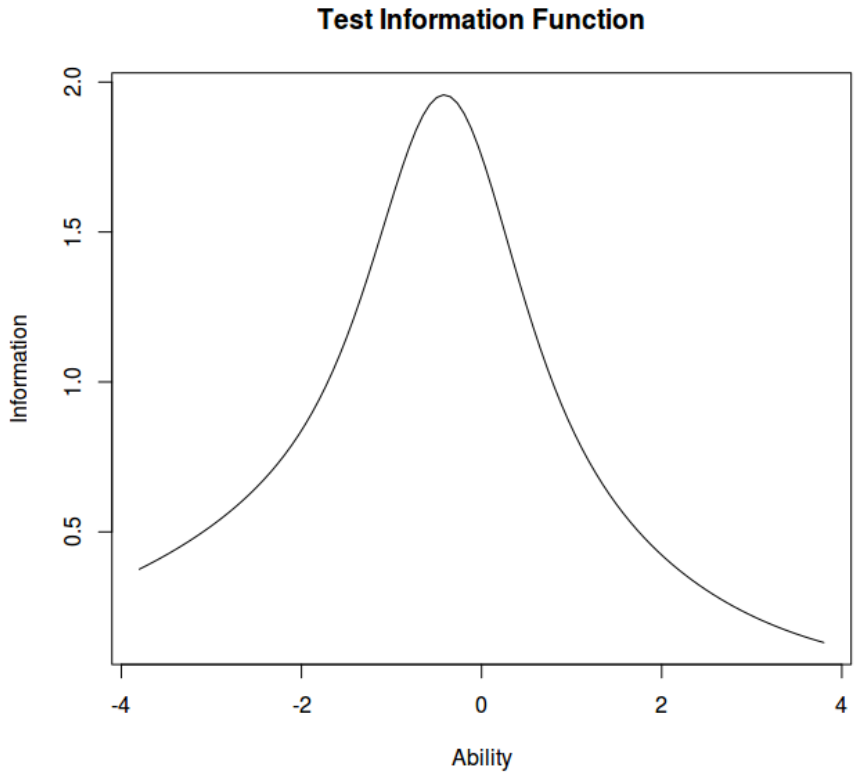

Figure 3: Test information function of the test.

```
Total Information = 7.46

Information in (-3, 3) = 5.87 (78.7%)

Based on all the items
```

This output indicates the amount of information given by the test is 78.7% for –3 to +3 ability range. You may change "–3, 3" in the command to other ability range to assess the amount of information tapped by the test. The –3 to +3 range was chosen in this example because it is the range considered as the typical range (14).

Using **irtoys** package, plot the test response function, which is also known as test characteristic curve,

```
plot(trf(est(data.mtf, model = "2PL", engine = "ltm")))
```

The interpretation for Figure 4 is straight forward. It is interpreted in similar way to that of the ICC, except that the *y*-axis is now replaced with "Expected score". You may say that the expected score for those with ability level of 4 is close to 10 (i.e. full mark). The expected score in this figure is also known as "true score" (14). On close inspection of the figure again, you will notice that the expected score for those with ability level of –4 is 2, which means that it is quite easy to obtain at least 2 marks on this test. Thus, it is worth considering the removal easy items (e.g. Q1b), while keeping other items with higher difficulty values.

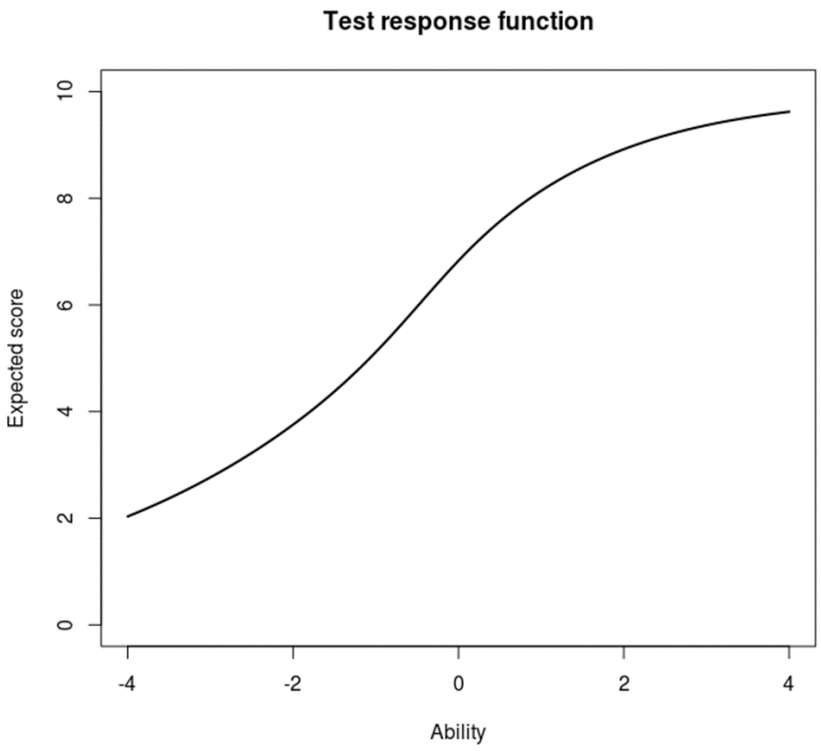

Figure 4: Test response function of the test.

Step 5. In this step, perform additional tests on the model fit, unidimensionality assumption and reliability.

The model fit assessment is done to ensure the values obtained (the difficulty and discrimination) from the IRT model accurately reflect the data that we have at hand. In **ltm**, the assessment can be done by item fit and fit for two-way margins.

At item level, we take a look at how accurately the ICC of an item (Figure 1; predicted proportions/probabilities of correct response versus ability levels) match the observed (i.e. from our data) proportions of correct response by ability groups of that item (i.e. as if we were to sort the students in ascending order by their test scores and divide them into 10 achievement groups). This is called item fit.

For the purpose of item fit assessment, perform chi-square tests to assess whether the observed proportions for an item correspond to the predicted proportions by 2-PL IRT model (i.e. along the ICC) (14).

```
item.fit(irt.mtf)
```

```
Item-Fit Statistics and P-values

Alternative: Items do not fit the model

Ability Categories: 10
```

	X^2	Pr(>X^2)
Q1A	14.3105	0.074
Q1B	24.0446	0.0023
Q1C	33.1838	0.0001
Q1D	14.9949	0.0592
Q1E	12.8225	0.1181
Q2A	16.4653	0.0362
Q2B	19.8424	0.0109
Q2C	15.9399	0.0432
Q2D	15.8849	0.0441
Q2E	15.2307	0.0548

To say that an item fit the model, we are looking for a *P*-value ≥ 0.05 (using the commonly used significance level of 0.05) based on chi-square goodness-of-fit test. Please note the degree of freedom for the chi-square (df) = number of ability categories –2. In our example, df = 10 (default in **ltm**) –2 = 8. Based on the output, six items, Q1b, Q1c, Q2a, Q2b, Q2c and Q2d do not fit the model well (*P*-values < 0.05). Although these items do not fit the model that well, please also consider keeping the items if the items' difficulty and discrimination values fall into acceptable ranges based on Table 1. For example, we may consider keeping Q2c because it has difficulty = –1.06, discrimination = 0.67 and *P*-value = 0.043 (slightly below the cutoff of 0.05).

Next, for assessment of goodness-of-fit for two-way margins (as we will explain below), perform margins() to assess chi-squared residuals as follows:

```
describe(data.mtf)[1]
```

	vars
Q1A	1
Q1B	2
Q1C	3
Q1D	4
Q1E	5
Q2A	6
Q2B	7
Q2C	8
Q2D	9
Q2E	10

This output is obtained to match the item names with the item numbers, so that it is easier to read the subsequent output.

```
margins(irt.mtf)
```

```
Fit on the Two-Way Margins
```

Response: (0,0)

	Item *i*	Item *j*	Obs	Exp	(O-E)^2/E
1	5	6	14	7.24	6.32***
2	2	5	13	7.11	4.87***
3	4	7	15	20.18	1.33

Response: (1,0)

	Item *i*	Item *j*	Obs	Exp	(O-E)^2/E
1	2	5	13	18.89	1.84
2	5	6	26	32.78	1.40
3	4	7	28	22.83	1.17

Response: (0,1)

	Item *i*	Item *j*	Obs	Exp	(O-E)^2/E
1	5	6	12	18.77	2.44
2	2	5	28	33.89	1.02
3	7	10	23	19.30	0.71

Response: (1,1)

	Item *i*	Item *j*	Obs	Exp	(O-E)^2/E
1	5	6	108	101.22	0.45
2	4	7	67	72.09	0.36
3	2	5	106	100.10	0.35

Note: * denotes a chi-squared residual greater than 4.**

We believe that the readers are familiar with chi-square test, which compares the observed cell counts with the expected cell counts to make conclusion on the association between two categorical variables. The statistics (Obs = observed counts, Exp = expected counts by the IRT model) that we read here in the output are basically 2 by 2 tabulation of Item *i* vs Item *j*, thus the term *two-way margins*. For example if we tabulate Q1e vs Q2a,

```
table(data.mtf[,5], data.mtf[,6])
```

	0	1
0	14	12
1	26	108

Note the same count under the respective response combination in the output of **margins(irt.mtf)**. We then check the discrepancies (residuals) between the observed and expected counts, in the words the goodness-of-fit between the counts.

As a rule of thumb, chi-squared residual > 4 indicates poor fit on two-way margin (12). In the output, Q1e-Q2a and Q1b-Q1e item pairs show poor fit.

As for the unidimensionality, IRT makes a strong assumption of unidimensionality (8). To the readers who are familiar with factor analysis, it means there should be only one factor to explain the relationship between the items. In simpler words, it means that the items can be suitably summed up as a total score.

For the purpose of testing unidimensionality, perform modified parallel analysis (25),

```
unidimTest(irt.mtf)
```

```
Unidimensionality Check using Modified Parallel Analysis

Alternative hypothesis: the second eigenvalue of the observed data is substantially larger than the second eigenvalue of data under the assumed IRT model

Second eigenvalue in the observed data: 1.1476

Average of second eigenvalues in Monte Carlo samples: 0.9631

Monte Carlo samples: 100

p-value: 0.1089
```

The way we interpret the *P*-value here is almost similar to that of item fit as discussed before, as we are looking for a *P*-value ≥ 0.05 to say the data are unidimensional. The *P*-value here is more than 0.05, thus indicates that the unidimensionality assumption is met, thus unidimensional 2-PL IRT can be applied to the data.

COMPARISON BETWEEN IA AND IRT

The comparison of IA and IRT analysis are summarised in Table 2. Note the discrepancy between the analyses for Q1b in term of the difficulty and discrimination; item Q1b is a poor item based IRT analysis, but a good item based on IA. Also note the similarity for Q1e; both analyses show that Q1e is an easy item. For the rest of the items, IRT gives more refined cutoff values for the discriminative ability of the items, while IA only shows that all the items are very good at discriminating high and low scorers. It is also easy to decide on the difficulty of any items based on IRT analysis results because the value ranges from negative to positive that represent the intuitive progression from easy to difficult.

Table 2: Comparison of results by item analysis and 2-parameter logistic model of item response theory

MTF questions	Difficulty		Discrimination	
	IA	2-PL IRT	IA[a]	2-PL IRT
Q1a	0.69	–1.35	0.64	0.66
Q1b	0.74	**–4.20**	0.48	**0.26**
Q1c	0.63	–0.40	0.82	2.09
Q1d	0.59	–0.53	0.70	0.81
Q1e	**0.84**	**–3.97**	0.48	0.43
Q2a	0.75	–2.65	0.52	0.43
Q2b	0.73	–2.05	0.55	0.52
Q2c	0.66	–1.06	0.66	0.67
Q2d	0.53	–0.14	0.73	0.81
Q2e	0.52	–0.09	0.77	0.91

Potentially problematic items (based on criteria in Table 1) are highlighted in bold.

IA = item analysis, 2-PL IRT = 2-parameter logistic item response theory model, MTF = multiple true-false. [a]Because the data set used in this article has a relatively small number of items, there are a number of ties after sorting the responses by the total scores. To obtain the discrimination index values as presented here, the responses have to be sorted in descending order by the total scores followed by the scores of a particular item before calculating the discrimination index of that item.

Table 3: 2-PL IRT parameter estimates and item fit statistics

Items	Difficulty (*b*)	Discrimination (*a*)	χ^2 (df = 8)	*P*-values
Q1a	–1.35	0.66	14.31	0.074
Q1b	–4.20	0.26	24.04	**0.002**
Q1c	–0.40	2.09	33.18	**< 0.001**
Q1d	–0.53	0.81	14.99	0.059
Q1e	–3.97	0.43	12.82	**0.118**
Q2a	–2.65	0.43	16.47	**0.036**
Q2b	–2.05	0.52	19.84	**0.011**
Q2c	–1.06	0.67	15.94	**0.043**
Q2d	–0.14	0.81	15.88	**0.044**
Q2e	–0.09	0.91	15.23	0.055

Items with *P*-values < 0.05 on item fit assessment are highlighted in bold. On assessment of fit for two-way margins, Q1e-Q2a and Q1b-Q1e item pairs showed poor fit. Modified parallel analysis supported unidimensionality.

2-PL IRT = 2-parameter logistic item response theory model, df = degree of freedom.

Based on our outputs from RStudio, we may report the 2-PL IRT analysis as displayed in Table 3. 2-PL IRT analysis shows that items Q1b and Q1e are relatively easy. Q1b has very low discrimination, and Q1e, Q2a and Q2b have low discrimination based on Baker (14); six items (Q1a, Q1b, Q1e, Q2a, Q2b and Q2c) are outside the "good" range as recommended by de Ayala (18). Six items (Q1b, Q1c, Q2a, Q2b, Q2c and Q2d) do not fit the 2-PL IRT model based the item-fit statistics. Modified parallel analysis supports unidimensionality assumption of the model. Based on these findings, researchers may select the best items to represent a test, while keeping in mind the difficulty, discrimination and item-fit. There could be a balance in the selection, for example, by keeping Q2c because it has acceptable difficulty and discrimination, although the item does not fit the model. The decision is relatively easy for Q1b (easy, very low discrimination, poor item fit), while it gets difficult for Q1c (average difficulty, very high/good discrimination, poor item fit).

CONCLUSION

In this article, we showed how to obtain the parameter estimates of difficulty and discrimination from the IRT analysis and how to interpret the results. We also compared the results of IA and 2-PL IRT analyses. 2-PL IRT analysis provides more information and refined cutoff values as compared to the commonly used IA. Although admittedly the analysis is slightly more complicated than IA and requires good understanding of the statistical analysis, compounded with the need to learn R software, it is our intention in writing this article to make it easier to the majority of the readers. It is hoped that medical educationists will seriously consider using 2-PL IRT analysis to evaluate test items.

NOTES

The version numbers of software and packages used in this article are: R version 3.3.2, RStudio version 1.0.136, **psych** version 1.6.12, **ltm** version 1.0.0, and **irtoys** version 0.2.0. The files described in this article can be downloaded from https://researchgate.net/profile/Wan_Nor_Arifin

REFERENCES

1. AERA. Standards for educational and pyschological testing. 2nd ed. Washington, US: American Educational Research Association; 2004.

2. Epstein RM. Assessment in medical education. The New England Journal of Medicine. 2007;356(4):387–96. https://doi.org/10.1056/NEJMra054784.

3. Newble D. Assessing clinical competence at the undergraduate level. Medical Education. 1992;26(6):504–11. https://doi.org/10.1111/j.1365-2923.1992.tb00213.x.

4. Taib F, Yusoff MSB. Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. Journal of Taibah University Medical Sciences. 2014;9(2):110–4. https://doi.org/10.1016/j.jtumed.2013.12.002.

5. van der Vleuten C. Validity of final examinations in undergraduate medical training. BMJ. 2000;321(7270):1217–9. https://doi.org/10.1136/bmj.321.7270.1217.

6. Miller G. The assessment of clinical skills/competence/performance. Academic Medicine. 1990;65(9):S63–7. https://doi.org/10.1097/00001888-199009000-00045.

7. Cook DA, Beckman T. Current concepts in validity and reliability for psychometric instrument: theory and application. The American Journal of Medicine. 2006;119(166):e7–e16. https://doi.org/10.1016/j.amjmed.2005.10.036.

8. Streiner D, Norman G, Cairney J. Health measurement scales: a practical guide to their development and use. 5th ed. Oxford: Oxford University Press; 2014. https://doi.org/10.1093/med/9780199685219.001.0001.

9. Dixon R. Evaluating and improving multiple choice papers: true-false questions in public health medicine. Medical Education. 1994;28(5):400–8. https://doi.org/10.1111/j.1365-2923.1994.tb02551.x.

10. Rahim AFA. What those number mean? 1st ed. Kubang Kerian: KKMED Publications; 2010.

11. Hassan S, Mohd Amin R, Mohd Amin Rebuan H, Aung MMT. Item analysis, reliability statistics and standard error of measurement to improve the quality and impact of multiple choice questions in undergraduate medical education in faculty of medicine at UNISZA. Malaysian Journal of Public Health Medicine. 2016;16(3):7–15.

12. Bartholomew DJ, Steele F, Moustaki I, Galbraith JI. 2nd ed. Analysis of multivariate social science data. Boca Raton, FL: CRC Press; 2008.

13. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Quality of Life Research. 2007;16(1):5. https://doi.org/10.1007/s11136-007-9198-0.

14. Baker FB. The basics of item response theory. 2nd ed. USA: ERIC Clearinghouse on Assessment and Evaluation; 2001. Available from: http://ericae. net/irt/baker

15. Linn RL, Gronlund NE. Measurement and assessment in teaching. 7th ed. New Jersey: Prentice-Hall; 1995.

16. Sim S, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. Annals Academy of Medicine Singapore. 2006;35(2):67–71.

17. Ebel RL, Frisbie DA. Essentials of educational measurement. 5th ed. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.; 1991.

18. de Ayala RJ. The theory and practice of item response theory. New York: The Guilford Press; 2009.

19. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available from: https://www.r-project.org/

20. RStudio Team. RStudio: integrated development for R. Boston, MA: RStudio, Inc; 2016. Available from: http://www.rstudio.com/

21. Revelle W. Psych: procedures for personality and psychological research. Evanston, Illinois, USA: Northwestern University; 2016. Available from: https://CRAN.R-project.org/package=psych

22. Rizopoulos D. ltm: an R package for latent variable modelling and item response theory analyses. Journal of Statistical Software. 2006;7(5):1–25. Available from: http://www.jstatsoft.org/v17/i05/

23. Partchev I. Irtoys: a collection of functions related to item response theory (IRT). [R package]. 2016. Available from: https://CRAN.R-project.org/package=irtoys

24. Chalmers RP. Mirt: a multidimensional item response theory package for the R environment. Journal of Statistical Software. 2012;48(6):1–29. Available from: http://www.jstatsoft.org/v48/i06/. https://doi.org/10.18637/jss.v048.i06.

25. Drasgow F, Lissak RI. Modified parallel analysis: a procedure for examining the latent dimensionality of dichotomously scored item responses. Journal of Applied Psychology. 1983;68(3):363–73. https://doi.org/10.1037/0021-9010.68.3.363.