Probabilistic Approach in Examining Quality of Survey Response Data in Statistical Education Research

Zamalia Mahmud, Nor Azila C. Musa, Nor Azura M. Ghani, Rosli A. Rahim

***Abstract*—** Obtaining and maintaining quality data in a survey investigation has becoming a continuing concern among the statistical education researchers. Rasch probabilistic measurement model had been used to identify inappropriate survey items in many other instruments but it has not been extensively used in many survey investigations involving statistics education research. This study had employed Rasch dichotomous and rating models to examine the quality of survey response data, namely on the students' attitude towards and their competency in learning elementary statistics. Students' attitude was measured by the 24 items of 5-point Likert scale while statistical competency was measured by their ability to answer correctly or incorrectly based on three statistical elementary topics. This study used secondary data which was formerly gathered from 139 secondary school students over several occasions, at two different points of time (prior to statistics class teaching and end of class teaching). The outcome was investigated based on both item and person misfit response strings and PIDM map. Rasch analysis had shown that quality of items and persons can be enhanced with proper validation techniques namely, through identification of fit statistics on the items and misfit response strings. Generally, Rasch probabilistic model is able to diagnose the unusual response patterns which otherwise could not be detected using the general deterministic model.

***Keywords*—** Quality response data**,** Rasch probabilistic models, Logit, Differential Item Functioning, Attitude towards statistics, Statistical competency

I. Introduction

Obtaining and maintaining quality data in a survey investigation has becoming a continuing concern among the statistical education researchers. Rasch probabilistic model had been used to identify inappropriate survey items in many other instruments but it has not been extensively used in many survey investigations involving statistics education research. Rasch probabilistic model is known to examine and validate psychometric proportion of measurement instrument and test forms [1]-[4]. According to [1]-[2], Rasch model is able to produce a reliable repeatable measurement instrument and focuses on constructing the measurement instrument with accuracy. As Rasch model is not commonly used in statistical education research, this study shall attempt to explore the quality of response data using reliability and construct validity. The emphasis shall be on examining the quality of survey response relating to students' attitude towards learning and competency in elementary statistics. This shall be accomplished using the appropriate Rasch measurement tools such as Person-Item Distribution Map (PIDM), Scalogram, Fit Statistics, Misfit Response Strings and Characteristic Curves.

This study is expected to expose the statistical education researchers with other perspectives of validity measurement tools and reveal the advantages of using Rasch measurement model in an effort to maintain the quality of survey response data.

II. Methods

This study use secondary data obtained from the Survey of Attitude Towards Learning and Competency in Elementary Statistics among 139 upper secondary school students at SMK Bandar Baru Sungai Buloh [5]. Data was collected over a period of 4 weeks using structured questionnaires and test forms to assess students' attitude and competency in elementary statistics. The data were classified into several categories, which include type of class, gender, and race.

The study had administered two instruments namely, (1) Attitude towards learning statistics and (2) Competency test forms. Students' attitude towards learning statistics which comprises of 16 items were measured on a 5-point Likert-scale ranging from 1=Strongly Disagree, 3=Neutral to 5=Strongly Agree across three demographic variables which are gender, race, and type of classes. The competency test form A comprises of 8 questions relating to pictogram, bar graph, line graph, and pie chart. Test form B comprises of 6 questions relating to class interval, mode and mean, and histogram and test form C comprises of 6 questions relating to cumulative frequency distribution or ogive. These instruments were

Manuscript was submitted for review on 26 October 2011. This work was supported by the Ministry of Higher Education of Malaysia under the Fundamental Research Grant Scheme.

Dr Zamalia Mahmud is an Associate Professor at the Center of Studies for Statistics and Decision Science, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, MALAYSIA (phone: +603-55435367; fax: +603-55435301; e-mail: zamalia@tmsk.uitm.edu.my)

Nor Azila Che Musa is a lecturer at the University of Malaysia Pahang (e-mail: nor_azila772@yahoo.com)

Dr. Nor Azura Md Ghani is a senior lecturer at the Center of Studies for Statistics and Decision Science, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, MALAYSIA (e-mail: azura@tmsk.uitm.edu.my).

Rosli A. Rahim is an Associate Professor at the Faculty of Business Administration, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, MALAYSIA (e-mail: rosliar@salam.uitm.edu.my).

administered at two different points, that is, at the beginning (pre) and end of class (post).

The quality of data were analyzed using Winsteps 3.6.3. and results are presented in several sections namely, data exploration, validation and calibration of items and person responses. The analysis include producing item and person fit statistics in the form of mean square values (MNSQ) and standardized z-scores for infit and outfit. The analysis was carried out in order to calibrate between items difficulty and person ability on students' attitudes toward learning and competency in statistics using PIDM and ICCs for the endorsement of items and person responses.

Students' attitude towards learning statistics were measured by the Rasch Rating Scale Model as follows. The general probability of person *n* scoring *x* on item *i* given β_n and δ_i at different threshold level F_j is given by:

$$P_{ni}\{X = x \mid \beta_{n,}\delta_i, F_j\} = \frac{\exp\sum_{k=0}^{x}(\beta_n - \delta_i - F_j)}{\sum_{h=0}^{m_i}\exp\sum_{k=0}^{h}(\beta_n - \delta_i - F_j)ni} \qquad (1)$$

β_n defined as person ability, δ_i as item difficulty.

On the other hand, Rasch Dichotomous Model is used to measure students' competency in learning statistics. This instrument provide items with two alternative answers namely, "Yes" or "No". The probability of person and item is defined mathematically as follows:

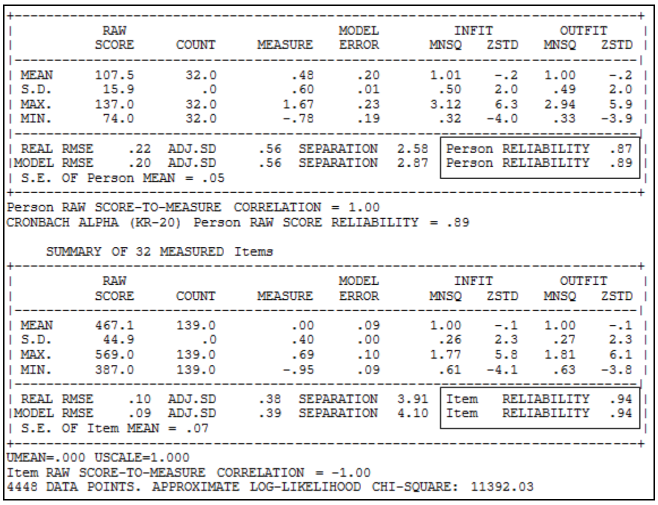
$$P_{ni}\{x = 1 \mid \beta_{n,}\delta_i\} = \frac{\exp(\beta_n - \delta_i)}{[1 + \exp(\beta_n - \delta_i)]} \qquad (2)$$

where $P_{ni}\{x = 1 \mid \beta_{n,}\delta_i\}$ is the probability of person n on item i scoring as correct (x=1) response rather than an incorrect (x=0) response, given person ability β_n and item difficulty δ_i.

III. ANALYSIS AND RESULTS

The summary statistics was obtained for person and items as shown in Table 1.

TABLE I
SUMMARY STATISTICS FOR PRE AND POST-ATTITUDE



```
+-----------------------------------------------------------------------------+
|          RAW                              MODEL         INFIT        OUTFIT  |
|         SCORE      COUNT      MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     107.5       32.0         .48      .20      1.01    -.2   1.00    -.2 |
| S.D.      15.9         .0         .60      .01       .50    2.0    .49    2.0 |
| MAX.     137.0       32.0        1.67      .23      3.12    6.3   2.94    5.9 |
| MIN.      74.0       32.0        -.78      .19       .32   -4.0    .33   -3.9 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .22  ADJ.SD     .56  SEPARATION  2.58  Person RELIABILITY  .87 |
|MODEL RMSE    .20  ADJ.SD     .56  SEPARATION  2.87  Person RELIABILITY  .89 |
| S.E. OF Person MEAN = .05                                                    |
+-----------------------------------------------------------------------------+
Person RAW SCORE-TO-MEASURE  CORRELATION = 1.00
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .89

     SUMMARY OF 32 MEASURED Items
+-----------------------------------------------------------------------------+
|          RAW                              MODEL         INFIT        OUTFIT  |
|         SCORE      COUNT      MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     467.1      139.0         .00      .09      1.00    -.1   1.00    -.1 |
| S.D.      44.9         .0         .40      .00       .26    2.3    .27    2.3 |
| MAX.     569.0      139.0         .69      .10      1.77    5.8   1.81    6.1 |
| MIN.     387.0      139.0        -.95      .09       .61   -4.1    .63   -3.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .10  ADJ.SD     .38  SEPARATION  3.91  Item   RELIABILITY  .94 |
|MODEL RMSE    .09  ADJ.SD     .39  SEPARATION  4.10  Item   RELIABILITY  .94 |
| S.E. OF Item MEAN = .07                                                      |
+-----------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000
Item RAW SCORE-TO-MEASURE  CORRELATION = -1.00
4448 DATA POINTS. APPROXIMATE LOG-LIKELIHOOD CHI-SQUARE: 11392.03
```

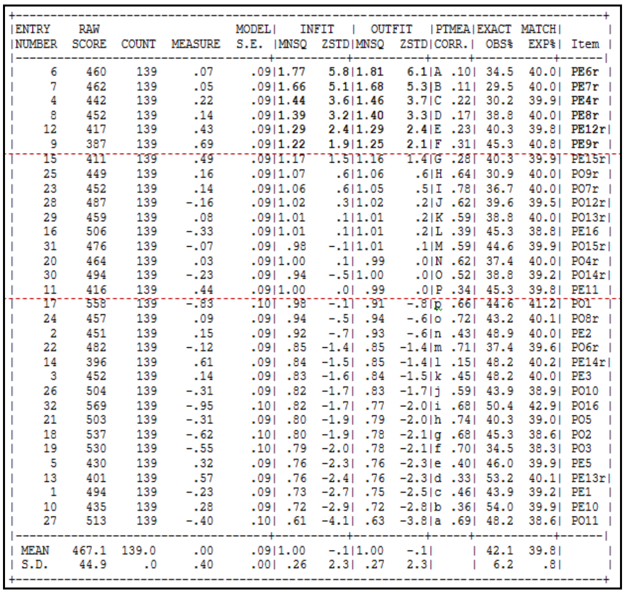
Table 1 presents a moderately high person reliability index (0.87) and a high item reliability index (0.94). These are considered good index for items and persons. The mean infit and outfit for person and item mean squares are expected to be 1.00, and for this data, they are all close to 1.00. The mean standardized infit and outfit are expected to be near 0.0.

However, the table shows that the z-scores for infit and outfit are -0.2 for persons and -0.1 for items, respectively. This indicates that the items are overfit. It also represents that the data fit the model somewhat better than would be expected which could be due to some redundant items. The data shows an overall acceptable fit as the value for standardized infit standard deviation for person is 0.50 while for item is 0.26.

The separation index for person is 2.58, a moderately good spread of items and person along a continuum. For item separation index, it shows a large index of 3.91 which indicates that a broader continuum for items than for person, and broader range of item difficulties.

Table of misfit responses was examined to identify misfit responses in the data set as shown in Table 2.

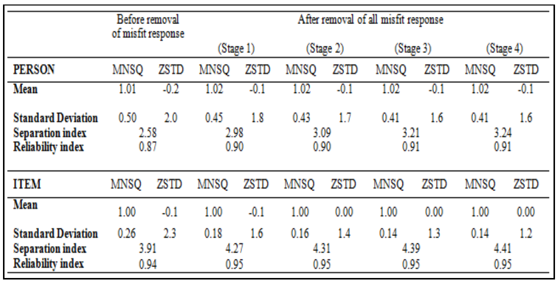
TABLE 2
FIT STATISTICS FOR PRE- AND POST-ATTITUDE ITEMS PRIOR TO REMOVAL OF MISFIT RESPONSES



```
+------------------------------------------------------------------------------------+
|ENTRY    RAW                          MODEL|    INFIT   |   OUTFIT   |PTMEA|EXACT  MATCH|      |
|NUMBER  SCORE   COUNT   MEASURE  S.E. |MNSQ   ZSTD|MNSQ   ZSTD|CORR.| OBS%   EXP%| Item |
|------------------------------------+----------+----------+-----+-----------+------|
|     6    460     139      .07      .09|1.77    5.8|1.81    6.1|A .10| 34.5   40.0| PE6r |
|     7    462     139      .05      .09|1.66    5.1|1.68    5.3|B .11| 29.5   40.0| PE7r |
|     4    442     139      .22      .09|1.44    3.6|1.46    3.7|C .22| 30.2   39.9| PE4r |
|     8    452     139      .14      .09|1.39    3.2|1.40    3.3|D .17| 38.8   40.0| PE8r |
|    12    417     139      .43      .09|1.29    2.4|1.29    2.4|E .23| 40.3   39.8| PE12r|
|     9    387     139      .69      .09|1.22    1.9|1.25    2.1|F .31| 45.3   40.8| PE9r |
|    15    411     139      .49      .09|1.17    1.5|1.16    1.4|G .28| 40.3   39.9| PE15r|
|    25    449     139      .16      .09|1.07     .6|1.06     .6|H .64| 30.9   40.0| PO9r |
|    23    452     139      .14      .09|1.06     .6|1.05     .5|I .78| 36.7   40.0| PO7r |
|    28    487     139     -.16      .09|1.02     .3|1.02     .2|J .62| 39.6   39.5| PO12r|
|    29    459     139      .08      .09|1.01     .1|1.01     .2|K .59| 38.8   40.0| PO13r|
|    16    506     139     -.33      .09|1.01     .1|1.01     .2|L .39| 45.3   38.8| PE16 |
|    31    476     139     -.07      .09| .98    -.1|1.01     .1|M .59| 44.6   39.9| PO15r|
|    20    464     139      .03      .09|1.00     .1| .99     .0|N .62| 37.4   40.0| PO4r |
|    30    494     139     -.23      .09| .94    -.5|1.00     .0|O .52| 38.8   39.2| PO14r|
|    11    416     139      .44      .09|1.00     .0| .99     .0|P .34| 45.3   39.8| PE11 |
|    17    558     139     -.83      .10| .98    -.1| .91    -.8|p .66| 44.6   41.2| PO1  |
|    24    457     139      .09      .09| .94    -.5| .94    -.6|o .72| 43.2   40.1| PO8r |
|     2    451     139      .15      .09| .92    -.7| .93    -.6|n .43| 48.9   40.0| PE2  |
|    22    482     139     -.12      .09| .85   -1.4| .85   -1.4|m .71| 37.4   39.6| PO6r |
|    14    396     139      .61      .09| .84   -1.5| .85   -1.4|l .15| 48.2   40.2| PE14r|
|     3    452     139      .14      .09| .83   -1.6| .84   -1.5|k .45| 48.2   40.0| PE3  |
|    26    504     139     -.31      .09| .82   -1.7| .83   -1.7|j .59| 43.9   38.9| PO10 |
|    32    569     139     -.95      .10| .82   -1.7| .77   -2.0|i .68| 50.4   42.9| PO16 |
|    21    503     139     -.31      .09| .80   -1.9| .79   -2.0|h .74| 40.3   39.0| PO5  |
|    18    537     139     -.62      .10| .80   -1.9| .78   -2.1|g .68| 45.3   38.6| PO2  |
|    19    530     139     -.55      .10| .79   -2.0| .78   -2.1|f .70| 34.5   38.3| PO3  |
|     5    430     139      .32      .09| .76   -2.3| .76   -2.3|e .40| 46.0   39.9| PE5  |
|    13    401     139      .57      .09| .76   -2.4| .76   -2.3|d .33| 53.2   40.1| PE13r|
|     1    494     139     -.23      .09| .73   -2.7| .75   -2.5|c .46| 43.9   39.2| PE1  |
|    10    435     139      .28      .09| .72   -2.9| .72   -2.8|b .36| 54.0   39.9| PE10 |
|    27    513     139     -.40      .10| .61   -4.1| .63   -3.8|a .69| 48.2   38.6| PO11 |
|------------------------------------+----------+----------+-----+-----------+------|
| MEAN   467.1   139.0      .00      .09|1.00    -.1|1.00    -.1|     | 42.1   39.8|      |
| S.D.    44.9      .0      .40      .00| .26    2.3| .27    2.3|     |  6.2     .8|      |
+------------------------------------------------------------------------------------+
```

From Table 2, there are four items (4, 6, 7, 8) which are underfit as the MNSQ values fall between 1.4 and 2.0. These are items in which students gave unusual or inappropriate responses and hence considered as misfit. These misfit responses were removed from the data set and subjected to another reliability and validity analysis. This process continue until all misfit responses were removed. In the process of identifying unusual responses of students' attitude towards statistics, the removal process was done in four stages until there was no misfit responses. The summary statistics after each removal is summarized in Table 3.

TABLE 3
SUMMARY STATISTICS FOR ATTITUDE CONSTRUCTS



	Before removal of misfit response		After removal of all misfit response							
			(Stage 1)		(Stage 2)		(Stage 3)		(Stage 4)	
PERSON	MNSQ	ZSTD	MNSQ	ZSTD	MNSQ	ZSTD	MNSQ	ZSTD	MNSQ	ZSTD
Mean	1.01	-0.2	1.02	-0.1	1.02	-0.1	1.02	-0.1	1.02	-0.1
Standard Deviation	0.50	2.0	0.45	1.8	0.43	1.7	0.41	1.6	0.41	1.6
Separation index	2.58		2.98		3.09		3.21		3.24	
Reliability index	0.87		0.90		0.90		0.91		0.91	
ITEM	MNSQ	ZSTD	MNSQ	ZSTD	MNSQ	ZSTD	MNSQ	ZSTD	MNSQ	ZSTD
Mean	1.00	-0.1	1.00	-0.1	1.00	0.00	1.00	0.00	1.00	0.00
Standard Deviation	0.26	2.3	0.18	1.6	0.16	1.4	0.14	1.3	0.14	1.2
Separation index	3.91		4.27		4.31		4.39		4.41	
Reliability index	0.94		0.95		0.95		0.95		0.95	

Table 3 shows that the person mean and standard deviation values give a more acceptable fit at value around 1.0 while standard deviation decreases as more misfit responses were removed. Separation index shows an increasing value from 2.58 to 3.24. The indices indicate that person ability level can be categorized into 2 to 3 level spread of person positions. The initial person reliability index was estimated at 0.87, and it increases to 0.91 when all misfit responses were removed. For item summary statistics of items, there is an improvement in the fit statistics as compared to before removal of misfit responses. Overall item reliability index is estimated at 0.94 but slightly increases to 0.95 as more misfit responses were removed. Items spread also shows more variability as separation index increases from 3.91 to 4.41.

The distribution of students responses were tabulated as in Fig. 1. Generally, there are about 36.7% of the students who indicate their agreement with all the items, where more than half of them are female students (72.5%). Based on race and type of classes, Malay students are more likely to agree with all the items compared to non-Malay students, while about 84.3% students from the science class agree with all the items in the questionnaire.

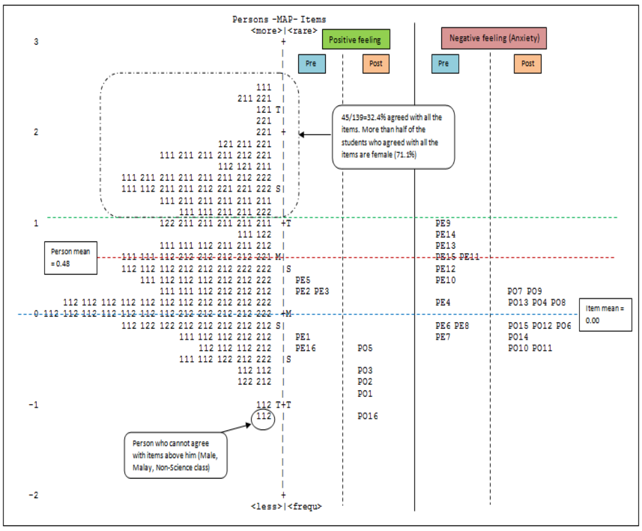


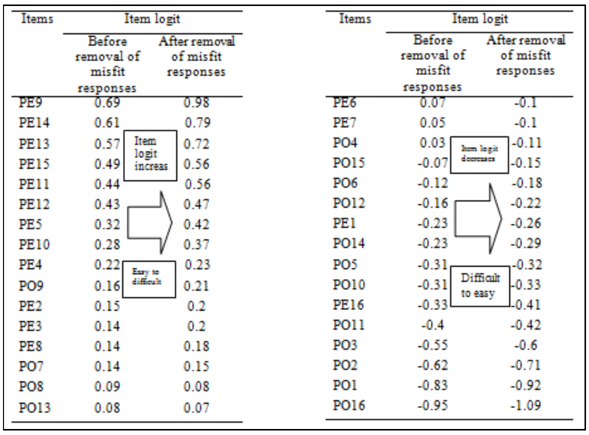
Fig. 1 PIDM for Pre and Post-Attitude towards learning elementary statistics (After Removal of misfit responses)

Pre-attitude items are considered as difficult questions to endorse or to agree with because more than half of the items are located above the item mean (0.00). The most difficult is Item 9, that is "I will tend to make lots of errors in statistics calculation", which is the factor for negative feeling towards statistics. Most students thought this item is difficult to agree with because they lack the confidence in the computational part. Items 13 and 14 are also difficult to endorse, and fall under the negative feeling category. Post-attitude items 16 and 1 are considered as easy to endorse. Items 16 and 1 ask about "Statistics is relevant in my life" and "Learning statistics is exciting", respectively. Students find these items most easy to agree with, as it measures students' positive feeling towards statistics.

Post-attitude items are mostly located at the bottom of the map compared to the pre attitude items. There are 11 out of 16 items located below the item mean logit 0.00. The easiest items to endorse are items which measure students' positive feeling towards statistics. These items are PO1, PO2, PO3, and PO16, while the most difficult items to endorse are PO7, PO9, PO13, PO4, and PO8. These items measure students' negative feeling towards statistics. Overall, most post-attitude items which are easy to endorse with are items PO1, PO2, PO3, PO5, PO6, PO10, PO11, PO12, PO14, PO15, and PO16, in which all fall below the item mean logit 0.00. Students are observed to be more able to endorse post-attitude items than the pre-attitude items. This indicates that students have more positive feeling towards statistics after attending the statistics lessons.

Items logit before removal of misfit responses and after removal of misfit responses were observed. Table 4 revealed the overall increment of logit values for all items after removal of misfit responses. After removal of the misfit responses all items logit change from either easy to difficult to endorse or from difficult to easy to endorse the items. For example, item PE9 has item logit 0.69 before removal of misfit items and increase to 0.98 logit after removal of misfit response. As the logit values increases, the difficulty of the items also increases. In this case, item PE9 changes from easy to difficult to endorse by students. For item PO16, the item logit value decrease from -0.95 before removal of misfit responses to -1.09 after removal of misfit responses. This suggests that the item changes from difficult to easy to endorse by students.

TABLE 4
ATTITUDE-ITEM LOGIT BEFORE AND AFTER REMOVAL OF MISFIT RESPONSES



Items	Item logit		Items	Item logit	
	Before removal of misfit responses	After removal of misfit responses		Before removal of misfit responses	After removal of misfit responses
PE9	0.69	0.98	PE6	0.07	-0.1
PE14	0.61	0.79	PE7	0.05	-0.1
PE13	0.57	0.72	PO4	0.03	-0.11
PE15	0.49	0.56	PO15	-0.07	-0.15
PE11	0.44	0.56	PO6	-0.12	-0.18
PE12	0.43	0.47	PO12	-0.16	-0.22
PE5	0.32	0.42	PE1	-0.23	-0.26
PE10	0.28	0.37	PO14	-0.23	-0.29
PE4	0.22	0.23	PO5	-0.31	-0.32
PO9	0.16	0.21	PO10	-0.31	-0.33
PE2	0.15	0.2	PE16	-0.33	-0.41
PE3	0.14	0.2	PO11	-0.4	-0.42
PE8	0.14	0.18	PO3	-0.55	-0.6
PO7	0.14	0.15	PO2	-0.62	-0.71
PO8	0.09	0.08	PO1	-0.83	-0.92
PO13	0.08	0.07	PO16	-0.95	-1.09

Further analysis was carried out to look at the endorsement of the items and responses based on the Item Characteristics Curve (ICC). The selected expected and empirical ICC are chosen for the easiest and most difficult items in measuring students' attitude towards statistics. The pre-attitude items which are considered as most difficult is item 9 while post-attitude item 16 is considered as easiest item.

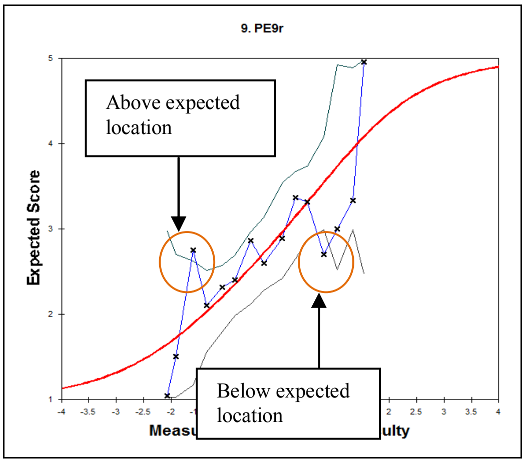


Fig. 2 The expected empirical ICC for Item 9 (Pre-attitude)

Fig. 2 shows that this item asks students about their tendency of making errors in statistics calculation. It can be seen that some expected score of responses does not lie along the sigmoid curve. There are some scores which scattered far away from the curve and located outside the range of the confidence interval (indicated by grey line). It was found that more able student tend to score highly beyond the 95% confidence interval while less able students tend to score below the lower confidence interval. There are misfit data or better known as unusual response. This may be due to guessing of answers or some other unexplained reasons. On the other hand, most students have 0.5 probability of answering this item correctly. For the easiest item, post-attitude item 16 as illustrated in Fig. 4 shows that this item is easy to endorse as the distribution of the responses are located within the confidence limit and also fall along the sigmoid curve. All the points are also located well above zero logit and along the upper half of the curve.

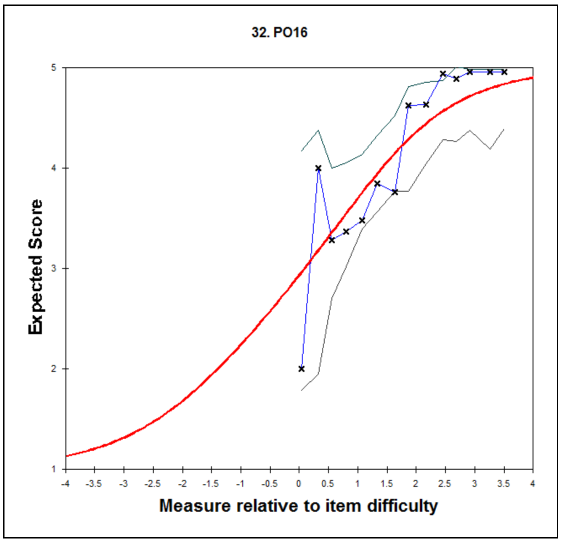


Fig. 4 The expected empirical ICC for Item 16 (Post-attitude)

The analysis continue by checking for existence of Gender Differential Item Functioning (GDIF) in the attitude survey items. Gender Differential Item Functioning (GDIF) Contrast index was used to show the difference of gap confirmation level for each item when comparing between male and female students. [6], [7], and [8] suggest that significant individual t test less than 0.5 was considered as unimportant or DIF is negligible.

The analysis demonstrated that 7 items from 40 items in the attitude survey items show the significance of GDIF in t value of greater than 2.0 logit and p value < α (0.05) as shown in Table 5. These items are pre-attitude items PE4, PE9, PE10, and post-attitude items PO2, PO3, PO5, and PO6. This indicates that these items have significant differences between male and female students.

TABLE 5
DIF BASED ON GENDER-ATTITUDE TOWARD STATISTICS

DIF Measure Group 1	DIF Measure Group 2	DIF Contrast	t	probability	Item
-0.12	0.41	0.53	±2.58	0.0112*	PE4
0.74	1.16	0.42	±2.02	0.0457*	PE9
0.11	0.58	0.47	±2.34	0.0208*	PE10
-0.4	-1.02	-0.62	±2.85	0.0051*	PO2
-0.34	-0.86	-0.52	±2.42	0.0167*	PO3
0.02	-0.62	-0.64	±3.07	0.0026*	PO5
0.25	-0.4	-0.65	±3.12	0.0022*	PO6

*p-value < α=0.05

The differences of items between male and female can be identified by examining the DIF chart. Fig. 5 shows DIF chart based on gender differential. It shows that items PE9 and PE10 are easier to endorse by female students as the DIF contrast indices is 0.42 and 0.47, respectively. Basically, most of pre-attitude items (except for item PE1 and PE15) seem to be easily endorsed by female students as the DIF contrast indices are positive values while male students were observed to endorse post-attitude items PO4, PO5, PO6, PO7, and PO16 easier than female students.

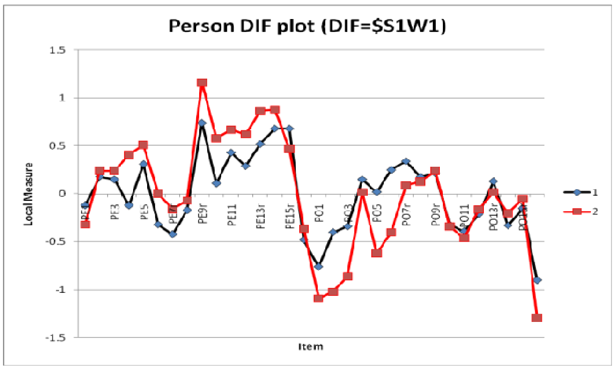
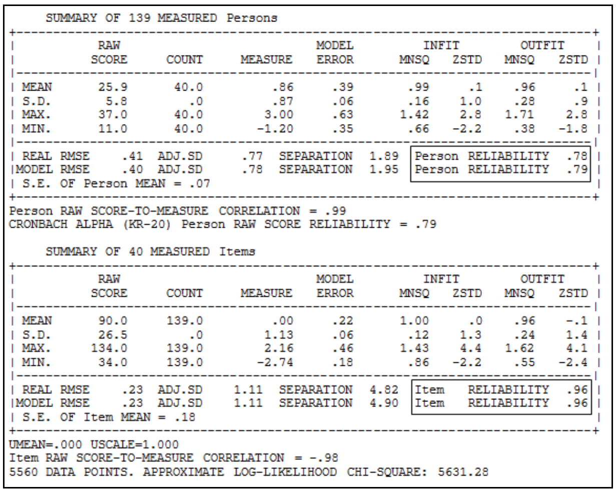


Fig. 5 DIF chart for Gender Differential Item Functioning (GDIF)-Attitude Items

The second part of the data was administered for students' competency in statistics. The result of summary statistics obtained is shown in Table 6.

TABLE 6
SUMMARY STATISTICS FOR PRE- AND POST COMPETENCY



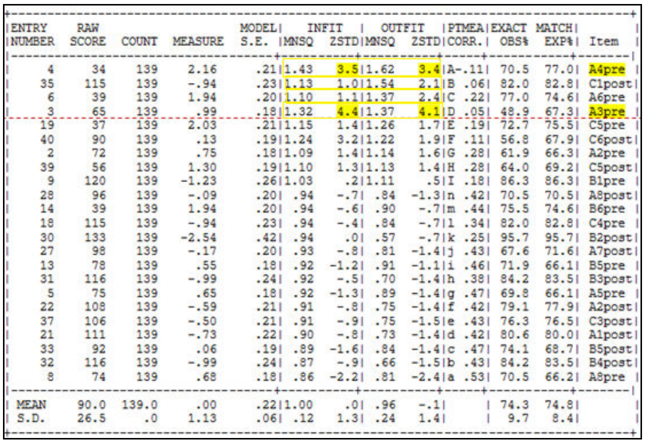
```
    SUMMARY OF 139 MEASURED Persons
+-----------------------------------------------------------------------------+
|          RAW                          MODEL         INFIT         OUTFIT    |
|         SCORE      COUNT    MEASURE   ERROR       MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      25.9       40.0        .86     .39        .99     .1    .96     .1 |
| S.D.       5.8         .0        .87     .06        .16    1.0    .28     .9 |
| MAX.      37.0       40.0       3.00     .63       1.42    2.8   1.71    2.8 |
| MIN.      11.0       40.0      -1.20     .35        .66   -2.2    .38   -1.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .41  ADJ.SD     .77  SEPARATION  1.89  Person RELIABILITY  .78 |
|MODEL RMSE    .40  ADJ.SD     .78  SEPARATION  1.95  Person RELIABILITY  .79 |
| S.E. OF Person MEAN = .07                                                   |
+-----------------------------------------------------------------------------+
Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .79

    SUMMARY OF 40 MEASURED Items
+-----------------------------------------------------------------------------+
|          RAW                          MODEL         INFIT         OUTFIT    |
|         SCORE      COUNT    MEASURE   ERROR       MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      90.0      139.0        .00     .22       1.00     .0    .96    -.1 |
| S.D.      26.5         .0       1.13     .06        .12    1.3    .24    1.4 |
| MAX.     134.0      139.0       2.16     .46       1.43    4.4   1.62    4.1 |
| MIN.      34.0      139.0      -2.74     .18        .86   -2.2    .55   -2.4 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .23  ADJ.SD    1.11  SEPARATION  4.82  Item   RELIABILITY  .96 |
|MODEL RMSE    .23  ADJ.SD    1.11  SEPARATION  4.90  Item   RELIABILITY  .96 |
| S.E. OF Item MEAN = .18                                                     |
+-----------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000
Item RAW SCORE-TO-MEASURE CORRELATION = -.98
5560 DATA POINTS. APPROXIMATE LOG-LIKELIHOOD CHI-SQUARE: 5631.28
```

Table 6 shows a moderately high person reliability index (0.78) with a separation index of 1.89. This indicates the separation of person ability into two levels of ability. There is a high item reliability index of 0.96 and a separation index at 4.82. This indicates a separation of item into approximately five levels of difficulty. The mean infit and outfit for person and item mean squares are expected to be 1.00, and for this data, they are 0.99 for person and 1.00 for item. The mean standardized infit and outfit are expected to be 0.0. From the table, it shows that there are 0.1 and 0.00 for person and items, respectively. This indicates that the data shows an overall acceptable fit as the value for standardized infit standard deviation for person is 0.16 while for item is 0.12.

The analysis continue by examining table of item fit to identify misfit items. From Table 7, it shows that four items were considered misfit as the mean square value is either above 1.2 or standardized infit above 2.0 or both. This refers to items A4Pre, C1post, A6pre and A3Pre.

TABLE 7
ITEM FIT FOR POST-COMPETENCY



```
+----------------------------------------------------------------------------------+
|ENTRY    RAW                          MODEL|   INFIT  |  OUTFIT  |PTMEA|EXACT  MATCH|       |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| OBS%  EXP%| Item  |
|------------------------------------+----------+----------+-----+-----------+-------|
|     4     34    139     2.16    .21|1.43   3.5|1.62   3.4|A-.11| 70.5  77.0| A4pre |
|    35    115    139     -.94    .23|1.13   1.0|1.54   2.1|B .06| 82.0  82.8| C1post|
|     6     39    139     1.94    .20|1.10   1.1|1.37   2.4|C .22| 77.0  74.6| A6pre |
|     3     65    139      .99    .18|1.32   4.4|1.37   4.1|D .05| 48.9  67.3| A3pre |
|    19     37    139     2.03    .21|1.15   1.4|1.26   1.7|E .19| 72.7  75.5| C5pre |
|    40     90    139      .13    .19|1.24   3.2|1.22   1.9|F .11| 56.8  67.9| C6post|
|     2     72    139      .75    .18|1.09   1.4|1.14   1.6|G .28| 61.9  66.3| A2pre |
|    39     56    139     1.30    .19|1.10   1.3|1.13   1.4|H .28| 64.0  69.2| C5post|
|     9    120    139    -1.23    .26|1.03    .2|1.11    .5|I .18| 86.3  86.3| B1pre |
|    28     96    139     -.09    .20| .94   -.7| .84  -1.3|n .42| 70.5  70.5| A8post|
|    14     39    139     1.94    .20| .94   -.6| .90   -.7|m .44| 75.5  74.6| B6pre |
|    18    115    139     -.94    .23| .94   -.4| .84   -.7|l .34| 82.0  82.8| C4pre |
|    30    133    139    -2.54    .42| .94    .0| .57   -.7|k .25| 95.7  95.7| B2post|
|    27     98    139     -.17    .20| .93   -.8| .81  -1.4|j .43| 67.6  71.6| A7post|
|    13     78    139      .55    .18| .92  -1.2| .91  -1.1|i .46| 71.9  66.1| B5pre |
|    31    116    139     -.99    .24| .92   -.5| .70  -1.4|h .38| 84.2  83.5| B3post|
|     5     75    139      .65    .18| .92  -1.3| .89  -1.4|g .47| 69.8  66.1| A5pre |
|    22    108    139     -.59    .21| .91   -.8| .75  -1.4|f .42| 79.1  77.9| A2post|
|    37    106    139     -.50    .21| .91   -.9| .75  -1.5|e .43| 76.3  76.5| C3post|
|    21    111    139     -.73    .22| .90   -.8| .73  -1.4|d .42| 80.6  80.0| A1post|
|    33     92    139      .06    .19| .89  -1.6| .84  -1.4|c .47| 74.1  68.7| B5post|
|    32    116    139     -.99    .24| .87   -.9| .66  -1.5|b .43| 84.2  83.5| B4post|
|     8     74    139      .68    .18| .86  -2.2| .81  -2.4|a .53| 70.5  66.2| A8pre |
|------------------------------------+----------+----------+-----+-----------+-------|
| MEAN   90.0  139.0      .00    .22|1.00    .0| .96   -.1|     | 74.3  74.8|       |
| S.D.   26.5     .0     1.13    .06| .12   1.3| .24   1.4|     |  9.7   8.4|       |
+----------------------------------------------------------------------------------+
```

However, these items were not excluded as it was found that the misfit arises from the person not responding appropriately toward the test items. Hence, the misfit responses were removed from the data set and reanalyze till there are no misfit responses left. The summary of the results is shown in Table 8.

TABLE 8
SUMMARY STATISTICS FOR PERSONS

	Before removal of misfit responses		After removal of all misfit responses					
			(Stage 1)		(Stage 2)		(Stage 3)	
PERSON	MNSQ	ZSTD	MNSQ	ZSTD	MNSQ	ZSTD	MNSQ	ZSTD
Mean	0.99	0.1	0.99	0.10	0.99	0.10	0.99	0.10
Standard Deviation	0.16	1.0	0.16	1.00	0.16	1.00	0.16	1.0
Reliability index	0.78		0.78		0.80		0.80	
Separation index	1.89		1.90		2.01		2.02	
ITEM	MNSQ	ZSTD	MNSQ	ZSTD	MNSQ	ZSTD	MNSQ	ZSTD
Mean	1.00	0.00	1.00	0.10	1.00	0.10	1.00	0.10
Standard Deviation	0.12	1.3	0.11	1.3	0.10	1.20	0.10	1.1
Reliability index	0.96		0.96		0.96		0.96	
Separation index	4.82		4.85		5.01		5.01	

Table 8 shows that the person MNSQ and ZSTD score show a more acceptable fit at consistently around 1.0 and 0.10 at all three stages. Separation index for person shows an increasing trend from 1.89 to 2.02. The indices indicate that there is more spread of the person ability positions. The person reliability index estimated for this data is 0.78, and it increases to 0.80 when all misfit responses were removed. This is considered as sufficiently good to test the students' competency. Summary statistics for competency items is shown in Table 8. As for the items reliability index, there were no significant changes in the MNSQ and ZSTD. However, only a slight increase in separation index was observed at each stage.

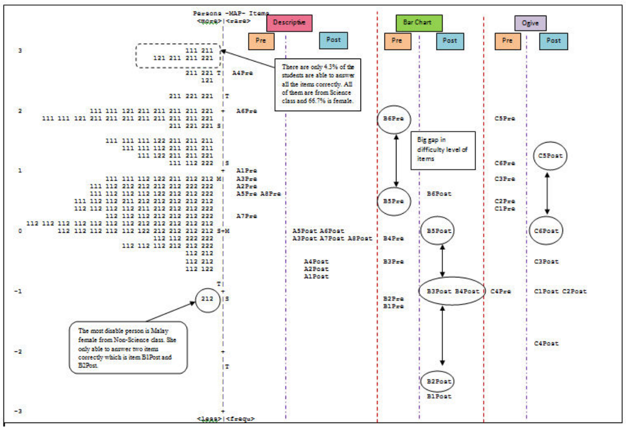


Fig. 6 PIDM for students' competency in learning statistics (After removal of misfit responses)

Fig. 6 shows that 6 students are classified as most able students as they are able to answer all items correctly in the competency test. Slightly more than half (66.7%) of them are female. All of the able students are from the Science class, and 66.7% are Malay students. The PIDM also shows that only one student was found to be least able as she was able to answer only two items correctly. This student was identified as Malay and came from the non-science class and able to answer two easy items from Section B (Bar Graph) which are B1Post and B2Post. Most post-competency items are located well below the pre-competency items. This suggests that students are able to answer the test items correctly after attending the statistics lessons.

Looking at the items within each topic, there were some gaps between items. For example, in the Bar chart section, there is a big gap between items such as B6 Pre and B5 Pre. These two items have large differences in the difficulty level, which means that item B6 Pre is more difficult compared to item B5 Pre. The same goes to Section C where moderately large gaps were observed in between the items indicating a large gap in the difficulty level.

Then, logit for competency test items before removal of misfit responses was compared with item logit after removal of misfit responses.

TABLE 9
COMPETENCY-ITEM LOGIT BEFORE AND AFTER REMOVAL OF MISFIT RESPONSES

Items	Item logit		Items	Item logit	
	Before removal of misfit responses	After removal of misfit responses		Before removal of misfit responses	After removal of misfit responses
A4pre	2.16	2.29	A5post	-0.05	-0.06
C5pre	2.03	2.03	A8post	-0.09	-0.09
A6pre	1.94	1.94	B4pre	-0.17	-0.17
B6pre	1.94	1.94	A3post	-0.17	-0.17
C5post	1.30	1.30	A7post	-0.17	-0.17
C6pre	1.12	1.12	B3pre	-0.50	-0.50
A1pre	0.99	0.99	C3post	-0.50	-0.50
A3pre	0.99	0.96	A4post	-0.54	-0.55
C3pre	0.82	0.82	A2post	-0.59	-0.59
A2pre	0.75	0.75	A1post	-0.73	-0.73
A8pre	0.68	0.68	C4pre	-0.94	-0.94
A5pre	0.65	0.65	C1post	-0.94	-0.94
B6post	0.62	0.61	C2post	-0.94	-0.94
B5pre	0.55	0.55	B3post	-0.99	-1.00
C2pre	0.48	0.48	B4post	-0.99	-1.00
C1pre	0.38	0.37	B2pre	-1.11	-1.11
A7pre	0.24	0.24	B1pre	-1.23	-1.24
C6post	0.13	0.13	C4post	-1.88	-1.88
B5post	0.06	0.06	B2post	-2.54	-2.55
A6post	-0.02	-0.02	B1post	-2.74	-2.74

Item Logit Increase ⇩ Easy to Difficult

Item Logit Decreases ⇨ Difficult to Easy

Table 9 shows some increment of logit values for all items after removal of misfit responses. It indicates that after removal of the misfit responses all items logit changes from either easy to more difficult to endorse or from difficult to easy to endorse the items. For an example, item A4pre has item logit 2.16 before removal of misfit items and increase to 2.29 logit after removal of misfit response. As the logit values increases, the difficulty of the items also increases. In this case, item A4pre change from easy to difficult to endorse by students. For item B3post, the item logit value decrease from -0.99 before removal of misfit responses to -1.00 after removal of misfit responses. This suggests that the item changes from difficult to easy to endorse by students.

The analysis was carried out by investigating Item Characteristic Curves (ICC) for the most difficult and easiest competency test items. Pre-competency item, namely A4 (Descriptive statistics) was considered difficult by students while post-competency items B1 (Bar graph) was considered easy by majority of students.

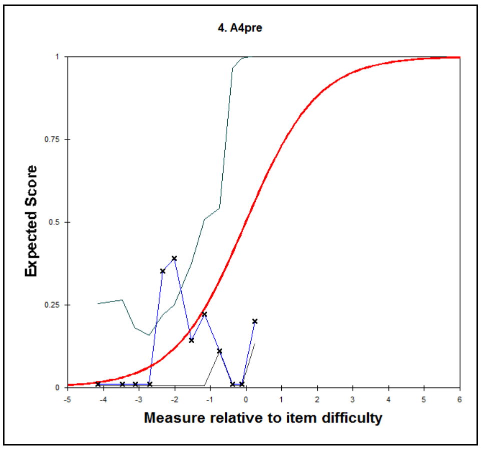


Fig. 7 The expected empirical ICC for Pre-competency Item A4 (Descriptive statistics)

Fig. 7 shows the ICC for the pre-competency Item A4 (Descriptive statistics). This item relates to the identification of nominal data. It can be seen that the pattern of responses does not lie along the sigmoid curve. Most of the responses were located at the bottom of the curve, and it indicates that most students failed to answer this test item correctly. Then for the easiest item which is post-competency B1, Fig. 8 shows that this item was easy to endorse as the pattern of the response was located within the expected location and falls at the top of the sigmoid curve. This suggests that students have nearly 0.95 probability of answering these test items correctly.

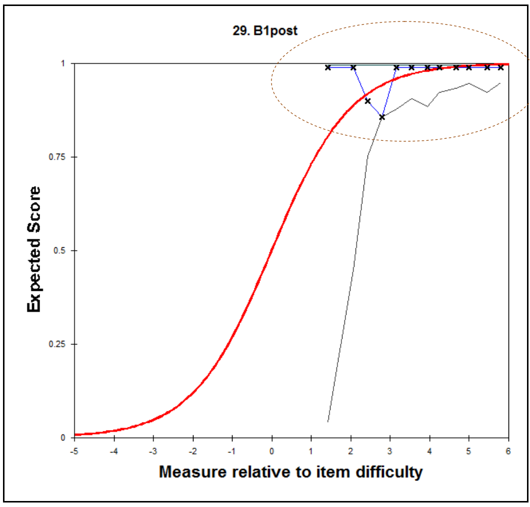


Fig. 8 The expected empirical ICC for Post-competency Item B1 (Bar graph)

The probability of answering the competency test items correctly can be determined by computing the probability of success for each item and person. Table 10 shows that high negative logit for person measure indicates that students are less able to attempt the items correctly while high positive person measures indicate that the students are more able. For item measure, negative logit suggest that items are considered easy for students to answer, while high positive logit indicates that items are more difficult for students to attempt.

TABLE 10
PROBABILITY OF SUCCESS FOR COMPETENCY IN LEARNING ELEMENTARY STATISTICS

	Person	Logit Person Measure	Items	Logit Item Measure	211	111	211	221	211	211	221	211	211	111	212	212	212	212	212	112	212	122	112	212
1	211	3.00	A4pre	2.16	0.70	0.70	0.62	0.55	0.49	0.44	0.39	0.35	0.31	0.28	0.22	0.20	0.18	0.15	0.12	0.09	0.08	0.07	0.06	0.03
2	111	3.00	C5pre	2.03	0.73	0.73	0.65	0.58	0.52	0.47	0.42	0.38	0.34	0.31	0.25	0.22	0.20	0.16	0.13	0.11	0.09	0.07	0.07	0.04
3	211	2.65	A6pre	1.94	0.74	0.74	0.67	0.60	0.54	0.49	0.45	0.40	0.36	0.33	0.27	0.24	0.22	0.18	0.14	0.11	0.10	0.08	0.07	0.04
4	221	2.36	B6pre	1.94	0.74	0.74	0.67	0.60	0.54	0.49	0.45	0.40	0.36	0.33	0.27	0.24	0.22	0.18	0.14	0.11	0.10	0.08	0.07	0.04
5	211	2.12	C5post	1.30	0.85	0.85	0.79	0.74	0.69	0.65	0.60	0.56	0.52	0.48	0.41	0.37	0.34	0.29	0.24	0.20	0.18	0.14	0.13	0.08
6	211	1.91	C6pre	1.12	0.87	0.87	0.82	0.78	0.73	0.69	0.65	0.60	0.56	0.52	0.45	0.42	0.38	0.33	0.27	0.23	0.20	0.17	0.15	0.09
7	221	1.72	A1pre	0.99	0.88	0.88	0.84	0.80	0.76	0.72	0.67	0.63	0.59	0.55	0.48	0.45	0.42	0.35	0.30	0.25	0.23	0.18	0.17	0.10
8	211	1.54	A3pre	0.99	0.88	0.88	0.84	0.80	0.76	0.72	0.67	0.63	0.59	0.55	0.48	0.45	0.42	0.35	0.30	0.25	0.23	0.18	0.17	0.10
9	211	1.37	A8pre	0.68	0.91	0.91	0.88	0.84	0.81	0.77	0.74	0.70	0.67	0.63	0.56	0.52	0.49	0.43	0.37	0.31	0.28	0.24	0.21	0.13
10	111	1.21	A5pre	0.65	0.91	0.91	0.88	0.85	0.81	0.78	0.74	0.71	0.67	0.64	0.57	0.53	0.50	0.44	0.38	0.32	0.29	0.24	0.22	0.14
11	212	0.92	B6post	0.62	0.92	0.92	0.88	0.85	0.82	0.78	0.75	0.72	0.68	0.64	0.57	0.54	0.51	0.44	0.38	0.33	0.30	0.25	0.22	0.14
12	212	0.78	A1post	-0.73	0.98	0.98	0.97	0.96	0.95	0.93	0.92	0.91	0.89	0.87	0.84	0.82	0.80	0.75	0.70	0.65	0.62	0.56	0.52	0.38
13	212	0.65	C1post	-0.94	0.98	0.98	0.97	0.96	0.96	0.95	0.93	0.92	0.91	0.90	0.87	0.85	0.83	0.79	0.75	0.70	0.67	0.61	0.58	0.44
14	212	0.39	C2post	-0.94	0.98	0.98	0.97	0.96	0.96	0.95	0.93	0.92	0.91	0.90	0.87	0.85	0.83	0.79	0.75	0.70	0.67	0.61	0.58	0.44
15	212	0.14	B3post	-0.99	0.98	0.98	0.97	0.97	0.96	0.95	0.94	0.93	0.91	0.90	0.87	0.85	0.84	0.80	0.76	0.71	0.68	0.62	0.59	0.45
16	112	-0.11	B4post	-0.99	0.98	0.98	0.97	0.97	0.96	0.95	0.94	0.93	0.91	0.90	0.87	0.85	0.84	0.80	0.76	0.71	0.68	0.62	0.59	0.45
17	212	-0.24	B1pre	-1.23	0.99	0.99	0.98	0.97	0.97	0.96	0.95	0.94	0.93	0.92	0.90	0.88	0.87	0.83	0.80	0.75	0.73	0.67	0.65	0.51
18	122	-0.50	C4post	-1.88	0.99	0.99	0.99	0.99	0.98	0.98	0.97	0.97	0.96	0.96	0.94	0.93	0.93	0.91	0.88	0.85	0.84	0.80	0.78	0.66
19	112	-0.63	B2post	-2.54	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.97	0.97	0.96	0.95	0.94	0.92	0.91	0.88	0.87	0.79
20	212	-1.20	B1post	-2.74	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.97	0.97	0.97	0.96	0.95	0.93	0.92	0.90	0.89	0.82

As a simple example, probability of success for person with ID 211 (Malay female from Science class) answering question A4 Pre correctly is shown below:

$$(Y = \beta_i - \delta_i) = \frac{\exp(3.00 - 2.16)}{1 + \exp(3.00 - 2.16)} = 0.6985 = 0.70 \qquad (3)$$

Person 211 was considered as one of the most able students since her logit person measure is 3.00, the highest among the students. The probability that this student was able to answer the most difficult item (A4 Pre) correctly is 0.70. She also obtained the highest probability of answering correctly items B2 Post, B1 Post (Bar graph), 0.99 for Item B1 Pre and C4 Post (Ogive).

Further analysis was carried out to investigate the signficant difference between the gender in their response towards the competency test items using differential item functioning (DIF). Based on Table 11, only 2 items from 40 items in the competency test indicate a significant difference between male and female students (GDIF in t value greater than 2.0 logit and p value < 0.05).

TABLE 11
DIF BASED ON GENDER-COMPETENCY IN LEARNING STATISTICS

DIF Measure Group 1	DIF Measure Group 2	DIF Contrast	t	probability	Item
1.59	2.94	1.35	±3.04	0.0029*	A4Pre
-1.77	-0.49	1.28	±2.44	0.0161*	B3Post

*p-value < α=0.05

Fig. 9 shows DIF chart based on gender for competency test items and it shows that items A4pre and B3post are misfit as these items have local measure (t-value) greater than 2.0 logit. For item A4pre, it shows that this question is easier to be endorsed by female students. Further analysis as shown on the ICC in Fig. 10 reveal poor understanding and knowledge about variables and types of data prior to the teaching of the topics and a tremendous improvement in students' understanding of the topics at the end of statistics lessons.

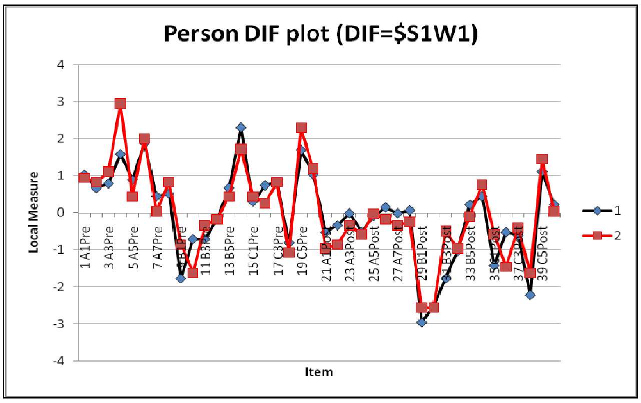


Fig. 9 DIF chart for Gender Differential Item Functioning (GDIF)

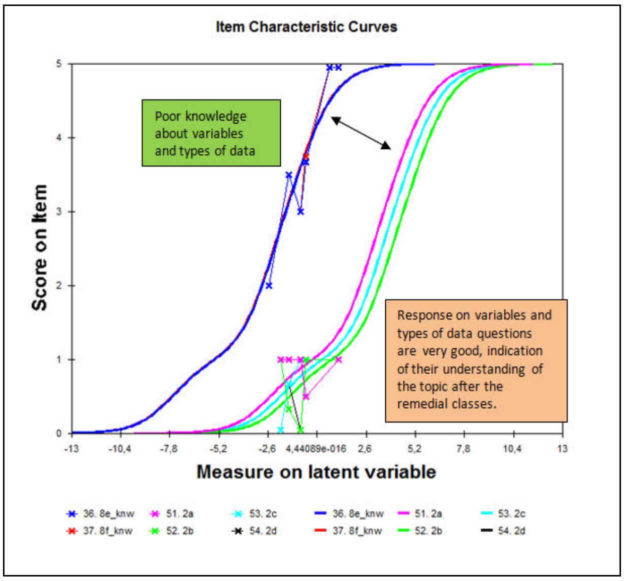


Fig. 10 DIF chart for Gender Differential Item Functioning (GDIF)

IV. Conclusion

Assessing quality of survey response using Rasch probabilistic model has shown to be an alternative and effective approach in improving the quality of survey data. In the study on students' attitude toward statistics and competency in learning statistics, Rasch model has demonstrated its ability to identify and exclude misfit items and inappropriate responses while maintaining other responses which are considered appropriate. The process of calibrating students' ability and item difficulty took place following the construction of the logit scale ruler which saw both the reliability and validity of the items and person response to increase upon deletion of misfit items and person response. The study reveals that reliability index for person responses and items can be enhanced when misfit responses were identified and excluded from the sample.

The results in this study is parallel with [8]-[10] which highlights the application of Rasch measurement model in overcoming measurement hurdles in statistical education research and in the appraisal of course learning outcomes. In the measurement of students' attitude towards statistics, this study has revealed that there exist anxiety towards making errors in statistics computation among male students who came from the non-science class. While in measuring students' competency in learning statistics, a high percentage of the students were not able to attempt the test items correctly prior to the lesson of each statistics topic. Test questions which involve more computations were mostly difficult to endorse by students.

Based on the study outcome, it is recommended that statistics topics or concepts be given more attention at the secondary school level. More emphasis on learning statistics should also be given to the non-science students who were also found to be academically weak. Rasch measurement which is based on the conditional probabilistic model is recommended as an alternative and effective tool in assessing the reliability and validity of items or constructs. It can also be used to assess and improve the quality of survey data to ensure that only true items and true response are included for further analysis. Hence future research in assessing teaching and learning of statistics should include the use of Rasch measurement tools.

Acknowledgment

This research is funded by the Ministry of Higher Education of Malaysia under the Fundamental Research Grant Scheme. The authors also wish to express their gratitude to the Center of Engineering Education Research, Universiti Kebangsaan Malaysia for the collaboration and the survey researchers who were responsible for collecting the data. This investigation would not have been possible without their consent.

References

[1] J. M. Linacre. (2002). "Understanding Rasch measurement: Optimal Rating Scale Category Effectiveness". *Journal of Applied Measurement*, Volume 3, pp 85-106, 2002.

[2] B. D. Wright, G. N. Masters. *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press, 1982.

[3] A. A. Azrilah, M. Azlinah, A. Noorhabibah, Z. Sohaimi, Z. Azami, Z, Hamza, A. G., M. Saidfuddin. "Application of Rasch Model in Validating the Constructs of Measurement Instrument." *NAUN International Journal of Education and Information Technologies*, Volume 2, Number 2, pp 105-112, 2008.

[4] N. L. Abu Kassim, N.Z. Ismail, Z. Mahmud , M.S. Zainol. "Measuring Students Understanding of Statistical Concepts using Rasch Measurement." *International Journal of Innovation, Management and Technology*, Volume 1, Number 1, pp 13-19, 2010.

[5] O. Nor Hasmaniza, M. Zamalia. "Relationship between the Attitude towards and Competency in Learning Statistics among Upper Secondary School Students at SMK Bandar Baru Sg Buloh." *Unpublished Resport*, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia, 2009.

[6] E.V. Smith. "Detecting and evaluation the impact of multidimensionality using item fit statistics and principal component analysis of residuals." *Journal of Applied Measurement*, Volume 3, pp 205-231, 2002.

[7] A. Tennant, J.F. Pallant. " DIF matters: A Practical Approach to test if Differential Item Functioning Make a Difference." *Rasch Measurement Transaction*, Volume 20, Number 4, pp 1082-1084, 2007.

[8] A. Siti Rahayah, I. Rodiah, M.I. Noriah. "Differential Item Functioning in Malaysian Generic Skills Instrument (MyGSI)." *Jurnal Pendidikan Malaysia*, Volume 35, Number 1, pp 1-10, 2010.

[8] Z. Mahmud, M.S. Masodi, and A.A.Aziz, "Overcoming Measurement Hurdles in Statistical Education Research", *WSEAS Computers and Simulations in Modern Science*, vol. 5, pp. 78-85, 2011.

[9] M.S.Masodi, A. Mohamed, A.A. Aziz, N.Arshad, & S.Zakaria, "Appraisal of Course Learning Outcomes using Rasch measurement: A case study in Information Technology Education", *International Journal of Systems International Journal of Systems Applications, Engineering & Development*; Issue 4, vol.1, University Press, UK. pp.164-172, 2007.

[10] S.A. Osman, W.H.W.Badaruzzaman, R.Hamid, K.Taib, A.R. Khalim, N.Hamzah, and O. Jaafar, "Assessment on Students' Performance Using Rasch Model in Reinforced Concrete Design Course Examination", *Recent Researches in Education* in Computers and Simulations in Modern Science, vol. 5, pp. 193 – 198, 2011.