Exploring students’ perceived and actual ability in solving statistical problems based on Rasch measurement tools

Nor Azila Che Musa¹, Zamalia Mahmud¹ and Norhayati Baharun²

¹Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.
²Universiti Teknologi MARA Cawangan Perak, Kampus Seri Iskandar, 32610 Seri Iskandar, Perak, Malaysia.

E-mail:norazilachemusa@gmail.com

Abstract. One of the important skills required from any student who are learning statistics is knowing how to solve statistical problems correctly using appropriate statistical methods. This will enable them to arrive at a conclusion and make a significant contribution and decision for the society. In this study, a group of 22 students majoring in statistics at UiTM Shah Alam were given problems relating to topics on testing of hypothesis which require them to solve the problems using confidence interval, traditional and p-value approach. Hypothesis testing is one of the techniques used in solving real problems and it is listed as one of the difficult concepts for students to grasp. The objectives of this study is to explore students’ perceived and actual ability in solving statistical problems and to determine which item in statistical problem solving that students find difficult to grasp. Students’ perceived and actual ability were measured based on the instruments developed from the respective topics. Rasch measurement tools such as Wright map and item measures for fit statistics were used to accomplish the objectives. Data were collected and analysed using Winsteps 3.90 software which is developed based on the Rasch measurement model. The results showed that students’ perceived themselves as moderately competent in solving the statistical problems using confidence interval and p-value approach even though their actual performance showed otherwise. Item measures for fit statistics also showed that the maximum estimated measures were found on two problems. These measures indicate that none of the students have attempted these problems correctly due to reasons which include their lack of understanding in confidence interval and probability values.

1. Introduction

Solving statistical problems requires the knowledge of various statistical techniques such as computing, analysing, interpreting and drawing conclusions. One of the important skills that is required from any student who are learning statistics is knowing how to solve statistical problems correctly using appropriate statistical methods. This will enable them to arrive at a conclusion and make a significant contribution and decision for the society. Hypothesis testing is recognized as one of the techniques used in solving real problems and it is listed as one of the difficult concepts for students to grasp. How these skills are measured are explained in this study. Firstly, the purpose is to explore students’ ability in solving statistical problems using relevant statistical techniques. Next is to determine which of the problems that student find difficult to solve. Students’ perceived and actual ability were measured based on the instruments developed based on the respective topics.
Hence, the assessment of students’ ability on knowledge and skills in problem solving should not just be based on asking them questions and measure their performance based on raw scores. The assessment should cover the process of arriving at the solutions at every stage, the links between the stages and ability to entertain the fact that there may be more than one solution to a statistical problem [1].

There are measurement approaches that can be used to measure the properties of this kind of skill. Measurement, in its broadest sense, is the process of assigning numbers to categories of observations in such a way as to represent quantities of attributes [2]. However, simply assigning numbers at these different levels does not guarantee that the resulting measures are indeed at those corresponding levels [3]. Therefore, this study used probabilistic model namely, the Rasch Measurement Model which basically referred to as latent trait model and a technique developed for the purposes of item based analysis.

2. Research Design and Analysis Methods

This study employed a case study method which involved only a small number of 22 undergraduate students majoring in Statistics at Universiti Teknologi MARA, Shah Alam. Data was collected using two different instruments which were developed for the purpose of the study. One of the instruments was used to measure students’ perceived ability in solving statistical problems which contained questions about the level of agreement on their knowledge on subtopics in confidence interval and hypothesis testing topic. The other instrument was used to measure their actual ability in solving the problems using questions related to confidence interval and hypothesis testing. Data was analysed based on the Rasch measurement model using Winsteps 3.90. The analysis generally focused on examining the reliability of items and persons, observing the person and item distributions on the Wright map and identify statistical problems which are found difficult by students to solve based on item measures for fit statistics.

Rasch Polytomous model was used to measure students’ perceived and actual ability in solving statistical problems based on the data collected using the instruments. Perceived ability was measured using a 5-point Likert scale items ranging from (1) strongly disagree, to (5) strongly agree. On the other hand, actual ability was measured based on a given set of 11 statistical problems with designated marks allocated for each problem. Based on the corrected problems, total marks (or raw scores) were recorded and later converted to grade A, B, C, D or E. Each grade represents a scale such that grade A correspond to scale 5, grade B to scale 2 until grade E to scale 1. This process of data coding is necessary to enable the data to be analysed using Winsteps. The Rasch Polytomous model estimates the probability that a person will choose a particular response category or an item as:

\[ P_{nij} = \frac{e^{B_n + D_i + F_j}}{1 + e^{B_n + D_i + F_j}} \]  

The probability of respondent \( n \) scoring in category \( j \) for item \( i \) is represent by \( P_{nij} \) which is when person’s answer \( x \) is 1, the probability is calculated using the exponential of person measure or ability of respondent \( n(B_n) \), minus the summation of difficulty of item \( i(D_i) \) and the difficulty of category step \( j(F_j) \) divided to 1 plus the same exponential value. The Rasch analysis places persons \((B_n)\) and items \((D_i)\) on the same measurement scale where the unit of measurement is the logit (logarithm of odds unit). The person’s likely score is defined by the interaction between the person’s measure, the item’s difficulty, and the score’s category threshold.
Person reliability is reported as it uses an item response theory approach where standard error may vary at different points of the latent construct depending on the item information [4]. Wright map was used to calibrate the relationship between students’ perceived and actual ability where the persons (i.e. students) and items are plotted on the same logit scale or ruler. Finally, item measures was computed to identify which statistical problems were found to be difficult to solve by the students.

3. Analysis and results

3.1 Structural information and reliability

The analysis of perceived and actual ability reveals a summary measures that provide the general acceptable fit to the model. Table 1 shows the summary of the structural information including person/item reliability and separation index. The initial indication of data fitness are measured from mean square infit and outfit values. The acceptable value for mean square infit and outfit values for person and item should be close to 1.0 while for mean standardized infit and outfit are between 0 to -0.4. Items outside the range value are either considered overfit or underfit. Overfit means the items are too predictive while underfit means the items are unpredictable [5].

Table 1 shows that the mean square infit and outfit are both close to 0 while standardized infit and outfit are -0.5 and -0.5 respectively. This indicates that the items are slightly underfit and that the data fit the model somewhat better than we would expect which may signal possible redundancy of items [6]. Both standard deviation of mean and standardized infit and outfit for item and person shows some misfit for person and item as the values are far from 0 and exceed 2.0 as the cut-off criterion.

|           | Infit    | Outfit   | Index |
|-----------|----------|----------|-------|
|           | MNSQ ZSTD| MNSQ ZSTD|       |
| PERSON    |          |          | 1.32  |
| Mean      | 1.18 -0.5| 1.10 -0.6|       |
| Standard Deviation | 1.06 2.9 | 0.97 2.9|       |
| Separation|          |          |       |
| Reliability|         |          | 0.64  |
| ITEM      |          |          | 3.90  |
| Mean      | 1.18 -0.5| 1.10 -0.6|       |
| Standard Deviation | 1.06 2.9 | 0.97 2.9|       |
| Separation|          |          |       |
| Reliability|         |          | 0.94  |

Separation is the index of spread of the person positions or item positions [7]. Separation index measures the spread of both items and persons in standard error units [7] and [8]. It can be thought of as the number of levels into which the sample of items and persons can be separated. For an instrument to be useful, separation should exceed 1.0, with higher values of separation representing greater spread of items and persons along a continuum. Separation determines reliability [7] and [9]. Separation index for persons is 1.32 which indicates that students’ perception can be categorized into 1 to 2 level spread of person positions or small range ability. However, separation index for items is 3.90 which indicate that the items can be categorized up to four levels of item difficulty positions. Person reliability index is 0.64 which is moderately consistent and reliable. This reliability can be improved if more students perform better in the actual test. On the other hand, items indicate a strong reliability index at 0.94 which indicate that the survey items have a range of item difficulty that matches with the sample of respondents.

3.2 Wright map of perceived and actual ability

Figure 1 shows the Wright map displaying the distribution of students and survey items side by side as calibrated by the logit ruler. Students are distributed on the left side of the logit ruler (center vertical
Line) while items are distributed on the right side of the map. “M” marks the person and item mean logit, “S” is one standard deviation away from the mean logit, and “T” is two standard deviation away from the mean logit. Items at the top of the scale are not well perceived (or less agreeable) by the respondents while items at the bottom are well perceived (or agreeable) [6]. There are 14 items measuring perceived ability and 11 items measuring actual ability of students.

**Figure 1.** Wright map of perceived and actual ability.

Figure 1 shows that about half of the students are located below the perceived item mean logit 0.00 while the other half are located above the perceived item mean logit. This indicates that half of the students perceived solving the six questions (P3, P11, P9, P10, P8, P5) as rather difficult. Question P2 (Perceived ability of definition of significance level) and P12 (perceived ability of performing p-value approach) are perceived as easy by majority of the students. However, their actual ability revealed otherwise. From the map, items A1d (actual ability of performing p-value approach) and A1b (actual ability of constructing confidence interval) are located at the very top of the map which means that all students are not able to solve the questions correctly. In the actual ability column, there are six items which are located above all students that is, A2b, A1a, A3, A1c, A1b and A1d. These are questions in which all students are not able to solve correctly.
On the other hand, for students who perceived item P14 (perceived ability of performing ANOVA) as difficult to solve is able to perform well on the similar topic A4, which is located at the lowest position of the map (logit -2.0). For item P13 (perceived ability of constructing confidence interval), half of the students perceived they are able to solve this question correctly, however the actual test result shows otherwise where its logit value 4.90 is the highest and located at top right-hand side of the map.

3.3 Item measure and unidimensionality

Item dimensionality is measured by the raw unexplained variance (total) in standardized residual variance in eigenvalue units. The eigenvalue is 23.0 with observed percentage 52.3% which is greater than 40% [10]. This indicates that questions used in this study are related with the content of this subject. Next step is to determine which item does not fit into the Rasch measurement model.

Table 2. Item measure.

Table: 2.

| Entry | Total | Score | Count | Measure | S.E. | MNSQ | ZSTD | INFIT | OUTFIT |
|-------|-------|-------|-------|---------|------|------|------|-------|--------|
| 18    | 22    | 4.86  | 1.81  | MAXIMUM MEASURE | .00 | .00 | 100.0 | 100.0 | A10    |
| 21    | 22    | 4.86  | 1.81  | MAXIMUM MEASURE | .00 | .00 | 100.0 | 100.0 | A1d    |
| 20    | 22    | 2.64  | .47   | 1.35 | 2.5 | 1.73 | 1.1 | .38 | .17 | 95.5 | 84.9 | A1c    |
| 24    | 16   | 1.62  | .36   | 1.13 | .51 | .07 | .3 | .53 | .26 | 59.1 | 58.2 | A1a    |
| 24    | 18   | 1.30  | .26   | 1.81 | 2.1 | 1.86 | 1.2 | -.12 | .29 | 22.7 | 38.5 | A3     |
| 23    | 17   | 1.17  | .25   | 1.50 | 1.5 | 1.30 | 1.0 | .32 | .31 | 27.3 | 36.3 | A2b    |
| 3     | 22    | .47   | .21   | .87 | .41 | .81 | .7 | .25 | .35 | 40.9 | 31.4 | P3     |
| 11    | 55    | .34   | .20   | .34 | .33 | .35 | .33 | .72 | .36 | 68.2 | 33.7 | P11    |
| 9     | 57    | .26   | .20   | .40 | .29 | .40 | .29 | .70 | .36 | 63.6 | 33.8 | P9     |
| 10    | 58    | .22   | .20   | .46 | .25 | .47 | .25 | .58 | .36 | 59.1 | 34.9 | P10    |
| 15    | 58    | .22   | .20   | 3.28 | 5.7 | 3.26 | 5.7 | .21 | .36 | 21.2 | 36.3 | A1d    |
| 8     | 61    | .10   | .20   | .53 | .21 | .53 | .21 | .60 | .36 | 50.0 | 35.8 | P8     |
| 5     | 63    | .02   | .20   | .22 | .44 | .23 | .44 | .79 | .36 | 63.6 | 36.5 | P5     |
| 15    | 66    | .10   | .20   | .74 | .18 | .74 | .18 | .06 | .36 | 45.5 | 36.5 | P14    |
| 13    | 67    | .14   | .20   | .57 | .18 | .58 | .18 | .39 | .36 | 45.5 | 37.0 | P13    |
| 4     | 68    | .17   | .20   | .31 | .36 | .32 | .35 | .68 | .36 | 45.5 | 38.2 | P4     |
| 6     | 71    | .29   | .20   | .37 | .31 | .38 | .31 | .46 | .35 | 54.5 | 37.7 | P6     |
| 7     | 72    | .33   | .20   | .45 | .26 | .46 | .25 | .46 | .35 | 50.0 | 36.7 | P7     |
| 22    | 74    | .41   | .20   | 1.53 | 1.81 | 1.67 | 2.2 | .19 | .35 | 36.4 | 36.6 | A2a    |
| 1     | 81    | .71   | .23   | .40 | .28 | .39 | .29 | .43 | .33 | 59.1 | 31.9 | P1     |
| 2     | 82    | .75   | .23   | .37 | .31 | .38 | .29 | .33 | .33 | 68.2 | 31.9 | P2     |
| 17    | 82    | .75   | .23   | 3.43 | 5.7 | 3.56 | 5.9 | .07 | .33 | .0 | 31.9 | AA3    |
| 12    | 83    | .80   | .23   | .55 | .39 | .52 | .38 | .39 | .33 | 63.6 | 31.5 | P12    |
| 18    | 88    | .16   | .20   | 3.24 | 4.1 | 3.03 | 3.6 | .57 | .24 | 45.5 | 49.1 | A2b    |
| 25    | 102   | .84   | .34   | 1.46 | 1.1 | 1.36 | .6 | .20 | .20 | 63.6 | 64.5 | A4     |
| Mean  | 61.0  | .39   | .36   | 1.18 | .51 | 1.10 | .6 | .49 | .40 | 40.1 |
| P.SD  | 21.7  | .16   | .43   | 1.06 | 2.9 | .97 | 2.9 | .21 | 12.5 |
answer an easy question wrongly. This is considered as inappropriate or unusual response where MNSQ and standardized (ZSTD) values can be located outside the acceptable range. As proposed by the Rasch experts [6] and [10], an acceptable range value for MNSQ is between 0.5 to 1.5 while for ZSTD is between -2.0 to 2.0. In Table 2, all items except AA1, AA2 and AA3 are located within the acceptable range. Hence, it is recommend to do further analysis on those misfit items for the purpose of enhancing the instrument such as rephrasing or replacing the items. Maximum estimated measures appears for items A1b and A1d. This indicates that the sample obtained an extreme minimum (zero) score on these items. From the responses, all students answered these two questions incorrectly and gain no marks.

4. Discussion and conclusion
This study revealed that students who perceived themselves as able to solve the statistical problems may have difficulties in actually solving the questions. This is demonstrated by the test outcome of performing hypothesis test using the confidence interval and p-value approach, particularly in the computation of p-values. These findings can be considered as a guide for educators to strategize their teaching approach and prepare their lessons to focus more, for example on the computations of p-values both manually and using statistical softwares. Generally, Rasch measurement model is an effective tool to determine actual ability of the students and to diagnose exactly where students are having difficulty the most in understanding statistical concepts and in solving statistical problems [9].

Acknowledgement
We would like to thank Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam and UiTM Lestari Grant RMI file number 600-RMI/DANA 5/3/LESTARI (109/2015) for the support and providing us with the conference fund to present this paper at ICoAIMS 2017.

References
[1] Gibson L, Mariott J and Neville D 2007 Assessing Statistical Problem Solving. International Association for Statistics Education.
[2] Nunnally J C and Bernstein I H 1994 *Psychometric Theory* (3rd ed). (McGraw-Hill, Inc.)
[3] Michell J 1990 *An Introduction to the Logic of Psychological Measurement*. (Hillsdale, New Jersey: Lawrence Erlbaum)
[4] Makransky G and Bilenberg N 2014 Psychometric properties of the parent and teacher ADHD rating scale (ADHD-RS): Measurement invariance across gender, age and informant. *Assessment, 21*(6), 694-705.
[5] Siti Rahayah A, Suriana M A, Shafiza M, Shah Nazim S, Noriah M I, Rosseni D, Abdul Ghafur A, Hamidah Y, Rosadah M and Siti Fatimah M Y 2010 Validity of UKM1 intelligence test using Rasch analysis, *Procedia – Social and Behavioral Sciences* 7(C), 205-09.
[6] Bond T G and Fox C M 2007 *Applying the Rasch model: Fundamental measurement in the Human Sciences*, (New Jersey: Lawrence Erlbaum Associates)
[7] Wright B, Masters G 1982 Rating scale analysis: Rasch measurement. Chicago: MESA Press.
[8] Andrich D 2011 Rating scales and Rasch measurement. *Expert Rev Pharmacoeconomics Outcomes Res* 11, 571-85.
[9] Zamalia M, Nor Azura M G and Rosli A R 2013 Assessing students’ learning ability in a postgraduate Statistical course: A Rasch analysis, *Procedia – Social and Behavioral Sciences*. 89, 890-4.
[10] Linacre J M 2008 *User’s guide to Winsteps & Ministeps Rasch model computer programs*. Retrieved from www.winsteps.com.