Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 129 (2014) 133 – 139

Procedia
Social and Behavioral Sciences

ICIMTR 2013

International Conference on Innovation, Management and Technology Research,
Malaysia, 22 – 23 September, 2013

A Rasch Model Analysis on Secondary Students' Statistical Reasoning Ability in Descriptive Statistics

Shiau Wei Chan[a*], Zaleha Ismail[b], Bambang Sumintono[c]

[a]*Ph.D, University of Technology Malaysia, 81310 Skudai, Johor, Malaysia*
[b]*Associate Professor Dr, University of Technology Malaysia, 81310 Skudai, Johor, Malaysia*
[c]*Ph.D, University of Technology Malaysia, 81310 Skudai, Johor, Malaysia*

Abstract

Earlier studies provided evidence that students had misconceptions in statistical reasoning. This study was aimed to determine the statistical reasoning ability in descriptive statistics among the tenth-grade students from Malaysian secondary schools. There were 412 participants randomly selected for this study. An instrument called statistical reasoning test which consisted of five questions with 16 items was utilized. The results obtained were analyzed using Rasch measurement model including person reliability, variable map, and person analysis report. The findings indicated that statistical reasoning ability of tenth graders seem to be at a poor level. Thus, the instructors and researchers should put efforts to enhance students' statistical reasoning ability in future studies.

Selection and peer-review under responsibility of Universiti Malaysia Kelantan

Keywords:Statistical reasoning; descriptive statistics; Rasch measurement model

1. Introduction

Previous studies demonstrated that students from primary to university level harbored misconceptions in descriptive statistics, including reasoning about measures of central tendency (Cooper & Shore, 2008; Huck, 2009; Olani et al., 2011), representations of data (Lee and Meletiou-Mavrothesis, 2003; Sharma, 2005; Cooper and Shore, 2008), variability (delMas and Liu, 2005; Matthew and Clark,

* Corresponding author. Tel.: +6-012-781-1269
E-mail address: shiauweichan@yahoo.com

Selection and peer-review under responsibility of Universiti Malaysia Kelantan
doi:10.1016/j.sbspro.2014.03.658

2007; Huck, 2009), and distribution (Lee and Meletiou-Mavrothesis, 2003). Admittedly, such confounding situation implied that students are facing difficulties in learning statistics. Statistical reasoning is defined as “the way people reason with statistical ideas and make sense of statistical information. It involves making interpretations based on sets of data, or statistical summaries of data. Students need to be able to combine ideas about data and chance, which leads to making inferences and interpreting statistical results (Garfield and Chance, 2000, p. 101)”. Lovett (2001) described statistical reasoning as the usage of statistical ideas and tools is needed to summarize, draw assumptions and make conclusions from the data. During the 1950s and 1960s, Piaget and Inhelder (1975) commenced the seminal work on probability. Since then, the majority of the researchers focused on the probabilistic thinking and steadily shifted towards statistical reasoning. Their work has become the underpinning study of the growth of statistical reasoning.

Substantial studies had been done for statistics reasoning in other countries, but the actual situation in Malaysian school is yet to be understood. Thus, this study was conducted to scrutinize tenth-grade students’ statistical reasoning ability in descriptive statistics using Rasch measurement model. The research question driving this study is, ‘how is the statistical reasoning ability among the tenth-grade students in secondary schools?’

2. Rasch Measurement Model

This study employs Item Response Theory (IRT) in conjunction with Rasch measurement model. IRT is an alternative measurement framework apart from Classical Test Theory (CTT) (Gorin and Embretson, 2007). CTT is a psychometric technique that allows the presumption of test results, for instance the items’ difficulty and the individual’s aptitude (Alagumalai and Curtis, 2005). Meanwhile, IRT is a psychometric technique focusing on the response given by an individual to a particular test item as affected by the qualities of the item and individual background. IRT is more complex than CTT in terms of computation, but it has more advantages if compared to CTT (Gorin and Embretson, 2007). According to Magno (2009), the estimates of item difficulty in IRT remain the same for two different samples, b ut not so in CTT. Besides, the difficulty indices of items in IRT are more constant than CTT. Moreover, in IRT, the internal consistencies of test do not change for two different samples, but they become unstable in CTT. Furthermore, IRT has considerably fewer measurement errors than CTT. To conclude, IRT is regarded as the best techniques to determine students’ achievement from time to time among the test-equating procedures.

Rasch measurement model or one parameter model is the simplest IRT model and it has strong measurement properties (Afrassa, 2005). The probability of people getting the same item correctly resort to two parameters in Rasch model, which are item difficulty and person ability (Bond and Fox, 2007). According to Wright (1977), there are numerous benefits using Rasch model in the test measurement. Firstly, Rasch model can evaluate whether the item is fit and identify if item bias exists. Secondly, its item calibration is not influenced by the ability of sample, which means it is sample free. Thirdly, standard error of calibration can be exploited to examine the precision of each item. Fourthly, Rasch model can estimate the item difficulties from various samples and convert them into a common scale. Hence, the item banks can be equated automatically as a common calibration to be shared by all items. Fifthly, the ability of two people can be compared although they do not have any item in common by transforming the ability estimates into a common scale. This is called test-free person measurement (Tinsky and Dawis, 1977). Sixthly, Chi-square of person fit can be utilized to assess measurement quality. Lastly, by using Rasch model, it makes the construction and design of the best test as well as tailored testing easier to be governed.

3. Methodology

3.1 Participants

In this study, the participants were 412 tenth-grade students from nine secondary schools in Malaysia. They were chosen randomly from fourteen classes. There were 172 male students (41.74 %) and 240 female students (58.25 %). The participants comprised of 229 Malay students (55.58 %), 159 Chinese students (38.59 %), 22 Indian students (5.34 %), an Iban student (0.24 %), and a Kadazan student (0.24 %). Their ages are between 16 and 17 years old. This study was conducted at the end of November for year 2011. All the participants possessed prior knowledge of descriptive statistics as they had already studied about them. The statistical reasoning test was distributed to the students during the mathematics or additional mathematics period in the classroom. The students were given an hour to finish all the problems. The results of the test were then utilized as data in this study. Each student was labeled using code owing to the sensitive ethical issue, for instance the B in B033MC represents school, 033 refers to participants, M is male, and C is Chinese.

3.2 Instrumentation

The instrument employed in this study is a statistical reasoning test with five questions that comprised of 16 items of descriptive statistics. The purpose of this test was to examine students' statistical reasoning ability in descriptive statistics. The topics that were covered in this test were average, weighted mean, measure of central tendency, and standard deviation. In the first section of each question, students were required to interpret the data and link one concept to another and apply their knowledge. Students were also necessitated to provide their reasoning in the second section of each question.

In this study, a Rasch measurement model software named WINSTEPS version 3.73 was utilized for dichotomous responses, i.e. items with only two potential responses (True and False). In the summary statistics, Cronbach-alpha, α, informs us on the test reliability or internal consistency reliability. The Cronbach-alpha value for statistical reasoning test was 0.66 with valid responses of 74.3 %. According to Azrilah Abd Aziz, et al. (2008), the Cronbach-alpha value is acceptable because it goes beyond the minimum acceptable value which is 0.6 at 95 % confidence interval; $p = 0.05$.

Item reliability means the reproducibility of item placements if they are given to another sample that has the same characteristics (Bond and Fox, 2007). The item reliability was excellent, i.e. 0.99 based on the rating scale instrument quality criteria (Fisher, 2007). This indicated that the items had large difficulty range and sample of students (Linacre, 1991-2008). Item reliability also verifies that the instrument constructs validity. Separation is the distribution of position for the person and item along the variable. Since the item separation was 8.89, which was higher than 1.0, it showed that the items had adequate spread (Gracia, 2005).

4. Findings

4.1 Person reliability

In general, Cronbach alpha value gives us test reliability, but does not tell us if we have problem with the person or the item. Nevertheless, by using Rasch measurement model, it can inform us about the person reliability and item reliability. Person reliability refers to the reproducibility of each person's sequence of order if they are given another set of items assessing the same construct (Wright and Master, 1982). Based on the rating scale instrument quality criteria (Fisher, 2007), the person reliability in this study was relatively poor, i.e. 0.43. To increase the person reliability, the students' ability range ought to be widened (Linacre, 1991-2008) and more items of test should be added. The value of person separation was below 0.86 (below 1.0). Hence, it showed that the students could not be well -distinguished (Gracia, 2005).

4.2 Variable map

The variable map demonstrates the distribution of students' ability and item difficulty on a same logit scale. The ability of students is listed on the left side of the map while the item difficulty is on the right side of the map. Higher logits represents students with higher ability (left side) and more difficult items (right side) and vice versa (Iramaneerat, Smith, and Smith, 2008). Through the variable map, it allows us to identify if the items match the ability of the students.

Logits 0 is set as average of test items (Iramaneerat, Smith, and Smith, 2008). From the variable map, we can notice that most of the students are located below the average of test items. Only a few students with higher ability are placed at +2.66 logits and some weak students are placed at -3.78 logits. The values of logits obtained from the maximum measure and minimum measure.. Thus, we can state that students' ability in statistical reasoning was incredibly low because most of them could not solve the questions and the items were fairly difficult for them. In other words, the items are considered not functioning well enough to fit the ability of students and segregated them into discrepancy level of ability. The reason was that students were not familiar with statistical reasoning items in this study and they were not taught to answer those types of questions in school as well.

We also can observe that there are eight difficult items which are above zero logits which includes Q3d2, Q5b, Q3c1, Q4b, Q5a, Q4a, Q3d1, Q3c1 and Q3a2. Question 3b2 plotted at +2.73 logits is the most difficult item and Question 3b1 plotted at -3.11 logits is the easiest item in statistical reasoning test. The value of logits gained from the maximum and minimum measures. There are three items that measure the same level of difficulty including Q2a, Q3a1, and Q3a2. Hence, it is suggested that two of the items have to be eliminated and improved by adding other items that can discriminate the students better.

4.3 Person analysis report

All the students in this study were ranked according to the MEASURE value from highest to the lowest. From the value of MEASURE, we notice that the student with the highest ability at +2.99 logits was student N390FM. Meanwhile, the students with lowest ability at -5.17 logits included student G174FM, A028FC, G187MM, and K280FM. There were students who had gotten all items correct, such as student N390FM had tried 6 items and she got correct for all. There were also students who had all items answered wrongly, for example student K280FM attempted 16 items, but all answers were incorrect.

There are five good fitting students with their infit and outfit mean square values close to one. They were considered as an ideal Rasch model response string (Bond and Fox, 2007) and included student M343MC, D083FM, N409FI, F128FC, and A001FC. Besides that, ten students were categorized as overfitting to the model and demonstrated Guttman pattern because their infit and outfit mean square were too low (Bond and Fox, 2007), for instance student M348MC with infit MNSQ 0.56 and outfit MNSQ 0.25; student F137MC with infit MNSQ 0.49 and outfit MNSQ 0.38; and student H208FC with infit MNSQ 0.39 and outfit MNSQ 0.32.

Furthermore, there were ten students who were regarded as misfitting, i.e. student L306MK, M364FM, J251FM, H215MI, J250FM, N308FM, K295FM, L325FM, and L302FM. Their values of point-measure correlation were negative and had infit mean-square values greater than the sum of infit mean-square and standardized statistics. The total infit mean-square was 0.98 and the standard deviation was 0.51. Therefore, the range of infit for each person was from 0.47 to 1.49 by adding 0.98 and 0.51 as well as subtracting 0.51 from 0.98. For example, point-measure correlation for student L306MK was -

0.82 and its infit mean-square was 2.06, which was more than the range of infit. Another student K295FM had -0.28 for point-measure correlation and 2.16 for infit mean-square. Again, that was greater than the range of infit.

5. Discussion

The research question of this study was, 'how is the statistical reasoning ability among the tenth-grade students in secondary schools?' In the variable map of Rasch measurement model analysis, we notice that most of the students are located below the zero logits. Lower logits manifests students with lower ability. This proved that the students' statistical reasoning ability was absolutely poor. On the other hand, the measurement of person reliability using Rasch measurement model was poor at only 0.43 and person separation was also very low, i.e. 0.86. Students were not differentiated well in this study due to inadequate items and narrow ability range. Consequently, to gain good person reliability, it is recommended that more items should be added into the test and students' ability range needs to be broadened. In addition, for the person analysis report, only five out of 412 students belonged to good fitting type. There were 20 misfitting students where 10 of them were overfitting; ten of them had negative value of point-measure correlation which had impinged on the results. It is suggests that these ten students are to be eliminated from this study.

6. Conclusion

To conclude, the statistical reasoning ability among tenth-graders from secondary school in Malaysia is still at the unsatisfactory stage, particularly in descriptive statistics. Therefore, appropriate actions should be taken in order to enhance students' statistical reasoning ability throughout the age level. In future studies, similar .studies can be implemented in terms of gender, age, background information, and aptitude.

Acknowledgements

This research was funded by the Ministry of Higher Education of Malaysia and University /of Technology Malaysia (Vote No. R.J130000.7831.4L068).

References

Afrassa, T.M. (2005). Monitoring mathematics achievement over time. In S. Alagumalai, D.D.Curtis, and N. Hungi (Eds.), *Applied Rasch Measurement: A Book of Exemplars: Papers in Honour of John P. Keeves* (pp.61-77). Dordrecht, The Netherlands: Springer.

Alagumalai, S., & Curtis, D.D. (2005). Classical test theory. In S. Alagumalai, D.D.Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars: Papers in honour of John P. Keeves* (pp.1-14). Dordrecht, The Netherlands: Springer.

Aziz, A.A., Mohamed, A., Arshad, N., Zakaria, S., Zaharim, A., Ghulman, H.A., & Masodi, M.S. (2008). Application of Rasch model in validating the construct of measurement instrument. *International Journal of Education and Information Technologies*, 2(2), 105 – 112.

Bond, T. G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in human sciences*. New Jersey: Lawrence Album Associates.

Capraro, M.M., Kulm, G., & Capraro, R.M. (2005). Middle grades: Misconceptions in statistical thinking. *School Science and Mathematics*, 105(4), 165-174.

Cooper, L.L., & Shore, F.S. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education*, 16(2).

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, 4(1), 55-82.

Fisher, W. P. (*2007). Rating scale instrument quality criteria. Rasch Measurement Transactions*, 21(1), 1095.

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning*, 2(1&2), 99-125.

Gorin, J. S., & Embretson, S. E. (2007). Item response theory and Rasch models. In D. McKay (Ed.) *Handbook of research methods in abnormal and clinical psychology* (pp. 314-334). Los Angeles: Sage Publications.

Gracia, S. (2005). Analyzing CSR implementation with the Rasch model. *Faculty Publications*. Paper 271.
Huck, S.W. (2009). *Statistical Misconceptions*. New York: Psychology Press, Taylor & Francis.

Iramaneerat, C., Smith Jr. E. V., & Smith R.M. (2008). An introduction to Rasch measurement. In J.W. Osborn (Ed.), *Best practices in quantitative methods* (pp. 50-70). Thousand Qaks, California: Sage Publications, Inc.

Lee, C., & Meletiou-Mavrotheris, M. (2003). Some difficulties of learning histograms in introductory statistics. *Joint Statistical Meetings – Section on Statistical Education*. Retrieved on 17 October, 2011, from http://www.statlit.org/PDF/2003LeeASA.pdf

Linacre, J.M. (1991-2008). *A user's guide to Winsteps/Ministeps Rasch-model computer programs.* Chicago, IL: Winsteps.

Lovett, M. (2001). A collaborative convergence on studying reasoning processes: A case study in statistics. In D. Klahr & S. Carver (Eds.). *Cognitive and instruction*: *Twenty-five years of progress* (pp. 347-384). Mahwah, NJ: Lawrence Erlbaum.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data . *The International Journal of Educational and Psychological Assessment*, 1(1), 1-11.

Matthews, D., & Clark, J. (2007). *Successful students' conceptions of mean, standard deviation and the central limit theorem*. Retrieved on 6 April, 2012, from http://www1.hollins.edu/faculty/clarkjm/stats1.pdf

Olani, A., Hoekstra, R., Harskamp, E., & van der Werf, G. (2011). Statistical reasoning ability, self-efficacy, and value beliefs in a reform based university statistics course. *Electronic Journal of Research in Educational Psychology*, 9(1), 49-72.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. London: Routledge & Kegan Paul.

Sharma, S.V. (2005). High school students interpreting tables and graphs: Implications for research. *International Journal of Science and Mathematics Education*, 4, 241-268.

Tinsky, H.E.A., & Dawis, R.V. (1977). Test-free person measurement with Rasch simple logistic model. *Applied Psychological Measurement*, 1(4), 483-487.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14 (2), 97 – 116.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis.* Chicago: MESA Press.