Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 7(C) (2010) 205–209

Procedia Social and Behavioral Sciences

International Conference on Learner Diversity 2010

Validity of UKM1 Intelligence Test using Rasch Analysis

Siti Rahayah Ariffin[a,]*, Suriana Mohd Asari[a], Shafiza Mohamed[a], Shah Nazim Shahar[a], Noriah Mohd Ishak[a], Rosseni Din[a], Abdul Ghafur Ahmad[a], Hamidah Yamat[a], Rosadah Majid[a], Siti Fatimah Mohd Yasin[a]

[a]*Faculty of Education, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia*

Abstract

UKM1 is a test developed in Malaysia to evaluate the level of intelligence of the students. This study aims at identifying the reliability and validity of the UKM1 using the item analysis method. A random sample of 773 students was selected from the local schools. The test was administered via on-line. The study confirms that UKM1 is a multidimensional test with three correlated dimensions which are Visual Spatial (VS), Verbal Linguistic (VL) and Logical Mathematics (LM).

Keywords: UKM1 intelligence test; Visual spatial; Verbal linguistic; Logical mathematics; Rasch analysis; Fit statistics; Unidimensionality

1. Introduction

Charles Spearman in 1904 introduced the theory of intelligence. According to Charles, intelligence refers to the *g* factor on how one uses his mental capacity to perform daily life activities. The first intelligence test, Binet-Simon scale, was administered in France which aimed to identify the weak students in school. This test was developed by Alfred Binet and Simon Theophile in 1905. The concept of mental age had been introduced by Binet, Lichtenberger, and Kaufman (2009) whereas the IQ formula was introduced by William Stern in 1912 (Roid, 2003). Raven Standard Progressive Matrices scale was introduced by John C. Raven in 1936.

A group of Malaysian researchers had developed an intelligence test called UKM1. It was developed in 2009 and adapted from the Raven Standard Progressive Matrices for individuals ranging from 9 to 15 years old. The test was being used as an intervention method in learning at school. The outcome based learning can be meaningful if teachers can identify the strength and weaknesses of students abilities to increase their knowledge (Siti Rahayah Ariffin et al., 2009).

The purpose of this study is to determine the reliability and validity of UKM1 test items. Analysis of items using the item response theory (IRT) with the application of Rasch model of measurement was performed to check the psychometrics properties of UKM1. A sample of Malaysian students aged 9 years old was chosen in the study. The purpose of assessing the students is to modify the items in UKM1, in order to develop an appropriate item bank for the new version of UKM1.

* Corresponding author. Tel: +6013-320 4848; fax: +603-8925 4372
Email address: sitira@ukm.my

1877-0428 © 2010 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.sbspro.2010.10.029

The quality of UKM1 items is determined based on reliability, separation, unidimensionality, difficulty and fit statistics using Rasch model. Item analysis is a statistical and empirical method which provides information on the items (Doyle, Hula, McNeil, Mikolic, & Matthews, 2005; Wang & Chen, 2005; Forkmann, Boecker, Norra, Eberle, Kircher, Schauerte, Mischke, Westhofen, Gauggel, S. & Wirtz, 2009; Williams, Heinemann, Bode, Wilson, Fann, & Tate, 2009). The study of psychological tests is important in evaluating the quality of the instrument that has been developed. Construct validity is the most important factor to determine the reliability and quality of the test.

Item response theory (IRT) which is is a modern test theory, used to analyze the test items [2]. IRT is a statistical approach based on the item characteristics curve (ICC). ICC is the probability of different capabilities of individuals to answer an item correctly (Kubinger, 2005; Magno, 2009; Meyers, Miller, & Walter, 2009). The Rasch measurement model is an application used in the study of the reliability and validity of UKM1. Results from the item analysis will be used for deleting or modifying the overall test items with the application of IRT (Kubinger, 2005; Magno; 2009; Meyers, et al., 2009). Using the Rasch measurement model, the quality of the test items are based on the reliability and the separation of items and individuals (Doyle et al., 2005; Wang et al., 2005; Forkmann, et al., 2009; Williams et al., 2009) had performed item analysis study using the Rasch measurement model to determine the reliability and validity in the instruments in rehabilitation, education, psychology and health and medical. Reliability is the consistency of the results of assessment test over time. Validity refers to what a test wants to measure or the purpose of the test.

Cronbach alpha (KR 20) of Rasch model is used to test the reliability of item. Logits scale is a representative of the individual ability, who responses to the items in different magnitude of difficulty (Bond et al., 2003). Unidimensionality is a detection of construct validity in the tests that has been developed. Items should test constructs which measure a single dimension only. Local independence will occur when an item has no correlation with another item in the same test. Unidimensionality and local independence are very important to measure the internal consistency of the instrument using the principal component of analysis (PCA). Fit statistics is the criteria of mean square (MNSQ) to identify the information-weighted (Infit) and outlier-sensitive (Outfit). The MNSQ values are ranging from zero to infinity with expected value of 1. Items outside the range value of MNSQ are considered overfit or misfit. Overfit means the items are too predictive while misfit means the items are erratic. The study of UKM1 items analysis was conducted using the WINSTEPS 3.64.2 program. The study aims at analyzing the UKM1 test to identify the reliability and the separation of items, unidimensionality and fit statistics using Rasch model.

2. Literature Review

According to Sommer, Fink, & Neubauer, (2008), intelligence tests should be conducted to determine the level of intelligent of the students. Intelligence tests could measure the mental ability of a person (Noriah Mohd Ishak et al., 2009). Students' achievement can be measured through the Raven Progressive Matrices Intelligence Test in which the potential, ability and intelligence of students are identified and categorized.

Previous studies carried out by researchers showed that the removal of items could improve the degree of validity and reliability of the instrument. This could be done by analyzing the instrument using the Rasch model. According to Gallini (1983) a statistical analysis had been carried out using Rasch model where Raven Progressive Matrices test was taken by 157 seventh grade students. The results were 30 items of 60 items were removed from the test item. According to Kahn, (1985) 9 items were removed from the instrument after the analysis. According to Green & Kluever, (1991) test was administered on the 166 intelligent students. The findings of the analysis found 3 misfit items out of the 36 items and cannot be maintained to measure the intelligence of the students.

3. Methodology

A total number of 773 Malaysian undergraduate students were chosen as respondents using the stratified random sampling method. The sample of male students were 31.80% (n= 246) and female students were 68.20% (n = 547). The data comprised of the Malays 70.50% (n = 545), and Non-Malays 29.5% (n = 248). The sample also comprised 28.10% of science students and 71.90% of non science students.

The test contains 148 items consisting of 3 indexes. The Visual Spatial Index comprises 60 multiple choice symbol search items. The Verbal Linguistics Index comprises 44 open ended vocabulary items. The Logical Mathematics comprises 44 multiple choice vocabulary items. The test was administered for via online, and internet access was setup for the test administration in the computer lab. All students successfully completed the test. The responses of all items are dichotomous data.

The separation and reliability indexes of items and person are shown in Table 1. All items' separation indexes are greater than two. All items' reliability indexes are more than 0.80. However, separation indexes of persons are mostly below two.

Table 1 shows the principal components of analysis (PCA) for the standardized residuals. All of the indexes measured have variances of more than 60.00% and the additional dimension variance is less than 5.00%.

Table 1. Separation and Reliability Indexes, Principal Component Analysis (PCA) and Fit Statistics

Subtest	Separation		Reliability		Measured Dimension (%)	Additional Dimension (%)	Item		
	Item	Person	Item	Person			Fit	Overfit	Misfit
vs	5.58	1.74	0.98	0.97	85.30	0.6	54	0	6
vl	13.11	0.86	0.99	0.43	67.30	1.8	38	0	6
lm	10.34	1.28	0.99	0.62	76.60	1.3	41	0	3

Note: vs-Visual Spatial, vl-Verbal Linguistics, lm-Logical Mathematics

Statistical analysis for suitability of items were carried out to identify items that have values of Infit MNSQ and Outfit MNSQ which should be greater than 0.5 and less than 1.5 (0.5-1.5). First, the fit statistic was performed on the Outfit MNSQ then to the Infit MNSQ statistic. Table 1 shows the numbers of fit (0.5-1.5), overfit (<0.5) and misfit (>1.5) items based on the Infit MNSQ and Outfit MNSQ statistics. Overfit means 50% (1 – 0.5 = 0.5) lack the expected variance, while misfit means 50% (1.5 – 1 = 0.5) is over the expected variance.

There are 133 fit items, 0 overfit item and 15 misfit items from an overall total of 148 items administered in the UKM1 intelligence test. The numbers of fit items range from 54 for Visual Spatial Index to 38 for Verbal Linguistics Index. The numbers of misfit items range from 6 for Visual Spatial and Verbal Linguistics Index to 3 for Logic Mathematics Index. There is no existence of overfit items in the UKM1 test.

4. Findings

The study aims at identifying the validity of UKM1 test using Rasch analysis. Results show the et al. the indexes have high reliability values ranging from 0.98 to 0.99. Each indexes has a good reliability value near to 1 besides the items' separation that range from 5.58 to 13.11. Strata is a statistical method to identify the number of group exist in the test. This indicates that 6 to 13 groups of people exists in the UKM1 test. Both the reliability and separation values showed the al. l of the indexes are internally consistent. Construct validity asserted the effectiveness of the instrument developed overtime (Azrilah Abdul Aziz et al., 2009; Rosseni Din et al., 2009; Siti Rahayah Ariffin & Nor Azaheen Abdul Hamid, 2009).

The separation and number of strata of items mean all of the indexes showed that the items in the test are spreading fairly with individual abilities in logits. The reliability and separation of individuals are not discussed because the values were not linear. Moreover, the value of reliability and separation of individual in UKM1 shows an unusual result. Unidimensionality is an identification of a correct measure of item within its subtest by using the principal component analysis (PCA) (Tan Chue Poh et al., 2007). All the indexes are showing unidimensionality.

In examining the fit statistics of UKM1 test items, the Outfit MNSQ and Infit MNSQ statistics used the range from 0.50 to 1.50 for quality control purposes. A total of 133 items are found to be fit, 0 items are overfit and 15 items are misfit. This indicates that intervention must be done to check on the problems or weaknesses of items. The number of items that highly misfit were from Verbal Spatial, Verbal Linguistic and Logical Mathematics indexes. According to Bond & Fox (2007) low value of Outfit items suggest that the responses were very predictive when the individual's ability level is distant from the item. Low value of infit items suggest that the responses were too predictive for the item when individual's ability level is near to the item. Meanwhile, high value of Outfit items suggest for an unusual response for the item when individual's ability level is distant from the item. High value of Infit items suggest that there was an unusual response in the item when individual of ability level near to the item. Items in Visual Spatial index has high Infit (MNSQ<1.5) when individual with low ability not able to response to the items easy items. Items in Visual Spatial and Verbal Linguistics indexes have high Outfit (MNSQ>1.5) when individuals with high ability were not able to response to the easy items. Furthermore, items that have low Infit (MNSQ<0.5) when individual with low ability able to response to the difficult items. Clearly, if Overfit items

existed in the instrument, the item needs to be revised before developing the new version of UKM1 to avoid having too many overfit items in the test. Limited number of overfit items is advisable. The most highly misfit items which should be revised or modified by experts. Items with good level of psychometric are retained after removing the misfit items. High level of internal consistency of reliability of an instrument is correlated with the size of error of measurement. The Infit and Outfit values in UKM1 show et al. the fit items have measured the indexes correctly. UKM1 should be able to measure low and high ability individuals.

In this study, another concerned matter is how well the individual's response is reliable on online test compared to paper and pencil test. However, Baker (2001) found that there is no bad implication in the context of quality of responses using online method. Online method is proven to have advantages on reducing the time of collecting data and is timely effective (Xu, Iran-Nejad, & Thoma, 2007). Moreover, an evaluation of the UKM1 online test can be performed to increase the effectiveness and usability (Noraidah Sahari, et al, 2009).

5. Conclusion

Items analysis is the best method for determining the quality of developed test items. Besides that, the development of calibrated item bank will produce the substantial measurement. Using Rasch model, the aims of developing the adult version of UKM1 will give specific information to the item bank which is sufficient to improve the quality of the test. Students intelligence test should be able to predict the ability of individuals into the teaching and learning in school. Intelligence tests should measure the *g* factor rather than individual factor, means that single index is not sufficient to measure intelligence but more indexes are good to measure individual cognitive ability. In order to measure the *g* factor, each of the indexes must show its unidimensionality and be correlated to the intelligence test. Future study should identify the confirmatory factor analysis (CFA) of UKM1 to examine the theory of intelligence.

References

Azrilah Abdul Aziz., Azlinah Mohamed., Noor Habibah Arshad., Sohaimi Zakaria, & Mohd Saidfudin Masodi, (2007). Appraisal Of Course Learning Outcomes Using Rasch Measurement: A Case Study In Information Technology Education. International Journal of Systems Applications, Engineering & Development.4, 164-172.

Baker, F. B., (2001). The basics of Item Response Theory, 2nd ed. United States: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland. College Park, MD, http://edres.org/irt. 3 Mac 2010.

Bond, T. G., & Fox, C. M., (2007). Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 2nd ed, New Jersey: Lawrence Erlbaum Associated, London.

Doyle, P. J., Hula, W. D., McNeil, M. R., Mikolic, J. M., & Matthews, C., (2005). An Application of Rasch Measurement of Communicative Functioning. *Journal of Speech, Language, and Hearing Reasearch*. 48, 1412-1428.

Gallini, J.K., (1983). Rasch Analysis of Raven Item Data. *Journal of Experimental Education*. Fall 52, 27-32.

Green, K.E., & Kluever, R.C., 1991. Structural Properties of Raven's Coloured Standard Progressive Matrices For A Sample Of Gifted Children. Perceptual and Motor Skills. 72, 59-64.

Forkmann, T., Boecker, M., Norra, C., Eberle, N., Kircher, T., Schauerte, P., Mischke, K., Westhofen, M., Gauggel, S. & Wirtz, M., (2009). Development of an item bank for the assessment of depression in persons with mental illnesses and physical diseases using Rasch Analysis, *Journal of Rehabilitation Psychology*. 54, 186–197.

Kahn, H.F., (1985). Application of Latent Trait Theory To Raven's Standard Progressive Matrices. Dissertation Abstracts International. 46, 1559-1560.

Kubinger, K. D., (2005). Psychological test calibration using the Rasch model – some critical suggestions on traditional approach, *International Journal of Testing*, 5, 377-394.

Lichtenberger, E. O & Kaufman, S. A., (2009). Essentials of WAIS-IV Assessment, New Jersey: John Wiley and Sons.

Magno, C., (2009). Demonstrating the difference between Classical Test Theory and Item Response Theory using derived test data, *The International Journal of Education and Psychological Assessment*. 1, 1-11.

Meyers, J. L., Miller, G. E. & Walter, D. W., (2009). Item Position And Item Difficulty Change in an IRT-Based Common Item Equating Design, Applied Measurement in Education. 22, 38-60.

Noraidah Sahari, Hairuliza Mohd. Judi, Abdul Azim Abdul Ghani, Hasan Selamat & Aida Suraya Md. Yunus, (2009). Construction and Selection of Usefulness Evaluation Items, *International Journal of Education and Information Technologies*. 1, 28-35.

Noriah Mohd Ishak, Rosadah Majid & Siti Fatimah Mohd Yasin (pnyt.). (2009). Permata Pintar: Pengalaman UKM. Bangi : Pusat Permata Pintar Negara, Universiti Kebangsaan Malaysia.

Roid, G. H., (2003).Stanford-Binet Intelligence Scales, 5th edition, Itasca, IL. Author,

Rosseni Din, Mohamad Shanudin Zakaria, Khairul Anwar Mastor, Norizan Abdul Razak, Mohamed Amin Embi & Siti Rahayah Ariffin., (2009). Meaningful Hybrid e-Training Model via POPEYE Orientation, International *Journal of Education and Information Technologies*. 1, 56-66.

Siti Rahayah Ariffin., Noriah Ishak., Rohaty Majzub., Ramlee Mustapha., Rosadah Majid., Abd. Ghafur Ahmad., Norshidah Salleh., (2004). Manual Instrument Malaysian Adolescent Multiple Intelligences Test (MAMIT). Bangi, Universiti Kebangsaan Malaysia.

Siti Rahayah Ariffin.,(2008. Innovations in Measurement and Evaluation. Bangi: Faculty of Education, Universiti Kebangsaan Malaysia.

Siti Rahayah Ariffin., Roseni Ariffin., Hafsa Mohamed Makki., (2008). Contribution Factors in Multiple Intelligences Among Adolescene Student, *Journal of Education* 32, 35 – 46.

Siti Rahayah Ariffin., Nor Azaheen Abdul Hamid., (2009).Critical Thinking Skills Profile Between the Science and Non-Science Students, *Education Deans' Council Journal*, 3, 1- 22.

Siti Rahayah Ariffin., Rodiah Idris & Noriah Ishak. (2010). Differential Item Functioning in Malaysian Generic Skills Instrument, *Journal of Education* 32, 1-10.

Sommer, U., Fink, A. & Neubauer, A.C. (2008). Detection of high ability by teachers and parents: Psychometric quality of new rating checklists for assessment of intellectual, creative and social ability. Psychology Science Quarterly. 50, 189-205.

Tan Chue Poh, Nur Fateha Muhamad Lani, Lai Weng Kin., (2007). Multi-Dimensional Features Reduction of PCA on SVM Classifier for Imaging Surveillance Application, *International Journal of Systems Applications, Engineering & Development*. 3, 45-50.

Wang, W.C. & Chen, C.T., (2005). Item Parameter Recovery, Standard Error Estimates, and Fit Statistics of The WINSTEPS Program for The Family of Rasch Models, Education Psychology Measure. 65, 376–404.

Williams, R.T., Heinemann, A. W., Bode, R.K., Wilson, C. S., Fann, J.R. & Tate, R. G., (2009). Improving Measurement Properties of the Patient Health Questionnaire–9 with Rating Scale Analysis, Rehabilitation Psychology. 54, 198–203.

Wong, J. G., Cheung, E. P., Chan, K. K., Ma, K. K. & Tang, S. W.,(2006). Web Based Survey of Depression, Anxiety and Stress in First-Years Tertiary Education Student in Hong Kong, Australia New Zealand. *Journal of Psychiatry.* 40, 777-782.

Xu, Y., Iran-Nejad, A. & Thoma, S. J., (2007). Admistering defining issues test online: Do response mode matter?, *Journal of Interactive Online Learning*. 6,10-27.