PalArch's Journal of Archaeology of Egypt / Egyptology

ASSESSING MEASUREMENT GAP USING MODERN TEST THEORY: AN INVESTIGATION IN MALAYSIAN BUSINESS ORGANIZATIONS

Muhammad Shahar Jusoh[1,a)] , Rushami Zien Yusoff [2,b)], Zakaria Abas [2,c)] & Mohd Salleh Hj Din [1,,d)]

[1] Faculty of *Applied* and Human *Sciences*, University Malaysia Perlis (UniMAP), 01000 Kangar, Perlis, Malaysia

[2] College of Business, University Utara Malaysia (UUM), 06010 Sintok, Kedah, Malaysia

Corresponding author:
[a)]shahar@unimap.edu.my
[b)]rzy278@uum.edu.my
[c)]zakaria@uum.edu.my
[d)]sallehdin@unimap.edu.my

Muhammad Shahar Jusoh[1,a)] , Rushami Zien Yusoff [2,b)], Zakaria Abas [2,c)] & Mohd Salleh Hj Din [1,,d)]: Assessing Measurement Gap Using Modern Test Theory: An Investigation in Malaysian Business Organizations-- Palarch's Journal Of Archaeology Of Egypt/Egyptology 17(7). ISSN 1567-214x

Keyword: Construct validity, Rasch analysis, Goodness of measure, Principal Component Analysis

ABSTRACT

Instrument construct is one of the most important issues in conducting research. Without proper consideration in tackling the issue, it is difficult for the instrument to be considered as valid and reliable. If construct validity is accurate, then it will provide a clearer and more precise descriptive analysis on the concepts being investigated. The most important criteria that need to be considered in answering the construct validity are reliability and validity. In traditional measurement model, the understanding of reliability and validity is totally different from the one offer by Rasch model from the context of ordinal data and interval data related issues. Regardless of the difference, this study is giving and sharing the other options of measuring an instrument in business and management natures for measuring the proposed framework. Using this model, local dependence and item fit are most considered in getting valid, reliable, and consistent, hence its significance to measure the construct. Finally the study applies goodness of measure purposely to answer issues related to validity and reliability tests. Principal components analysis was carried out to test the construct of questionnaires used in the study.

Introduction

Reliability relates to ability of a measure to remain the same; consistently over time (Sekaran, 2003) or the same result is obtained when the same research is repeated or does it again once more. In order to get the reliability

of the test, *Cronbach*- α is used as the common value in estimating the internal consistencies of the items (Onwuegbuzie & Danial, 2002). The *Cronbach*-α value should considerably be higher than the acceptance level of 0.60 (Garson, 1998; Gliem & Gliem, 2003; Leedy & Ormrod, 2005) to be taken as reliable. Rasch Analysis further provides the reliability of person and item arising from the interaction between both subjects in the test. If the Person reliability registered 0.76, it is described as 'Fair' reliability (Fisher Jr., 2007; Azrilah, 2010). This indicates that, the Person in this assessment with reliability equal to or greater than 0.6, given another set of questionnaire, they stand a high chance to produce a repeat outcome for the next time (Andrich, 1988). Azrilah (2010) further explained, if selected respondents from the same population were to be given the same quality management principles, cost of quality and organizational performance instrument, the probability of the respondents' ability pattern would be similar.

The purpose of having validity is to make sure the instrument in use is measuring what it is supposedly to measure (Sekaran, 2003; Zickmund *et al.*, 2010). The criterion validity of such findings represents the actual likeliness of the situation. Since most of the questions were adapted from previous studies, the issue of face validity is to assure the meanings of the questions given are measuring the underlying concept (Sekaran, 2003). Though some researchers believed that face validity is not a valid component of content validity, it remains a very important process to determine the suitability of the questions given posted to the respondents. The content validity itself has to be done as it will ensures the questionnaire include the adequate and represent or sufficient and enough items to represent the subject matter or the concept ushered in the study (Sekaran, 2003). This will later be validated by the model hence answering the content validity issue. Similarly with the item reliability provided, it will determine whether the instrument is having sufficient number of questions for all range of respondents.

Realizing the importance's of issues, this study attempt to explore the potential of using other measurement method which can give better results in measuring validity and reliability of an instrument purposely in quality management natures of research. More recent study by Preece (2002), Acton (2003), Schumacker and Smith (2007), Battista *et al.* (2010) challenges against the use of Classical Test Theory in treating the raw scores using the scale formatted as interval data. Their findings proved that the scales formatted are considered as ordinal data and the analyses of variances or product-moment correlations are not permissible unless transformed into interval-scale measures on a ratio scale. The analysis on the issues related to ordinal data and interval data have been criticized by many other authors such as Ganglmair and Lawson (2003), Hambleton and Jones (2005), Pallant *et al.* (2007), Pana *et al.* (2007), Gothwal *et al.* (2009), Casteleijn (2010), who pointed out the needs to revisit the procedures of getting the valid and reliable of an instrument. Therefore, this study attempt to use Rasch measurement model approaches in transforming the ordinal data into interval data through the statistical and mathematical combinations (Salwana *et al.*,2019 and Rohani *et al.*, 2019) so that the study capable to

figure out the uni-dimensionality dispute so that it will give better instrument stability.

Literature Review

The Rasch Measurement Model generates important information about the items whether it is measuring in a single direction or behaving erratically by functioning in the opposite direction. In this case, it can be reported that the instrument is behaving in bi-direction. Rasch's 'zero-set' the instrument when the item is at a threshold point equals to mid-point 0.5 hence a situation of 50:50. Then, it calibrating the rating scale to ascertain the assumed rating is valid for use. If the threshold separation between any two ratings is less than 1.4, then the assumed rating is collapsed and re-calibrated to ensure better Infit Standard Deviation or invariance obtained. This is the crucial test involve as the procedures itself determines the reliability of the respondents and construct validity of the instrument hence valid data (Andrich, 1988; Bond & Fox, 2007; Fisher Jr., 2007; Linacre, 2008).

There are three indicators need to be fulfilled before one can claim that the instrument employed, i.e. the questionnaire as reliable and valid thus replicable and measuring what we are supposedly to measure. The explanation summarized by Azrilah (2010):

a) Cronbach- α value (should be more than 0.6)
b) Item Reliability value – to answer whether the question is valid or not. If the reliability value is > 0.6 ; then the questions asked in the study is sufficient for the expected range of respondents. If the score is less than <0.6, then the number of questions asked is insufficient thus invalid instrument construct.
c) Person Separation value – to show the ability of the instrument to discriminate the respondents into distinct groups. If the instrument cannot separate the respondents as expected, then the items need to be reviewed; either rephrased or new item added.
d) Person Reliability value – gives an indication of the person latent trait measures or psychometry. If the score > 0.6 meaning the person involve in the study is reliable and if the score<0.6 meaning that the person is not reliable. Meaning that if the result shows high reliability, it means that the location of person along the ruler will be the same for the second time if an instrument of the same construct were taken by the same respondents.

Goodness of measure is performed in answering the issues related to validity and reliability tests. Principal Components Analysis (PCA) was carried out to test the uni-dimensionality of the questions used in the study. Basically the purpose in Rasch is to "*perform a Principal Components Analysis of the residuals by item correlation matrix. The first factor reported here is really the second factor, because the Rasch dimension is the first factor overall. This second factor identifies the strongest pattern in local dependency among the items as reflected in their correlations* " (Wright, 1996). However, Sick (2011) views Principal Component Analysis as "*an extension of Rasch fit analysis used to confirm whether the Rasch difficulty dimension; (thus the construct) has adequately accounted all of the non-random variance in the data*".

As the items represent in this study fit the model then it supported the unidimensionality of the scale hence explain the goodness of content validity (Wright, 1996; Sick, 2011). Principal Component Analysis allows the researcher to refine the instrument construct further by conducting the process of elimination to choose which item fits best by looking at the item quality compliance. Hence, it allows only quality items to best describe the variables being used.

Tennant and Pallant (2006) further details the Principal Components Analysis capabilities to support the post-hoc testing, having undertaken the Rasch analysis and supposing fit the Rasch model requirements. Principal Components Analysis constituted of two major steps, started with choosing the item fits best from the Local Item Dependence (LID) requirement and the quality compliance to Item Measure standard. Yen (1993) and Zenisky, *et al.* (2003) suggested using the local item dependence is to detecting the dependency between pairs of items or persons.

Methodology

In order to select the relevant items to represent the construct, a criterion for local dependence loading had been discussed by few authors. Yen (1984) and Yen (1993) suggested a small positive adjustment to the correlation of size 1/(L-1) where L is the test length. Local dependence would be large positive correlation, with highly locally dependent items (Correlation > 0.7) suggesting that only one of the two items is needed for measurement based Table Principal Component Analysis : Largest Standardized Residual Correlations in Winsteps 2006 (Wright *et al.*, 2000).

Further analysis in fulfilling the Rasch criterion through misfit analysis will show the items that are inconsistent with the construction of a single measure. The first step involve in this process is sharing more than half of their random variance (Wright, 1999). As suggested by Wright (1999), the only one of these two items is needed for measurement thus one of the items has to be discarded or removed. Item from the same domain which exhibit measure of MNSQ further away than 1 and the z-Std further than '0' shall be dropped. Further justification by Tennant and Pallant (2006) suggested, the analysis of the residuals was conducted in detecting the second factors after the Rasch factor is removed due to understanding that, "*originally interpretation of this was difficult as the proportion of variance attribute to the first residual factor was reported but the total variation in the data was unknown*".

The magnitude data of the first residual factor using the Rasch factor can be determined easily as compared to other measurement model (Tennant & Pallant,2006; Sick, 2011; Wright, 1999). Reckase (1979) accept raw variance explained by measures greater than 20% but Rasch requires 40% as an indicator of uni-dimensionality. Generally as shown in table 1, it can be considered since the modeled variance is 32.4% and the unexplained variance is quite noisy at 8.1% nearing the limit of 15%. Thus, further test need to be done to improve the uni-dimensionality of the instrument.

The next steps are by examining the redundancy or possible multi-collinearity through item pairs. It is important to note the local dependency specifies that the value of one data has no influence on another once the underlying data has been accounted (Wright, 1996). As such, data from this

study were fitted to the model and tested for appropriate category ordering if local dependency and principal component analysis were done since this two methods are equally answered the uni-dimensionality and multicollinearity methods in classical test theory (Yen, 1993; Zenisky *et al*., 2003; Salzberger & Sinkovics, 2006; Pallant *et al*., 2006).

Further investigation known as fit statistic being tested to further evaluate person's responses to test items to the model" (Boone & Scantlebury, 2005). Bond and Fox (2007) reasoning of having fit statistics from the technical explanation as, "*the use of chi-square fit statistic to determine how well any set of empirical data met the requirements of this model*". In addition Wright (1977) and Linacre (2002), recommended similar steps toward detecting the dissimilarity among items. It is important to identify how respondent's pattern accurately or predictably fit the model by converting the mean-square statistics to the normally-distributed z-standardized. To abridge the findings, Azrilah (2010) recommended four (4) criteria as to check for any outliers or misfits data, as any misfits pattern to be considered are focused on the requirements given, that are:

(1) Point Measure Correlation:
(Pt-Mea Corr); 0.4 < PT-Mea Corr value < 0.85.
(2) Point Measure Correlation (PT-Mea Corr); gave a negative value (meaning that the person is predicted a misfit due to careless respond or guessing).
(3) Outfit Mean Square (MNSQ): 0.5 < Outfit MNSQ value < 1.5
(4) Outfit Z-Standard (Z-STD); -2 < Outfit Z-Std value <+2
If the items under investigation do not meet the above criteria hence, the items can be discarded due to poor quality fit.

Conclusion

Generally, the findings from the Rasch measurement model in answering the reliability, validity and the level of significance are totally different from the classical test theory (CTT) approach. In CTT, it is generally focused on items (or test) rather than persons or items independently (Bell, 1982). However, the Rasch model is concerned about what people do in testing or estimating the person's ability. This supports Andrich's (1982) argument on the shortfalls of CTT computation for Cronbach-alpha based on raw score as compared to Rasch analysis using the probabilistic model. Additionally, the observed variance among the person parameter estimates and the result thereof are used to construct measurement scales which yield a better value of reliability as compared to Cronbach alpha (KR-20). With the implementation of Rasch's principal component analysis a total of 101 unfit items are discarded from the original 334 items. Further checking on the content validity shows that the person reliability and item reliability increase as compared to before the elimination is made.

REFERENCES

Acton, G.S. (2003). What is good about Rasch measurement?. *Rasch Measurement Transactions*, 16, 902-903.

Andrich, D. (1982). An Index of Person Separation in latent Trait Theory, the traditional KR-20 Index, and the Guttmen Scale Response Pattern, *Education Research and Perspectives*, 9 (1), 95-104.

Andrich, D. (1988). Rasch Models for Measurement, Newbury Park, CA:SAGE.

Ary, D., Jacobs, L.C. & Razavieh, A. (2002). Introduction for research in education. Sixth Edition. Wadsworth Thomson Learning.

Azrilah, A.A. (2010). Rasch Model Fundamentals: Scale Construct and Measurement structure, Perpustakaan Negara Malaysia, Kuala Lumpur.

Battisti, F.D., Nicolini, G. and Salini, S. (2010). The Rasch Model in customer satisfaction survey data. *Quality Technology & Quantitative Management*, 7(1), 15-34.

Bell, R.C. (1982). Person Fit and Person Reliability, *Education Research and Perspectives*, 9 (1), 105-113.

Bond, T.G. & Fox, C.M. (2007). Applying the Rasch Model : Fundamental measurement in the Human Sciences. Second Edition

Boone, W.J. & Scantlebury, K. (2005). *The role of Rasch Analysis when conducting science education research utilizing multiple-choice tests*. Wiley InterScience, www.interscience.wiley.com, DOI 10.1002/sce.20106.

Casteleijn, J.M.F (2010). *Chapter 3: Measurement Theory*. South Africa: University of Pretoria.

Fisher Jr., W.P. (2000). Theory, Instrumentation and Data, *Rasch Measurement Transactions*, 14 (3), 760.

Fisher, W.P., Jr. (2007). Rating Scale Instrument Criteria Quality Criteria. *Rasch Measurement Transactions*, 21 (1), 1095.

Fisher Jr., W.P. (2008). The Cash Value of Reliability, *Rasch Measurement Transactions*, 22 (1), 1158-1160.

Fisher Jr., W.P. (2010). The standard Model in the history of the Natural sciences, Econometrics, and the social sciences, *Journal of Physics: Conference Series*, 238 (2010), 012016.

Fisher Jr., W.P., Elbaum, B. & Coulter, A. (2010). Reliability, Precision, and Measurement in the Context of data from Ability Tests, Surveys, and Assessments, *Journal of Physics: Conference Series*, 238 (2010), 012036.

Ganglmair, A. & Lawson, R. (2003). Measuring affective response to consumption using Rasch Modeling. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, 16, 198-210.

Garson, G.D. 1998. *Reliability analysis*. Available at http://faculty.chass.ncsu.edu/garson/PA765/reliab.htm.

Gliem, J. A., & Gliem, R. R. (2003). *Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales.* Paper presented at the Midwest Research to Practice Conference in Adult, Continuing, and Community Education, Columbus, OH.

Gothwal,V.K., Wright, T.A., Lamoureux, E.L. & Pesudovs, K. (2009). Validity of the adaptation to age-related vision loss scale in an

Australian Cataract population, *Journal Optom*, 2 (3) July-September, 142-147.

Hambleton, R.K. & Jones, R.W. (1993 republished 2005). Comparison of Classical Test Theory and Item Response Theory and their applications to test development. *Educational Measurement : Issues and Practice*, 12 (3), 38-47.

Leedy, P. D., & Ormrod, J. E. (2005). *Practical research: Planning and design* (8th ed.). New Jersey: Prentice Hall.

Linacre, J.M. & Wright, B.D. (1994). (Dichotomous Mean-Square) Chi-square Fit Statistics, *Rasch Measurement Transactions*, 8 (2), 360-363.

Linacre, J.M. (1994). Sample Size and Item Calibration (or Person Measure) Stability, *Rasch Measurement Transactions*, 7 (4), 328.

Linacre, J.M. (1996). True-score Reliability or Rasch Statistical Validity?, *Rasch Measurement Transactions*, 9 (4), 455.

Linacre, J.M. (1997). KR-20 or Rasch Reliability: Which Tells the "Truth", *Rasch Measurement Transactions*, 11 (3).

Linacre, J.M. (2001). Generalizability Theory and Rasch Measurement, *Rasch Measurement Transactions*, 15 (1), 806-807.

Linacre, J.M. (2002). What do Infit and Outfit, Mean-Square and Standardized Mean?, *Rasch Measurement Transactions*, 16 (2), 878-879.

Linacre, J.M. (2003). Rasch Power Analysis: Size vs. Significance Chi-Square Fit Statistic, *Rasch Measurement Transactions*, 17 (1), 918-919.

Linacre, J.M. (2006). *A User's Guide to WINSTEPS® MINISTEP Rasch-Model Computer Programs – Program Manual 3.68.0*. www.winsteps.com.

Linacre, J.M. (2007). Standard Errors and Reliabilities: Rasch and Raw Score, *Rasch Measurement Transactions*, 20 (4), 1086.

Linacre, J.M. (2008). The Expected Value of a point-Biserial (or Similar) Correlation, *Rasch Measurement Transactions*, 22 (1), 1154.

Onwuegbuzie, A.J. & Daniel, L. (2002). Uses and misuses of the correlation coefficient. *Research in the Schools*, 9, 73-90.

Pagani, L. & Zanarotti, M.C. (2010). Some Uses of Rasch Models Parameters in Customer Satisfaction Data Analysis. *Quality Technology & Quantitative Management*, 7 (1), 83-95, 2010.

Pallant, J.F., Miller, R.L. & Tennant, A. (2006). Evaluation of the Edinburgh Post Natal Depression Scale using Rasch Analysis. *BMC Psychiatry*, 28 (6).

Pana, A., Chung, L., Fife, B.L. & Hsiung, P. (2007). Evaluation of the psychometrics of the Social Impact Scale: A measure of stigmatization. *International Journal of Rehabilitation Research*, 30, 235–238.

Preece, P. F. W. (2002). Equal-interval measurement: the foundation of quantitative educational research. *Research Papers in Education*, 17(4), 363-372.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and
implications. *Journal of Educational Statistics, 4*, 207-230.

Rohani Mohd, Badrul Hisham Kamaruddin, Anizah Zainuddin, Azimah Daud, Rozita Naina Mohamad (2019). Halal Logo and Loyalty of Muslim Consumers: Reflection for Kopitiam Owners. *Malaysian Journal of Consumer and Family Economics*, 22, 66-80.

Salwana Hassan, Rohani Mohd, Geetha Subramaniam & Badrul Hisham Kamaruddin (2019). Smart Partnership and women Micro-Entrepreneurs in Tanjung Karang-A RAsch Model Analysis. *Malaysian Journal of Consumer and Family Economics*, 22 (S2), 203-219.

Salzberger, T. & Sinkovics, R.R. (2006). Reconsidering the problem of data equivalence in international marketing research – Contrasting approaches based on CFA and the Rasch model for measurement. *International Marketing Review*, 23 (4), 390-417.

Schumacker, R.E. & Smith Jr., E.V. (2007). Reliability – A Rasch perspective. *Educational and Psychological Measurement*, 67 (3), 394-409.

Sekaran, U. (2003). Research Methods For Business: A Skill Building Approach. Fourth Edition. John Wiley & Sons, Inc.

Sick, J. (2011). Rasch Measurement and Factor Analysis, *SHIKEN : JALT Testing & Evaluation SIG Newsletter*, 15 (1) March 2011, 15-17.

Tennant, A. & Pallant, J.F. (2006). Unidimensionality matters (A tale of two Smiths?). *Rasch Measurement Transaction*, 20 (1) Summer, 1048-1051.

Yen, W.M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model, *Applied Psychological Measurement*, 8, 125-145.

Yen, W.M. (1993). Scaling Performance Strategies for Managing Local Item Dependence, *Journal of Educational Measurement*, 30 (3), Performance Assessment (Autumn, 1993), 187-213.

Wright, B.D. (1977). Solving measurement problems with the Rasch Model. *Journal Of Educational Measurement*, 14 (2), 97-116.

Wright, B.D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions*, 10 (3), 509-511.

Wright, B.D. (1997). *Measurement for social science and education - A history of social science measurement*. http://www.rasch.org/memo62.pdf.

Wright, B.D. (1999). Common sense for measurement. *Rasch Measurement Transactions*, 13 (3), 704.

Wright, B.D., Perkins, K. & Dorsey, J.K. (2000). Rasch measurement instead of regression. *Multiple Linear Regression Viewpoints*, 26 (2), 36-41.

Zenisky, A. L., Hambelton, R. K. & Sireci, S. G. (2003). Effects of local item dependence on the validity of IRT item, test and ability statistics. *Association of American Medical Colleges (AAMC).* http://www.aamc.org/students/mcat/research/monograph5.pdf.

Zikmund W., Babin, B., Carr. J. & Griffin, M. (2010). *Business research methods* (8th ed.). South-Western: Cengage Learning.