Interpersonal Citizenship Motivation: A Rating Scale Validity of Rasch Model Measurement

Shereen Noranee, Noormala Amir Ishak, Raja Munirah Raja Mustapha, Rozilah Abdul Aziz, and Rohana Mat Som

***Abstract*—This paper investigated empirically the utility of rating scales in the development of high-quality measures. The present study utilized the Rasch rating scale model to examine and reassess the psychometric properties of the adapted Interpersonal Citizenship Motivation scales. It was argued that interpersonal citizenship behaviour (ICB) could be driven by prosocial motivation and impression management (IMM) motivation. ICB is commonly viewed as prosocial behavior, not impression management. However, a process by which employees are either motivated intrinsically or extrinsically, plays an important role in interpersonal behavior. As ICB was claimed to be more overt and visible than organizational citizenship behavior, the mediation of leader-member exchange quality may increase the likelihood of being more friendly and offering help among IM-oriented employees. The observed measures of the pilot study suggested a violation pattern, indicating a lack of monotonicity in the average measures of Interpersonal Citizenship Motivation scale. Hence, collapsing categories was recommended to create a more uniform frequency distribution. This paper aims to validate the rating scale that yields the highest quality measures for the construct of interest. The WINSTEPS software provides a wide variety of output formats that are virtually indispensable for investigating rating scale quality.**

***Index Terms*—Rating scale validity, Rasch measurement model, interpersonal citizenship behavior, employee motivation.**

I. Introduction

Rasch rating scale model does not presume the size of the step necessary to move across each threshold. It detects threshold structure of the Likert scale in the data set, and then estimates a single set of threshold values that apply to all the item stems in the scale [1]. The way each rating scale is constructed has a great influence on the quality of data obtained from the scale [2]. As some categorizations of variables yield higher quality measures than other categorizations, rating scales should reflect careful consideration of the question construct and unambiguous responses should be elicited from the scales. Therefore, categories and labels should be determined in rating scales to guarantee the quality of the measures. Empirical testing should be utilized on the rating scales to facilitate measures interpretations [1].

Manuscript received January 5, 2014; revised March 10, 2014.

Shereen Noranee is with the Universiti Teknologi Mara (e-mail: shereen@puncakalam.uitm.edu.my).

II. Pilot Study

A. Respondents

This pilot study examined the mediating effect of leader-member exchange (LMX) quality on employee motivation, which comprised of prosocial motivation (PSM) and impression management motivation (IMM), and the interpersonal citizenship behavior (ICB) relationship among Administrative Officers at public universities in Malaysia.

A pilot test was carried out among 33 paired-respondents at four local universities in Selangor, Malaysia. The respondents were among the Administrative Officers and their supervisors, from Professional and Management Group. Six filled-up questionnaires were dropped because there were problems of not matching and not filling up few pages of questionnaire. Therefore, the total of respondents for the pilot tests was 30 pairs. The purpose of conducting the pilot test is to ensure that the measures and meanings of the constructs were suitable for the whole Malaysian context.

III. Instrument Construct Validation

A. Reliability Indices of Summary Statistics

In this preliminary Rasch analysis, the item-person statistics Table I are derived to identity the general fit of the data to the Rasch model. It is important to see the measure of each item fits for further analysis.

TABLE I: Summary of Fit Statistics for Interpersonal Citizenship Motivation form Rasch Analsyis

(*N*=60, *I*=62)

	Item Person Interaction			
	Item Location		Person Location	
Mean	0.00		1.69	
SD	0.43		0.68	
Infit MNSQ	1.02		1.04	
Infit ZSTD	0.00		-0.30	
Maximum measure	2.35	(SE=0.42)	5.56	(SE=0.60)
Minimum measure	-2.47	(SE=0.23)	-0.34	(SE=0.20)
Reliability Indices				
Separation	2.92	(fair)	4.03	(very good)
Reliability	0.90	(good)	0.94	(very good)
Model error	0.29	(very good)	0.26	(very good)
Cronbach's Alpha	0.95 (excellent)			

A total of 2,425 data points arising from 60 respondents on 81 items was analyzed using WINSTEPS 3.72.3, a Rasch analysis software. Items difficulty and person measure

DOI: 10.7763/IJSSH.2015.V5.491

locations, or person ability, are expressed in logits through the transformation of the raw score percentage into its success-to-failure ratio or odds which was then converted to its natural log. The Rasch analysis provides indicators or statistics of how well the items fit within the underlying construct. The results yield a Chi-Square value of 4462.98 with 2282 degree of freedom. The test raw score Cronbach's alpha registers a reliability of 0.95, which allows for further analysis of the instrument.

The goodness of the survey instrument is described by the precision or errors in the item difficulty estimates and person ability estimates, item fit, person fit, and reliabilities of person and item estimates. Table I shows the summary of items estimates with the mean defaulted at 0 logit. Item reliability is very good at 0.90 on a 0 to 1 scale, similar to interpreting Cronbach's alpha, which is transformed to item separation index of 2.92, indicating a fair of item range [3]. The order of items is replicable across other sample.

The subordinates' and their supervisors' performance or ability estimates (Table I) mean of 1.69 logits indicated that the ability is comparatively easy. The maximum item measure is 2.35 logits (SE=0.42), while person ability is high which is at 5.56 logits (SE=0.60). Despite the good reliability, more difficult items, however, recommended to be introduced for that large gap of 3.21 logits. Nevertheless, there is a sufficient item for the easy task where the minimum item measure is at -2.47 logits as against the person of -0.34 logits.

The instrument has a small measurement model error of 0.29 and capable of yielding a good person separation of 4.06but the the person Infit MNSQ SD=0.68 is high. Both item and person Infit MNSQ and Z-STD values are close to the ideal 1 and 0, giving an indication of the goodness of fit of the instrument measuring what is to be measured.

IV. Quality Control

A. Unidimensionality

Then, the dimensionality which is crucial in determining construct validity is identified by using Principal Component Analysis (PCA; Table II).

TABLE II: Description Residual Variance (In Eigenvalues Units)

Description	Empirical (%)	Modeled
Variance explained by measures	45.60%	46.90%
Unexplained variance	54.40%	53.10%
Unexplained variance in contrast	5.6	

Note: The item means were constrained to zero by the measurement model; When the data fit the measurement model, the fit statistics approximates a distributuion with an MNSQ near 1, and a ZSTD near 0 (a good fit for both items and person measures of this scale); The person separation = 4.03 meaning that there was a very separation of measures along the scale, compared with the errors of measurement, which were comparatively smaller (.26). This also implied that the power of the tests of fit t the model was excellent.

To satisfy unidimensionality, the items in the instrument must measure the same composite of abilities; the performance of employee interpersonal citizenship motivation. As indicated in Table II, the principal component analysis (PCA) of the residuals in Rasch shows the raw variance explained by measures of 45.6%. This meets the minimum threshold of 40% as sought by Conrad, Conrad, Dennis, Riley, and Funk [4]. In addition, the unexplained variance in the first factor of 5.6%, rated the instrument as good [3].

The multicollinearity problem could also be determined apart from the production of the largest Standardized Residual Correlation Table where the items that were highly correlated and redundant were rephrased or be deleted. Local dependence test for the largest standardized residual correlation yields that out of 81 items, 9 items are suggested to be deleted, which breached the 0.7 limit [5] which have multicollinearity problems.

B. Fit Statistics

Item fit: Generally, the items in the survey have positive Point Measure Correlation and a small measurement error, with mean of SE=0.29. An item is a misfit when it has a larger MNSQ than the sum of the mean of MNSQ and SD; in this case is 1.45 for infit MNSQ and 1.48 for outfit MNSQ. Items that are misfits with both bold infit MNSQ > 1.45 and Z-Std > +/-2, and bold outfit MNSQ >1.48 and Z-Std > +/-2 were removed. Items that within the accepted range were remained. PT-Measure Correlation value for the items were positive can be remained [6]. Therefore, items left for actual survey are 80 items.

V. Rating Scale Validity

A. Reliability of Measures

A considerable amount of research literature examines the question of how to determine the appropriate number of rating scale categories. Typically, the criterion for judging the optimal number has been the reliability of the responses. Results have shown mixed conclusions, with the following range of assertions about reliability: that it is independent of number of response categories of various number of Likert-point scales.

The research instruments used in this study were adopted from various sources based on the suitability. ICB measure was adopted from Williams and Anderson [7] with reliability of .88, consisting of 7 items and also from Settoon and Mossholder [8] with reliability of .70, consisting of ten items. PSM measure was adopted from Grant [9] with reliability of 0.90, which consists of 4 items; other 5 items from Grant and Summanth [10] with reliability of 0.96, which consists of 5 items, and finally, 3 items from Grant [11] with reliability of 0.91. IMM measure was adopted from Rioux and Penner [12], which consists of ten items with reliability of 0.89, and another 4 items from Connell [13] with reliability of 0.89. Last but not least, the LMX quality measure was adopted from the 7-item construct of Scandura and Graen [14] with reliability of .84, and additional 12 items were adopted from Bernerth, Armenakis, Field, Giles, and Walker [15] with reliability of 0.9.

B. Rating Scale Diagnostics

An examination of Rasch measurement diagnostics managed to determine the optimal number of response

categories. It assesses the functioning categories to create an interpretable measure. The rating scale diagnostics yield the reliability of data for persons and items, sufficiency of the model-fitted categories, indication of hierarchical pattern thresholds of the rating scale, and sufficiency of stable estimation for each category. A general remedy is suggested if problems are diagnosed in the existing rating scale. The remedy is to reduce the number of response options by collapsing categories that cause the problems. After adjacent and better-functioning categories are suggested, reanalyzing the data should be executed to see variable definition improvement. Hence the gold is to produce the rating scale that yields the highest quality measures for the construct of interest. The WINSTEPS Version 3.72.3 software provides a wide variety of output formats that are virtually indispensable for investigating rating scale quality.

Scale calibration is crucial in any measurement. The validity of the scale ultimately affects measurement precision due directly to thinner spread of responses across response categories. Enough data in each category are required for stable estimates. Usually, collapsing problematic categories with adjacent more functioning categories is done to improve variable definition or clarification of the data. The verification process in Rasch is linked to the threshold values between each rating scale.

Average measures which increase monotonically, indicating that on average, those with higher ability endorse the higher categories, whereas those with lower ability endorse the lower categories. However, the observed average measures of the pilot study in Table III, suggested a violation pattern, indicating a lack of monotonicity in the average measures.

TABLE III: DIAGNOSTICS FOR 12345

Category Label	Observed Count	Average Measure	Infit Mean Square	Outfit Mean Square	Threshold
1	28	0.27	1.83	2.46	None
2	116	-0.03	1.10	1.13	-1.87
3	518	0.47	0.91	0.90	-1.27
4	1080	1.50	0.91	0.92	0.29
5	653	3.43	0.91	0.92	2.84

C. Response Category Curves

Finally, analyzing using Response Category Curves (Fig. 1-Fig. 4) highlight the probability of answering each response category by Interpersonal Citizenship Motivation (ICM) measure.

Fig. 1 shows that the probability of answering each response category by Interpersonal Citizenship Motivation (ICM) measure. The curves show that the respondents could not discriminate consistently between two categories ("strongly disagree" and "disagree"). Hence, collapsing categories was recommended. Several attempts in collapsing the categories were executed, namely, categorization 11345 (Fig. 2), 13345 (Fig. 3), and 11234 (Fig. 4).

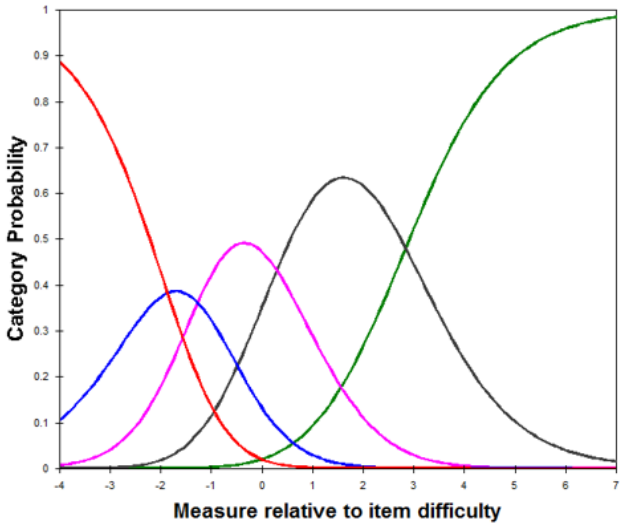

Fig. 1. Response category curve of 12345 categorization.

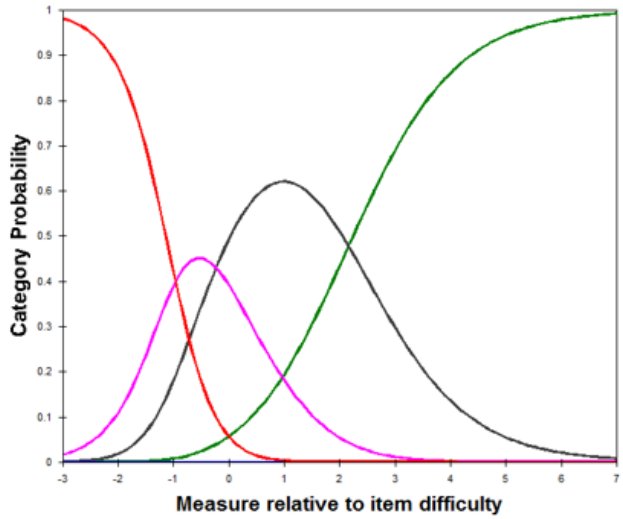

Fig. 2. Response category curve of 11345 categorization.

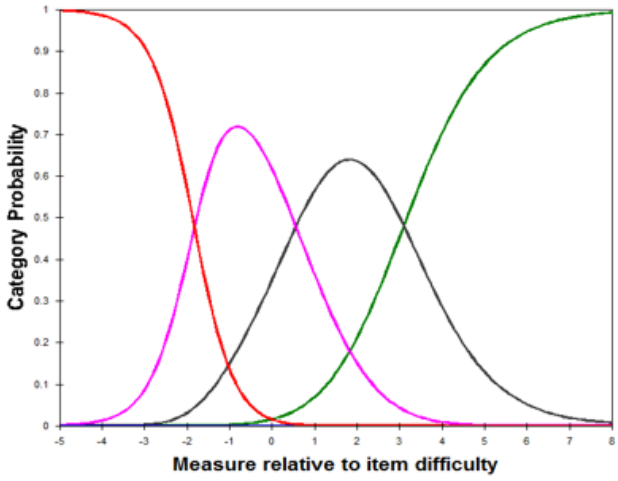

Fig. 3. Response category curve of 13345 categorization.

D. Comparison of Categorizations

As a guide, collapsing the data should create a more uniform frequency distribution. The Rasch-Andrich Threshold is where the transition of decision making occurs from one scale to another. This is captured in the structure calibration column where Linacre recommends that the

difference in threshold should be 1.4 logits apart but not exceeding 5 logits [1]. If the separation is less than 1.4, then it is recommended to collapse the affected rating into one and split the rating scale if it is more than 5.

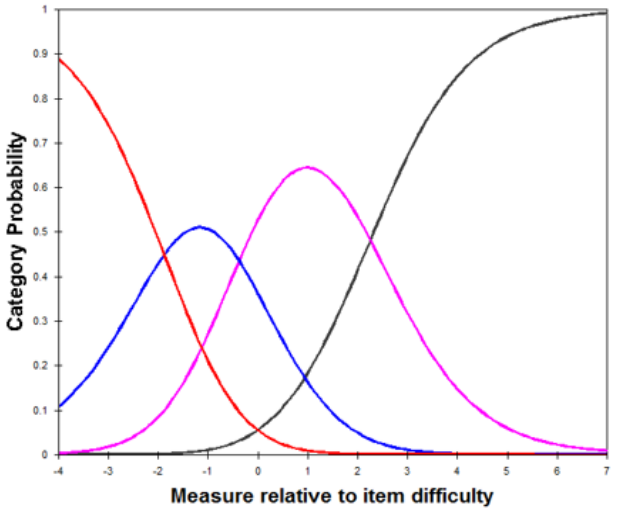

Fig. 4. Response category curve of 11234 categorization.

Several categories for comparison purposes were displayed in Table IV.

TABLE IV: COMPARISON OF FOUR CATEGORIZATIONS

Categorization	Average Measures	Fit	Step Calibrations	Person Separation	Item Separation	Infit MNSQ
12345	Disordered	> 2.0	Ordered	4.09	2.96	0.43
11345	Ordered	< 2.0	Ordered	4.09	2.92	0.45
13345	Disordered	> 2.0	Ordered	4.01	2.87	0.47
11234	Ordered	< 2.0	Ordered	4.18	3.10	0.36

It was found that the categorization of 11345 was not recommended to be used because item separation (2.92) is smaller and infit MNSQ is larger (0.45) than categorization of 12345. Categorization of 13345 also was not suitable to be used, as the same case as categorization of 11345, which yield smaller person and item separation (4.01, 2.87) with larger infit MNSQ (0.47). Categorization of 11234 is the most suitable one to be used as the item separation (3.10) is larger and infit MNSQ is smaller (0.36). Hence, categorization 11234 is the best categorization of rating scale to be used for the actual study.

VI. THE DISCUSSION FOR THE PILOT STUDY

A better fit instrument is constructed where the following parameters are met. Marked improvements are tabulated and analysed across the board on the fit statistics including the MNSQ, Z-STD, reliability, S.E. and variance measured. Values in Table V are the cleaned values with 80 items, followed by italic in bracket which are the original Interpersonal Citizenship Motivation (ICM) 81 items.

The new cleaned instrument has smaller person standard deviation and smaller person mean error. The PCA of explained variance also improved by 45.9% slightly higher than ICM-81. The new ICM-80 instrument shows that the items are sufficient and have very good items to cover the expected range of difficulty, hence validity. The controlling factor would be the measurement S.E.; the measurement precision.

TABLE V: CLEANED ICM-80 INSTRUMENT CONSTRUCT PROPERTIES

	Item	Person
Reliability	0.90 (0.90)	0.94 (0.94)
Separation	2.96 (2.92)	4.09 (4.06)
Infit MNSQ SD	0.43 (0.43)	0.69 (0.68)
Mean Error	0.29 (0.29)	0.26 (0.26)
PCA Variance Measured	45.9% (45.6%)	
Unexplained 1st Contrast	5.7 (5.6%)	

VII. CONCLUSION

A key point to emerge from Rasch analyses of item response data was that, measurement error was revealed and across the measured range. Judgement can be made about the interpretations and decisions (fitness for purpose) base on the data and the measurement error inherent in them [1]. The mean measurement error is influenced by how well or poorly targeted the instrument is to the test sample. The small error detected in the pilot survey for both item and person shows that the instrument has the capability to test the targeted sample.

The study also demonstrated specifically how the design of rating scales has a large impact on the quality of the responses elicited, and to show how the Rasch model provides an appropriate framework for carrying out such investigations.

REFERENCES

[1] T. G. Bond and C. M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, Psychology Press, 2013.
[2] H. H. Clark and M. F. Schober, "Asking questions and influencing answers," *Questions about questions*, 1992, pp. 15-48.
[3] W. Fisher, "Rating scale instrument quality criteria," *Rasch Measurement Transactions*, vol. 21, pp. 1095, 2007.
[4] K. Conrad, K. Conrad, M. Dennis, B. Riley, and R. Funk, "Validation of the Substance Problem Scale (SPS) to the Rasch Measurement Model, " GAIN Methods Report 1.1, ed. Chicago, IL: Chestnut Health Systems, 2011.
[5] S. Ferketich, "Focus on psychometrics: Aspects of item analysis," *Research in Nursing and Health*, vol. 14, pp. 165-168, 1991.
[6] A. A. Aziz, "Rasch model fundamentals: scale construct and measurement structure," *Kuala Lumpur: Perpustakaan Negara Malaysia*, 2010.
[7] L. J. Williams and S. E. Anderson, "Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors," *Journal of Management*, vol. 17, pp. 601-617, 1991.
[8] R. P. Settoon and K. W. Mossholder, "Relationship quality and relationship context as antecedents of person- and task-focused interpersonal citizenship behavior," *Journal of Applied Psychology*, vol. 87, pp. 255-267, Apr 2002.
[9] A. M. Grant, "Does intrinsic motivation fuel the prosocial fire? Motivational synergy in predicting persistence, performance, and productivity," *Journal of Applied Psychology*, vol. 93, pp. 48, 2008.
[10] A. M. Grant and J. Sumanth, "Mission possible? The performance of prosocially motivated employees depends on manager trustworthiness," *Journal of Applied Psychology*, vol. 94, pp. 927-944, 2009.

[11] A. M. Grant, "The significance of task significance: Job performance effects, relational mechanisms, and boundary conditions," *Journal of Applied Psychology*, vol. 93, pp. 108, 2008.
[12] S. M. Rioux and L. A. Penner, "The causes of organizational citizenship behavior: A motivational analysis," *Journal of Applied Psychology*, vol. 86, pp. 1306-1314, 2001.
[13] P. W. Connell, "Transformational leadersip, leader-member exchagne and organizational citizenship behavior: The role of motives," 2005.
[14] T. A. Scandura and G. B. Graen, "Mediating effects of initial leader-member exchange status on the effects of a leadership intervention," *Journal of Applied Psychology*, vol. 69, pp. 428-436, 1984.
[15] J. Bernerth, A. Armenakis, H. Feild, W. Giles, and H. Walker, "Leader-member social exchange (LMSX): Development and validation of a scale," *Journal of Organizational Behavior*, vol. 28, pp. 979-1003, 2007.

Shereen Noranee is a senior lecturer/Ph.D. candidate, Universiti Teknologi Mara Malaysia, Puncak Alam, Selangor, Malaysia. She holds a MSc. in human resource development from the Universiti Putra Malaysia, Malaysia and BSc. in business education from University of Nebraska-Lincoln, USA. Her research interest includes human resource management, organizational psychology and organizational behavior.

Noormala Amir Ishak is a professor at the Universiti Teknologi Mara (UiTM), Malaysia and currently is a professor of the faculty of Business Management. She is specialized in the area of human resource management. Her current research areas relate to organizational citizenship behavior, organizational justice, individual innovativeness, leader member exchange, emotional intelligence, human resource practices, self-esteem, and organizational learning.

Raja Munirah Raja Mustapha holds a PhD in Workforce Education and currently heads the Research Department, Institute of Graduate Studies, Universiti Teknologi Mara. Her research interests are organizational behavior, change management, technical and workforce education, office systems management and general management.

Rozilah Abdul Aziz was born in Kuala Lumpur, Malaysia. She is a postgraduate student at Universiti Teknologi Mara, Shah Alam and currently conducting a doctoral research in the field of Organizational Communication. Rozilah has obtained her MSc. in Corporate Communication, Universiti Putra Malaysia (UPM) and BBA in Business Administration, Western Michigan University, USA. She is currently a senior lecturer at the Faculty of Business Management in Universiti Teknologi Mara, Shah Alam. Her areas of interest include Organizational Communication, Human Communication, Organizational Behavior and Personality Development.

Rohana Mat Som was born in Kuala Lumpur, Malaysia . She is a postgraduate student at Universiti Teknologi Mara, Shah Alam and currently conducting a doctoral research in the field of Office Systems Management. Rohana has obtained her MSc. in Education (Vocational & Technical Education), Virginia Polytechnic and State University (USA) and BA in Business Education, The University of Toledo (USA). She is a senior lecturer at the faculty of Business Management, Universiti Teknologi Mara, Shah Alam. Her areas of interest includes Healthcare Management, Human Resource Management and Vocational and Technical Education