IRMM
INTERNATIONAL REVIEW OF MANAGEMENT AND MARKETING
EJ EconJournals
ISSN: 2146-4405

International Review of Management and Marketing

ISSN: 2146-4405

available at http: www.econjournals.com

International Review of Management and Marketing, 2016, 6(S7) 11-16.

Special Issue for "International Soft Science Conference (ISSC 2016), 11-13 April 2016, Universiti Utara Malaysia, Malaysia"

Rasch Model Scale Calibration Analysis for Islamic Value

Amal Hayati Ishak[1]*, Muhamad Rahimi Osman[2], Siti Khadijah Abdul Manan[3], Rafeah Saidon[4]

[1]Academy of Contemporary Islamic Studies, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia, [2]Academy of Contemporary Islamic Studies, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia, [3]Academy of Contemporary Islamic Studies, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia, [4]Academy of Contemporary Islamic Studies, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia.*Email: amalhayati@salam.uitm.edu.my

ABSTRACT

Rasch model is a robust technique for questionnaire validation. It offers a variety of analyses on the instrument's validity and reliability, as well as analysis on the rating scale design of a measurement scale. Thus, employing Rasch scale calibration analysis, the primary purpose of this article is to empirically analyze the rating scale categories applied in a scale to assess the application of Islamic values in quality management. In the literature, a plethora of Islamic values applied in the context of quality management have been conceptually elaborated in the literature. However, empirical data on that matter is scarce. Yet, an appropriate instrument could not be located. This article briefly explains the scale development process and initially proposes 60 items and eight dimensions. Applying the Rasch model, this article specifically analyses the appropriateness and effectiveness of the rating scale categories based on several indicators pointed by Rasch model. For that purpose, data from 59 responses was analyzed using Winstep. Based on the results, the initial five point Likert-scale is suggested to be modified.

Keywords: Islamic Values, Rasch Model, Scale Calibration.
JEL Classification: M0

1. INTRODUCTION

The Rasch measurement model hails from the item response theory (IRT) which defines how a scale measures latent variables (DeVellis, 2003; Singh, 2004). IRT is a family of measurement models used to measure latent variables. One of the model is Rasch measurement model. It is a probabilistic model which uses logit as measurement units, obtained by transforming ordinal data into interval data where the data can be mapped into a linear scale (Bond and Fox, 2015). Rasch Model has been increasingly used especially in scale development studies (DeVellis, 2003). Though it has first been initiated in educational studies, IRT has been widely applied in testing and validating instruments in various branches of social sciences, such as Bechtel (1985), Albano (2009) as well as Salzberger and Koller (2013). Bechtel (1985) used Rasch model for a consumer rating scale, Albano (2009) used Rasch for individual happiness scale and while Salzberger and Koller (2013) for their marketing scale.

Scholars including Bond and Fox (2015), Azrilah et al. (2013), Tennant and Conaghan (2007) as well as Singh (2004) have agreed that Rasch provides sufficient parameters for a good measurement with the ability of; (1) Providing a linear scale by transforming scores into probabilistic model using logit as measurement units; (2) transforming ordinal data into interval data which enable further statistical analyses to be performed. Mathematical functions which are used to calculate various analyses require interval data in order to produce unbias and accurate results; (3) providing suggestion for missing data by its probabilistic model. Rasch estimates a person's probable response to an item, by considering the person's ability and the item's difficulty; (4) assessing items' quality by detecting misfit and outliers, which may be evaluated by three measures; the point measure correlation, the infit and outfit mean square (MNSQ) and the Z standard; (5) providing distinct measures for item difficulty and person ability, which may be arranged or ranked according to items difficulty or persons ability.

In similar vein, Rasch model was employed to validate a scale to assess the application of Islamic values in quality management, named the Islamic Quality Management Scale (IQMS). However, this article only reports on the validity of rating scale design via Rasch scale calibration analysis. Prior to that, the gap which necessitates the development of the measurement scale, i.e., the IQMS, is explained. This is followed with the methodology and data analysis, which is conducted using Winstep software version 3.72.3.

2. LITERATURE REVIEW

The mainstream quality management initiated in the West by players of the industrial revolution. Hence, its philosophical foundation has been largely dominated with the western values, which have been criticized as narrowly focused to outputs (Syed, 1996; Naceur, 2005; Muhammad, 1985). Nevertheless, the rise of Japanese after disastrous World War II had been an eye opener to the non-western influence in quality management. This is due to the fact that the Japanese implemented it within the scope of their cultural values (Ishikawa, 1985; Naceur, 2005). In a book written by Ishikawa (1985), entitled "Quality Management the Japanese Way," they were described to favor collectivism in work, loyal and family centered. They also initiated the practice of quality circles as a platform to disseminate knowledge and experience between organizational members (Khaliq and Shamim, 1996).

In similar vein, a series of research have concluded on the importance of values in supporting successful implementation of quality management. For instance, in a research conducted by Baird et al. (2011), they reported on significant positive association between cooperation, outcome orientation and innovativeness with quality management practices. In parallel, Prajogo and McDermott (2011) explained on the influence of cultural values on organizational performance in three measures; product quality, product innovation and process innovation.

However, these empirical studies have narrowly analyzed values based on general framework of Hofstede's cultural dimensions (Baird et al., 2011), competing value framework of O'Reilly (Prajogo and McDermott, 2005; 2011; Gambi et al., 2013), organizational culture profile of Quinn and Rohrbaugh (Denison and Spritzer, 1991) and Detert's framework (Detert et al., 2003; Detert et al., 2000). None have made specific relation to values underpinned in religious sources. However, contemporary scholars, such as Khaliq (1996), Abulhasan and Khaliq (1996), Muhammad (2005), Siti et al. (2010), Sany et al. (2011) as well as Siti and Ilhaamie (2011), have consistently elaborate on quality management from Islamic perspectives. The major similarity in their works is the conceptual elaboration on a list of values embedded in the practice of quality management. However, empirical data on that matter is lacking. Yet, an appropriate instrument could not be located.

Based on the explained gap, this article proposes a scale to assess Islamic values application in quality management context. The scale development procedure will be briefly explained in the proceeding section. This is followed with elaboration on the research methodology and data analysis which reports on the application of Rasch model in investigating the appropriateness and effectiveness of rating scale.

3. MEASUREMENT SCALE

Based on an extensive systematic literature review by Ishak and Osman (2015) this article proposes 60 items vested under eight dimensions as a measurement scale to assess Islamic values application in quality management context. These proposed dimensions and items had followed a systematic procedure of scale development process which involved the pooling of items and followed by a refinement process to filter the redundant and irrelevant items. Then, the selected dimensions and items were endorsed by an expert review panel followed by Fuzzy Delphi analysis. Both expert review and Fuzzy Delphi involved different sets of expert panels.

In this study, the experts are selected based on their qualification and experience from the academia and industry. A total number of nine experts were involved in the process. Later, the items and dimensions were statistically confirmed via Fuzzy Delphi analysis, conducted within 17 experts and all items are accepted. Upon agreement by experts, these items were then tested on actual respondents.

A five point Likert scale was selected for Section A ranging from 1 (no implementation), 2 (very minimal implementation), 3 (minimal implementation), 4 (moderate implementation) and 5 (complete implementation). Such five scaling is the most frequently used scale in surveys (Lozano et al., 2008). On top of that, since five or seven point Likert scale produces similar results (Dawes, 2008), the current study decided to use a five point scale.

4. METHODOLOGY

Rasch model provides several empirical evidences on items and persons fit, reliability, and rating scale compatibility, among others. However, this study only reports on diagnostics of rating scale design, the appropriate remedies, and the effect to the overall scale reliability.

The questionnaire was administered among participants of ISO9001 training conducted by SIRIM. They consist of management representatives, document controller or quality division staff who are directly involved with quality management tasks. Out of 100 questionnaires distributed, 59 were returned with a response rate of 59%. The instrument has 61 items which resulted from a systematic literature review focusing on the topic of Islamic values in quality management context (Ishak and Osman, 2015). The data were then analyzed using Winstep, a software of Rasch measurement model.

Rasch model performs the assessment based on the response of a sample of respondents to a set of measurement scale. In Rasch, each person is categorized based on ability, while items are categorized based on difficulty. The categorization is resulted from the interaction between person ability and item difficulty,

which utilizes log odd values. Rasch transform responses into log odd units based on the probability of success, which depends on the differences between person ability and item difficulty. The log odd units enable the person ability and item difficulty to be mapped in a log ruler. The mapping is based on two assumptions; (1) A more developed (or able) person has greater likelihood of endorsing all items, and (2) easier items have greater likelihood to be endorsed by all respondents. Based on these two assumptions, Rasch model predicts the location of items and persons in a map. Besides that, Rasch also capable of analyzing the effectiveness of rating scale design (Bond and Fox, 2015; Azrilah et al., 2013), which is crux of this article.

5. DATA ANALYSES

Scale calibration in Rasch provides empirical evidence to detect whether the respondents understand and are able to differentiate the scaling labels. Ideally, Linacre (1999) points four indicators to diagnose a problematic rating scale, as summarized in Table 1. Based on the table, any value below 1.4 for the difference of structure calibration between categories is a sign of overlapping or inability of respondents to differentiate the scale categories. Rasch assumes that for a normal response, the lowest scale should be the least being selected, and the number of responses for a scale should be increasing, from the least to the highest scale (Azrilah et al., 2013; Bond and Fox, 2007).

Table 1: Indicators for a well-functioning rating scale

Indicators	Descriptions of a well-functioning rating scale
Observed count	High and stable observed count as low values often indicate unnecessary or redundant categories
Observed average	Expected to increase in size as the variable (the category) increases
Structure calibration	Expected to increase in size as the variable (the category) increases Expected difference between threshold is 1.4< × <5
Probability curves	Each category is expected to have distinct peak

Source: Linacre (1999)

The violation of these indices suggests the rating scale to be collapsed, or combined (Linacre 1999). However, Bond and Fox (2007) assert that collapsing categories either upward (for example collapsing Category 4 into Category 5), or downward (for example collapsing Category 4 into Category 3), should only be done if it is sensible based on their labels. For example, it is insensible to collapse agree and disagree, rather than moderately agree and agree. However, scale calibration is only appropriate for pilot test (Bond and Fox, 2007; Azrilah et al., 2013). An effective scale calibration can be detected from increased item reliability and separation (Azrilah et al., 2013).

5.1. The Diagnostics

Rasch provides several indices (Table 1) which empirically detect either the respondents are able to differentiate the scaling labels. The violation of these indices suggests the rating scale to be collapsed (Linacre, 1999). However, Bond and Fox (2015) assert that collapsing categories either upward (for example collapsing Category 4 into Category 5), or downward (for example collapsing Category 4 into Category 3), should only be done on sensible grounds. For example, it is insensible to collapse agree and disagree, rather than moderately agree and agree. In this study, the respondents seemed unable to differentiate Category 2 (very minimal implementation). This had been indicated empirically as shown in Figures 1 and 2.

In Figure 1, the observed count for Category 1 and 2 is much less as compared to other categories. Meanwhile, the structure calibration between Category 2 and 3 was decreasing. The difference was only 0.99 (−1.74 – [−2.73]), which was not between the acceptable range of 1.4< × <5. Linacre (1999) points that the distance between categories should be at least 1.4, but not exceeding 5, to be declared as a well-functioning category. A value below 1.4 is a sign of overlapping between categories and the respondents are unable to differentiate the scales.

The problematic rating scale of Category 2 can also be detected in Figure 2 as its probability curve is redundant and overshadowed with Category 1 and 2. It has no distinct peak as compared to other categories. This is a sign that Category 2 was not well functioning. The respondents were unable to distinctly differentiate it from other categories. It is also a sign that they might not understand it well.

Figure 1: Diagnostics for problematic scaling categories (pre-collapsing)

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE CALIBRATN	CATEGORY MEASURE	
1	1	32	1	-1.51	-1.64	1.06	1.29	NONE	(-3.55)	1
2	2	59	2	-.40	-.61	1.17	1.25	-1.74	-2.28	2
3	3	889	23	.61	.62	1.00	1.03	-2.73	-.78	3
4	4	1919	50	2.14	2.16	.94	.94	.61	2.24	4
5	5	918	24	3.93	3.89	.96	.96	3.86	(4.98)	5

5.2. The Remedy

As explained earlier, the initial rating scale labels are 1 (no implementation), 2 (very minimal implementation), 3 (minimal implementation), 4 (moderate implementation) and 5 (complete implementation). Thus, based on the scale labeling, it is sensible that the respondents might not be able to clearly differentiate between Category 2 and 3.

Furthermore, following the guidelines of Bond and Fox (2007) that collapsing categories should be logical, Category 2 is more logical to be collapsed with Category 3, rather than Category 1. Figure 3 shows the results of collapsing Category 2 into 3, followed by Figure 4.

In Figure 3, the average observed count is consistently increasing. Such increment is referred as monotonic ordering by Bond and Fox (2015) which reflects on the well-functioning of each category. The difference of structure calibration between categories is also above 1.4 and below 5.0. Thus, the results depicted that it is more suitable to use four categories instead of five. Based on extensive review with several respondents, it is suggested that the scaling to be relabeled into 1 = Not implemented, 2 = Slightly implemented, 3 = Moderately implemented and 4 = Highly implemented.

Similarly, Figure 4 shows distinct probability curves for each category, reflecting no redundancies between categories. Therefore, it is more suitable to use four scaling instead of five. The usage of four categories is supported by Lozano et al. (2008), claiming that the ideal number of response category is between four and seven. He also points that a scale is good when the

Figure 2: The probability of curve of rating categories (pre-collapsing)

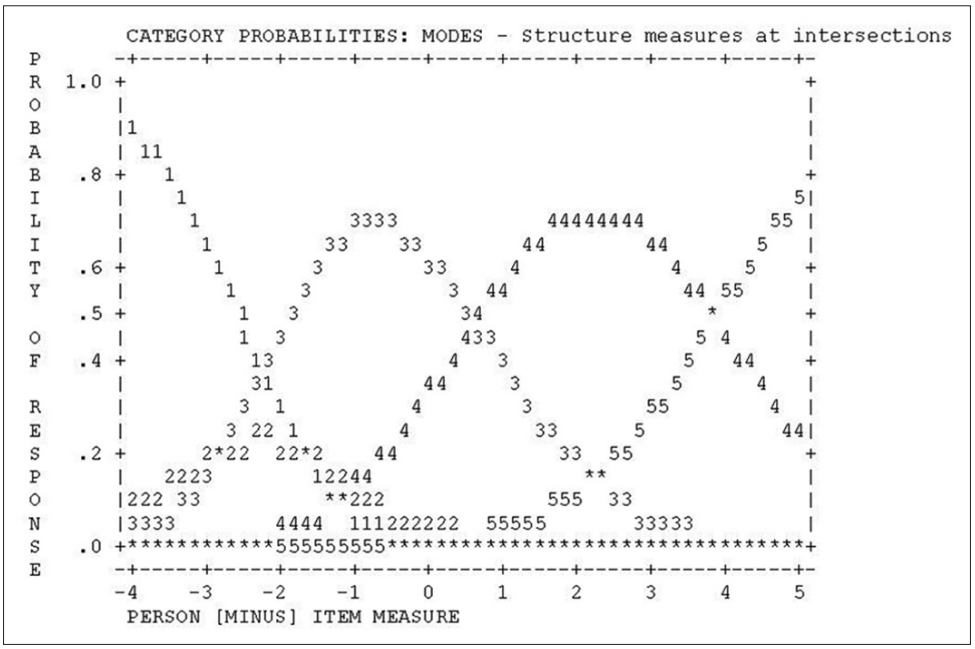

Figure 3: Summary of category structure for post-collapsing

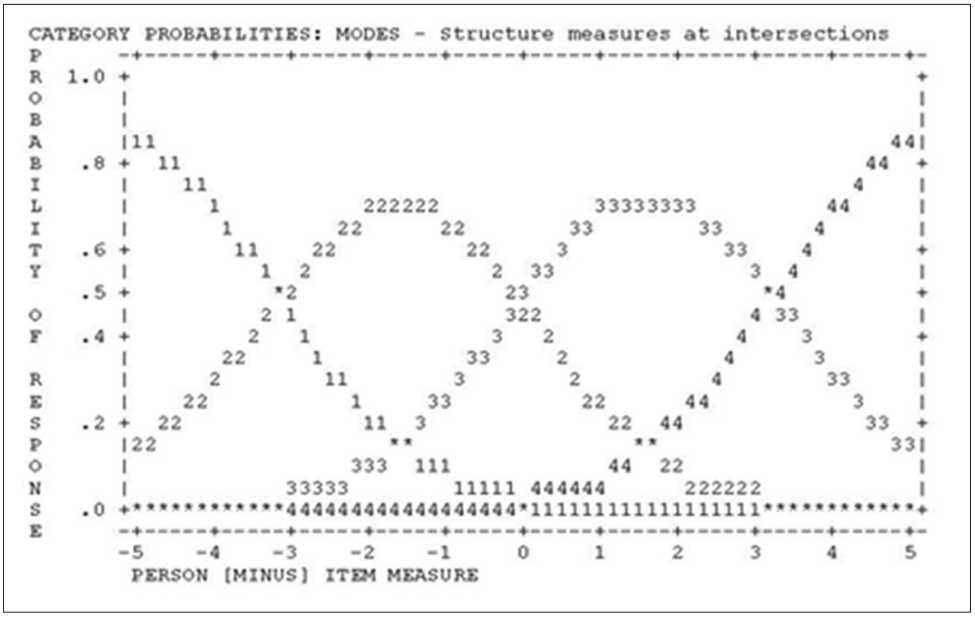

Figure 4: Probability curve for post-collapsing

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE CALIBRATN	CATEGORY MEASURE	
1	1	32	1	-2.20	-2.01	1.10	.92	NONE	(-5.37)	1
2	2	948	25	.27	.23	1.06	1.07	-4.26	-1.90	2
3	3	1919	50	2.04	2.07	.95	.96	.48	2.13	3
4	4	918	24	3.88	3.84	.96	.97	3.79	(4.91)	4

Table 2: Summary of scale collapsing effects

Measures	Pre-collapse	Post-collapse
Person Infit MNSQ SD	0.61	0.54
Item Infit MNSQ SD	0.30	0.28
Person separation	4.47	5.11

SD: Standard deviation, MNSQ: Mean square

respondents are capable of discriminating it. Ironically, Bond and Fox (2015) also support a four-point Likert-scale.

6. DISCUSSION

An accurate scale is a well understood scale (Lozano et al., 2008). It is a sign that the respondents understand the latent trait being tested (Bond and Fox, 2015). Rasch model assumes the respondents understand the scale based on several indicators; acceptable differences between structure calibration of categories and distinct peaks of probability curves for each category (Bond and Fox, 2015; Azrilah et al., 2013).

Initially, the measurement scale used a five-point Likert-scale. Then, Rasch scale calibration analysis reported that scale two was not well distinguished by the respondents. Thus, the scale was improved into a four-point Likert-scale, where Rasch suggests collapsing scale two and three, based on principles elaborated in Section 5.1.

According to Azrilah et al. (2013), the decision either to collapse a rating scale upward or downward does not only depend on the structure calibration or the probability curves, as explained in Section 5.2. Additionally, the decision either to collapse or not should also be made by comparing the pre-collapsing and post-collapsing values of Infit MNSQ standard deviation (SD) for both person and item, as well as the person separation index. According to Azrilah et al. (2013), the best selection is the scale calibration which produces the smallest infit MNSQ SD and the largest person separation.

In the current study, the scale was collapsed upward (Category 2 into 3) into four categories and the data was rerun again. Table 2 shows the difference in the person and item Infit MNSQ SD and person separation of pre- and post-collapsing, which reflects the effectiveness of scale calibration.

Based on Table 2, both person and item Infit MNSQ SD have smaller values after collapsing. Meanwhile, the person separation reported a slight increase of 0.64 from 4.47 to 5.11. These indicators confirmed that the upward scale collapsing is the best option.

7. CONCLUSION

The current study demonstrated analyses on rating scale diagnostics, the indication, remedy, as well as its effects. Rasch provides empirical evidence on rating scale design via the summary of category structure supported by probability curves. In addition, the accuracy of collapsing scale either downward or upward can be detected in the value of person separation and infit MNSQ for both person and item. These measures can confirm that the respondents are capable of differentiating the scales; i.e., understand the scale categories substantially. In this study, the original rating scale had five categories. However, Rasch model detected some distortion in the second category. The issue was further confirmed in the probability curves. Thus, the decision to collapse the category upward was proven accurate as it produced larger person separation index and smaller person and item infit MNSQ SD. Therefore, this article suggests that the developed scale should be administered using a four point Likert-scale.

8. ACKNOWLEDGMENT

This research is funded by the Fundamental Research Grant Scheme (FRGS) managed by the Research Management Center of Universiti Teknologi MARA. The FRGTS is granted by the Malaysian Ministry of Higher Education.

REFERENCES

Abulhasan M S., Khaliq A. (1996), Quality Management Islamic Perspectives. Kuala Lumpur: Leeds Publications.

Albano, J.F. (2009), Developing a measure and an understanding of the individual experience of happiness at work. Oakland, California: Saybrook Graduate School and Research Center, Saybrook University.

Azrilah, A.A., Mohd, S.M., Azami, Z. (2013), A sas Model Pengukuran Rasch Pembentukan Skala & Struktur Pengukuran. Bangi: Penerbit

UKM.
Baird, K., Hu, K.J., Reeve, R., (2011), The relationships between organizational culture, total quality management practices and operational performance. International Journal of Operations and Production Management, 31(7), 789-814.
Bechtel, G.G., (1985), Generalizing the rasch model for consumer rating scales. Marketing Science, 4(1), 62-73.
Bond, T.G., Fox, C.M. (2007), Applying the Rasch Model Fundamental Measurement in the Human Sciences. 2nd ed. New York: Routledge Taylor & Francis Group.
Bond, T.G., Fox, C.M., (2015), Applying the Rasch Model Fundamental Measurement in the Human Sciences. 3rd ed. New York: Routledge Taylor & Francis Group.
Dawes, J., (2008), Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. International Journal of Market Research, 50(1), 61-77.
Denison, D.R., Spritzer, G.M., (1991), Organizational culture and organizational development. Research in Organizational Change and Development, 5, 1-21.
Detert, J.R., Schroeder, R.G., Cudeck, R., (2003), The measurement of quality management culture in schools: Development and validation of the SQMCS. Journal of Operations Management, 21(3), 307-328.
Detert, J.R., Schroeder, R.G., Mauriel, J.J., (2000), A framework for linking culture and improvement initiatives in organizations. Academy of Management Review, 25(4), 850-863.
DeVellis, R.F., (2003), Scale Devlopment: Theory and Applications. California: Sage Publications.
Gambi, L.D.N., Gerolamo, M.C., Carpinetti, L.C.R., (2013), A theoretical model of the relationship between organizational culture and quality management techniques. Procedia - Social and Behavioral Sciences, 81, 334-339.
Ishak, A.H., Osman, M.R., (2015), A systematic literature review on Islamic values applied in quality management context. Journal of Business Ethics. DOI 10.1007/s10551-015-2619-z.
Ishikawa, K. (1985), What is total quality control? The Japanese Way. London: Prentice Hall.
Khaliq, A., (1996), Quality management foundation: An agenda for Islamization of management knowledge. Malaysian Management Review, 31(1), 44-52.
Khaliq, A., Shamim, A. (1996), Islamic values in management: A comparative study. In: Abulhasan, M., Sadeq, A. M., Khaliq, A., editors. Quality Management Islamic Perspectives. Kuala Lumpur: Leeds Publications.
Linacre, J.M., (1999), Investigating rating scale category utility. Journal of Outcome Measurement, 3(2), 103-122.
Lozano, L.M., García-Cueto, E., Muñiz, J. (2008), Effect of the number of response categories on the reliability and validity of rating scales. Methodology, 4(2), 73-79.
Muhammad, A.B. (1985), Management and Administration in Islam. Saudi Arabia: King Fahd University of Petroleum & Minerals.
Muhammad, A.B., (2005), Management principles derived from the sources of Islam. In: Mazilan, M., Salleh, S.S.S., editors. Quality Standard from the Islamic Perspectives. Kuala Lumpur: IKIM.
Naceur, J. (2005), Islam and Management. Riyadh: International Islamic Publishing House.
Prajogo, D.I., McDermott, C.M. (2005), The relationship between total quality management practices and organizational culture. International Journal of Operations and Production Management, 25(11), 1101-1122.
Prajogo, D.I., McDermott, C.M. (2011), The relationship between multidimensional organizational culture and performance. International Journal of Operations and Production Management, 31(7), 712–735.
Salzberger, T., Koller, M. (2013), Towards a new paradigm of measurement in marketing. Journal of Business Research, 66(9), 1307-1317.
Sany, S.M.M., Rushami, Z.A., Zakaria, A., Hartini, M.H., Muhammad, N.M. (2011), Aplikasi Sistem Pengurusan Kualiti dari Perspektif Islam. Sintok: Penerbit UUM.
Singh, J. (2004), Tackling measurement problems with item response theory. Journal of Business Research, 57, 184-208.
Siti, A.B., Bharudin, C.P., Raja H.R.S. (2010), Suntikan nilai-nilai Islam ke atas perlaksanaan penambahbaikan berterusan dalam konteks sistem pengurusan kualiti ISO9000. Journal Syariah, 18(1), 91-122.
Siti, A.B., Ilhaamie, A.G.A. (2011), Malaysian islamic quality management system MS 1900 from an Islamic perspective: An implementation model. Shariah Journal, 19(2), 85-106.
Syed, O.A. (1996), Quality and productivity conciousness: An Islamic approach. In: Quality Management Islamic Perspectives. Kuala Lumpur: Institut Kefahaman Islam Malaysia.
Tennant, A., Conaghan, P.G., (2007), The rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a rasch paper? Arthritis Care and Research, 57(8), 1358-1362.