

Investigating Rating Scale Design via Rasch Measurement Model: Analysis on a Scale to Assess Islamic Values Application in Quality management Context

Amal Hayati Ishak^{a*}, Muhamad Rahimi Osmana^a, Siti Khadijah Ab Manan^a,
Rafeah Saidon^a, Ghafarullahuddin Din^a

* Corresponding author: Amal Hayati Ishak, amalhayati@salam.uitm.edu.my

^aAcademy of Contemporary Islamic Studies, Universiti Teknologi MARA Shah Alam, 40450
Malaysia, amalhayati@salam.uitm.edu.my

Abstract

<http://dx.doi.org/10.15405/epsbs.2016.08.9>

The primary purpose of this article is to empirically analyze the rating scale categories applied in a scale to assess application of Islamic values in quality management context. In the literature, a plethora of Islamic values applied in quality management context 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 Rasch Measurement 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.

© 2016 Published by Future Academy www.FutureAcademy.org.uk

Keywords: Rasch Measurement Model; scale calibration; Islamic values; quality management.

1. Introduction

The mainstream of quality management has been started since the advent of industrial revolution. Though now it does not take the form business environment of the previous era, the tools refer to the similar philosophy which aims for customer satisfaction and overall participation of organizational

resources. Since then, research on quality management has been widely conducted, involving different aspects including organizational performance, culture, or specific industrial practices.

Since the pioneers of quality management were Western players, its philosophical foundation has been largely dominated with the western values, which have been criticized as narrowly focused to outputs (Syed Othman, 1996; Naceur, 2005; al-Buraey, 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 were also known to initiate the practice of quality circles as a platform to disseminate knowledge and experience between organizational members (Khaliq & Shamim 1994).

In similar vein, a series of research have concluded on the importance of values in supporting successful implementation of quality management. 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 & 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 analysed values based on general framework of Hofstede's cultural dimensions (Baird, Hu & Reeve, 2011; Baird et al. 2011), Competing Value Framework (CVF) of O'Reilly (Prajogo & McDermott, 2005; 2011; Gambi, Gerolamo & Carpinetti, 2013), Organizational Culture Profile (OCP) of Quinn & Rohrbaugh (Denison & Spritzer, 1991), and Detert's framework (Detert, Schroeder & Mauriel, 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), al-Buraey (2005), Siti Arni, Baharudin and Raja Hisyamudin (2010); Sany et al. (2011) and Siti Arni 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.

This article proposes a scale to assess Islamic values application in QM context. The scale development 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 a rating scale.

2. The measurement scale

Based on an extensive 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 an expert review followed by Fuzzy Delphi analysis.

In this study, the experts are selected based on their qualification and experience from the academia and industry. A total number of 11 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, García-cueto, & Muñiz, 2008). On top of that, since five or seven point Likert scale produces similar results (Dawes, 2008), thus the current study decided to use a five point scale.

3. 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, 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 staffs that directly involve with quality management tasks. Out of 100 questionnaires distributed, 59 were returned with a response rate of 59%. The instrument have 61 items which resulted from a systematic literature review focusing on the topic of Islamic values in quality management context (Ishak & Osman, 2015). The data was then analyzed using the Winstep 3.68.2, 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 values based on the probability of success, which depends on the differences between person ability and item difficulty. The value enables 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 & Fox 2007; Azrilah et al., 2013), which is the focus in this article.

4. Data analysis

Scale calibration in Rasch provides empirical evidence to detect whether the respondents understand and able to differentiate the scaling labels. Ideally, Linacre (1999) points four indicators to diagnose a problematic rating scale as summarised 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 & Fox, 2007).

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 & Fox, 2007; Azrilah et al., 2013). An effective scale calibration can be detected from increased item reliability and separation (Azrilah et al., 2013).

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 $1.4 < x < 5$
Probability curves	Each category is expected to have distinct peak

*Source: Linacre (1999)

4.1. Diagnosis

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). In this study, the respondents seemed unable to differentiate category 2 (disagree). This had been indicated empirically as shown in Figure 1 below.

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

```

TABLE 3.2 Pilot88.sav                                ZOU410WS.TXT Apr  8 17:14 2015
INPUT: 88 PERSON 44 ITEM REPORTED: 88 PERSON 44 ITEM 5 CATS WINSTEPS 3.72.3
-----
SUMMARY OF CATEGORY STRUCTURE. Model="R"
-----
|CATEGORY  OBSERVED|OBSVD  SAMPLE|INFIT  OUTFIT||STRUCTURE|CATEGORY|
|LABEL SCORE COUNT %|AVRGE  EXPECT|  MNSQ  MNSQ||CALIBRATN| 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
|-----+-----+-----+-----+-----+-----+-----+
|MISSING      55  1|  3.26      |          |          ||          |          |
-----
    
```

Figure 1. Diagnostics for five rating categories (Pre-collapsing)

4.2 The Remedy

As explained, 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 category 3.

Following the guidelines of Bond and Fox (2007) that collapsing categories should be logical. Thus, category 2 (very minimal implementation) is more logical to be collapsed with category 3 (minimal implementation), rather than category 1 (no implementation). Figure 2 shows the results of collapsing category 2 into category 3.

In Figure 2, the average observed count is increasing consistently. Such increment is referred as monotonic ordering by Bond & Fox (2007). Additionally, the difference of structure calibration between categories is also within acceptable range; above 1.4 but below 5.0. Therefore, it is more suitable to use four scaling 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.

```

TABLE 3.2 Pilot88.sav                                ZOU917WS.TXT  Apr 22 23:45 2015
INPUT: 88 PERSON 44 ITEM  REPORTED: 88 PERSON 44 ITEM 4 CATS  WINSTEPS 3.72.3
-----
SUMMARY OF CATEGORY STRUCTURE.  Model="R"
-----
|CATEGORY|OBSERVED|OBSVD SAMPLE|INFIT  OUTFIT||STRUCTURE|CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT| MNSQ  MNSQ||CALIBRATN| MEASURE|
|-----|-----|-----|-----|-----|-----|-----|-----|
| 1 1 91 2| -1.56 -1.65| 1.16 1.18|| NONE |(-4.29)| 1
| 2 2 889 23| -.09 -.10| 1.03 1.05|| -3.16 | -1.62 | 3
| 3 3 1919 50| 1.50 1.52| .94 .95|| -.06 | 1.59 | 4
| 4 4 918 24| 3.31 3.27| .96 .96|| 3.22 |( 4.35)| 5
|-----|-----|-----|-----|-----|-----|-----|
|MISSING| 55 1| 2.60 | | | | |

```

Figure 2. Results of Collapsing Categories

5. Discussion

An accurate scale is a well understood scale (Lozano, García-cueto & Muñoz, 2008). Thus, it is a sign that the respondents understand the latent trait being tested (Bond & Fox, 2007). Initially, the measurement scale used a five-point Likert-scale in Pilot I. 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 rating scale upward or downward does not only depend on the probability curves. Additionally, the decision either to collapse or not should also be made based on comparison of two values; the Infit MNSQ SD for both person and item, and the Person Separation index. The researcher should select 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.

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

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 scale upward calibration is the best option.

6. 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 analysis. 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. understands 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. 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 4 point Likert-scale.

Acknowledgement

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

References

- Abulhasan M. Sadeq. (1996). Quality management in The Islamic Framework. In Abulhasan M. Sadeq & Khaliq Ahmad (Eds.), *Quality management Islamic Perspectives*. Kuala Lumpur: Leeds Publications.
- Azrilah Abdul Aziz, Mohd Saidfudin Masodi, & Azami Zaharim. (2013). *Asas Model Pengukuran Rasch Pembentukan Skala & Struktur Pengukuran*. Bangi: Penerbit UKM.
- 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.
- Chang, P., & Hsu, C. (2011). An assessment model for hydrogen fuel cell applications : Fuzzy Delphi Approach. *International Journal of Social Science and Humanity*, 1(3), 218–223.
- Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Clinical Methods*, 5, 194–197.
- 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.
- Ishak, A. H., & Osman, M. R. (2015). A systematic literature review on Islamic values applied in quality management context. *Journal of Business Ethics*.
- Abulhasan M. Sadeq. (1996). Quality management in The Islamic Framework. In Abulhasan M. Sadeq & Khaliq Ahmad (Eds.), *Quality management Islamic Perspectives*. Kuala Lumpur: Leeds Publications.
- Azrilah Abdul Aziz, Mohd Saidfudin Masodi, & Azami Zaharim. (2013). *Asas Model Pengukuran Rasch Pembentukan Skala & Struktur Pengukuran*. Bangi: Penerbit UKM.
- 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.
- Chang, P., & Hsu, C. (2011). An assessment model for hydrogen fuel cell applications : Fuzzy Delphi Approach. *International Journal of Social Science and Humanity*, 1(3), 218–223.
- Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Clinical Methods*, 5, 194–197.

- 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.
- Ishak, A. H., & Osman, M. R. (2015). A systematic literature review on Islamic values applied in quality management context. *Journal of Business Ethics*. Volume?(Issue?), xx-xx page number?
- Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, 3(2), 103–122.
- Linacre, J. M. (2011). *Winsteps Rasch Measurement Computer Program User's Guide*. Beaverton. Retrieved from winsteps.com
- Lozano, L. M., García-cueto, E., & Muñoz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, 4(2), 73–79.
- Muhammad A. Al-Buraey. (2005). Management Principles Derived from the Sources of Islam. In Mazilan Musa & S. S. S. M. Salleh (Eds.), *Quality Standard from the Islamic Perspectives*. Kuala Lumpur: IKIM.
- Nik Mustapha Nik Hassan. (1996). An Islamic approach to quality and productivity. In Abulhasan M. Sadeq & Khaliq Ahmad (Eds.), *Quality management Islamic Perspectives*. Kuala Lumpur: Leeds Publications.
- Siddiq Fadzil, Abd Malek Awang Kechil, Malek Shah Mohd Hassan, Mohd Ezani Mat Yusoff, Mohd Fauzan Samsudin, & Abibullah Noordin. (2010). *Pengurusan Islami Menghayati Prinsip dan Nilai Qurani*. Kuala Lumpur: Akademi Pengurusan YaPEIM Sdn. Bhd.
- Syed Othman Alhabshi. (1996). Quality and Productivity: An Islamic approach. In Abulhasan M. Sadeq & Khaliq Ahmad (Eds.), *Quality management Islamic Perspectives*. Kuala Lumpur: Leeds Publications.