APPLICATION OF RASCH MODEL MEASUREMENT IN THE CONTENT VALIDATION OF THE INTENTION TO STAY SCALE (ITSS) AMONG MEDICAL ACADEMICS AT PUBLIC UNIVERSITIES

Wan Ismahanini Ismail 1,2, Roziah Mohd. Rasdi 1, Rahinah Ibrahim 3, Bahaman Abu Samah 1

1 Faculty of Educational Studies, Universiti Putra Malaysia, 43400 Serdang, Malaysia
2 Registrar Office, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia
3 Faculty of Design and Architecture, Universiti Putra Malaysia, 43400 Serdang, Malaysia

Abstract. The number of medical academics leaving public universities is increasing; thus, an investigation into their intention to stay in the academia is crucial. As such, development of the Intention to Stay Scale (ITSS) to measure the existing medical academics’ intention to remain in service at public universities is timely. However, a sound scale should be guarded against threats of validity to ensure that the findings and generalizations would be valid. This study aimed to examine content validity, which is one aspect of validity in this study, comprising five principal dimensions with 52 proposed items. The five dimensions of ITSS, namely, Feelings about Behaviour, Behavioural Beliefs, Normative Beliefs, Control Beliefs and Efficacy Beliefs, were generated from the Integrated Behavioural Model. Review and feedback from nine experienced medical academics as panel of experts were assessed using Rasch measurement model. All 52 items have been found acceptable for next stage of analysis, that is, the pilot study. However, several items were suggested to be discarded to ensure better clarity in achieving the objective of the study. Furthermore, additional items would be required for Normative Beliefs which had fewer than five items to ensure that there would be sufficient items for testing should any item be discarded after the pilot test. It can be concluded that the panel of experts’ decision and acceptability of the quality of each category that were able to explain intention to stay would be one of the critical steps in developing such a scale. The findings of this study would help researchers improve the current scale to ensure that it would measure what should be measured and achieve the intended objectives.

Keywords: Intention to Stay, Integrated Behavioural Model, Rasch Measurement Model, Content Validity Test.

Introduction. There are many reasons or factors that have been identified which contributes towards the resignation of medical academics both nationally and internationally (Hagander et al., 2013; Pati et al., 2013; Smith and Bunton, 2012; Wai, Dandar, Radosevich, Brubaker, and Kuo, 2014). However, there lacks a mechanism to measure their intention to stay. Few scales have been developed to measure concepts like turnover intention, intention to leave, and intention to stay. Nevertheless, there is always some misunderstanding between variables involved and its suitability to be used in different industries, sectors and regional applications (Dileep and Normala, 2014). It is not easy to develop a scale, a sound scale should be protected against threats of validity to ensure that the findings and generalizations would be valid. Therefore, the aimed of this study is to examine content validity, which is one aspect of validity in this study, comprising five principal dimensions with 52 proposed items to measure medical academics’ intention to stay.

Principal dimensions were generated from Integrated Behavioural Model (IBM) and the items were derived from review of related literatures, resignation report (Medical Deans’ Council of Public Universities, 2015) and promotion guidelines for medical academics (Ministry of Higher Education, 2015). Empirical evidences are essential to ensure validity and reliability of a scale, especially in ensuring that the outcomes are quality items as there has been evidence of items not being reviewed and their validity not being verified systematically (Zain et al., 2011).

Many aspects need to be considered in the development of a scale. One of these aspects is validity. According to Messick (1995), construct validity can be described in six distinctive aspects that embrace the key concept of validity. This key concept represents a “unified concept” (Messick, 1996). However in this study, focus was given to the aspect of content validity as an initial step in scale development.

Content validity addresses the applicability and reflection of the content upon which the items are based and the technical quality of those items (Wolfe and Smith, 2007b). Devellis (2003) refers to content validity as item sampling sufficiency, which is further defined as how the content domain is explained by a particular group of items. Meanwhile, Wolfe and Smith (2007b), suggest that evidence of validity relating to content can be generated from multiple ways, one of which includes expert evaluation. The content validity test is an assessment of items in a scale by a group of experienced panel of experts. It involves a systematic review of the scale’s content to ensure the items address the intended latent variable and excludes unnecessary items that should not be included (Zain et al., 2011).

Expert evaluation plays an important role in scale development, especially for content validation (Cheng, Chen, Liu, & Chen, 2011; Joo and Yeon Lee, 2011; Zhao et al., 2014). The duties of experts include providing input into the definition of the purpose of the scale, the definition of the construct, and the design and content of the scale as well as providing insights into fairness, accessibility, and sensitivity (Wolfe and Smith, 2007). Besides assessing and checking item quality and construct relevance, experts also play a role in hierarchical ranking of the items (Graham and Beltyukova, 2015). Expert review is primarily aimed at discovering problems with an instrument in order to eliminate the problems before the instrument is used in the main study for data collection purposes as well as to identify, group and arrange the items in a sequence which would indicate errors of measurement (Olson, 2010).
In other cases, experts also play important roles in providing explanations in support of achieving the scale’s objective (Schmiedel, Vom Brocke, & Recker, 2014), and judging the suitability of every indicator as suggested by survey researchers (Mat-jizat and Mckay, 2011). Normally, there are not many expert panel review members, which may only be very few, that is, two or three experts to several, up to 20 reviewers (Olson, 2010). However, there is a growing trend whereby the number of experts increases up to 27 (Schmiedel et al., 2014) and 30 (Zhao et al., 2014) which plays a role in reviewing proposed items for scale development. Although it has been claimed by Kitchenham and Pfleeger (2008) that there is no content validity statistic, such claim has been countered by Aziz et al. (2008) who have indicated that assessment of content validity can be made accurately by using Rasch analysis. The Rasch measurement model considers both person ability and item difficulty (Bond & Fox, 2015). In the present study, Rasch model was used to determine whether the quality criteria under each of the five dimensions had been agreed by the panel of experts and therefore confirm their content validity. If this was achieved, it can be said that the content aspect of the construct validity for the scale had been empirically tested. The adopted framework and guideline employed were as proposed by Messick (1996) and Wolfe and Smith (2007a, 2007b), respectively, in demonstrating validity claims using evidence from Rasch analysis.

There are five common types of validity aspects applied in the development of scale, namely, content validity, substantive validity, structural validity, generalizability and interpretability. However in this initial stage, the development of scale emphasized on the aspect of content validity. Content validity involves a few phases: (1) identifying the purpose of the scale; (2) defining the construct; and, (3) initial scale development. Nevertheless, the discussion in this article focuses on the third phase, that is, the involvement of the expert panels in the initial development of the scale. The next section in this paper offers a brief description of the Rasch measurement model. The subsequent section then provides an explanation of the underpinning theoretical framework of this study. This is then followed by methodology and the results. This paper concludes with the discussion of the results, the limitations of the present study and future directions in this area and conclusion drawn from this study.

**Rasch measurement model.** Intention to stay (ITS) is a latent variable of the human behaviour, or an abstract psychology construct (Graham, 2012) that is not directly observable. An instrument with items developed for the measurement of latent variable (i.e., intention to stay or leave) is designed to gauge its actual magnitude for each person measured at the time and place of measurement. The “actual magnitude” refers to the true score (Devellis, 2003). Therefore, Rasch measurement model was applied in this study because of its objective measurement in which the correct score could be used to infer the unobservable trait (Bond & Fox, 2015), and in this case, the medical academics’ intention to stay in service at Malaysian public universities.

A key advantage of scale development employing Rasch analysis is that it can be used repeatedly to measure latent variable, just as the same as in the measurement for weight and height (Azrilah, Mohd Saidfudin, & Azami, 2013). Moreover, Rasch analysis allows the researcher to modify the scale based on empirical judgments in discarding some items or modifying others, and altering the scale to produce a measurement that is unidimensional. Such changes at the level of survey scale and also that of the items will lead more towards a scale with optimised outcomes as a result of data collection, and therefore minimising errors of measurement. This will be highly valuable because the researcher will have an opportunity to communicate with research participants and understand their feelings, beliefs or attitudes through their responses to the survey items (Sondergeld & Johnson, 2014).

The justification for using Rasch measurement model instead of other models such as classical test models is that it would provide a mathematical framework against which scale developers can compare their data with. The Rasch model is based on the idea that worthwhile measurement contains investigation of only one human attribute at a time (Bond and Fox, 2015). Therefore, person and item performance that deviate from the line of inquiry (whether more than or less than) (Bond and Fox, 2015) gives a sign to the scale developer to reconsider the wording of the item and interpretations of the score of the data. The uniqueness of this model is that it has the ability to estimate each item difficulty and person ability on a common logit scale, and that each of these estimates has a degree of error connected with it. Another key advantage of using Rasch analysis is its ability to show the relation between person and item in the form of a meaningful pictorial ‘map’ (Bond and Fox, 2015). In addition to the use of Rasch analysis to test content validity test, this study was based on relevant model derived from sound theories which form the underpinning theoretical framework.

**Theoretical framework.** In the present study, the key dimensions for intention to stay scale were derived from the Integrated Behavioural Model (IBM) which is an integrative model that focuses primarily on the determinants of behavioural intention (Fishbein, 2000; Fishbein and Cappella, 2006). IBM is an extension of Theory of Reasoned Action (TRA) (Fishbein and Ajzen, 1975) and Theory of Planned Behaviour (TPB) (Ajzen, 1991). This model contains the elements from TRA and TPB as well as other influential theories (see Figure 1).
The most important determinant of behaviour in the IBM is similar to that of the TRA and also that of the TPB, which is intention to perform the behaviour (Montano and Kasprzyk, 2008). This model focuses on a person’s aim to behave in a specific manner (in this case, ITS) as both a dependent variable as well as an independent variable (Bleakley and Hennessy, 2012). This theory chiefly models formation of intentions, emphasizing on the roles of attitudes, norms and efficacy with regard to one another in indicating variations of the intention. This model posits that behavioural intention is characterized by the following three types of construct: attitude toward the behaviour, perceived norm and personal agency (Montano and Kasprzyk, 2008).

Each of these three types of construct, or the direct measures, is characterized by a matching group of prominent and fundamentally behavioural, normative and control beliefs (Ajzen, 1991). However, no definition is given for feelings about behaviour. Therefore, the definition of this dimension in this study was adapted from Watson, Clark, and Carey (1988). Five principal dimensions generated for the Intention to Stay Scale (ITSS) were feelings about behaviour, behavioural beliefs, control beliefs and efficacy beliefs, as illustrated in Figure 2. However, items in the scale were generated not solely from the aforementioned theories and model but also derived from other relevant theories, literatures and promotional guidelines for medical academics. All the items developed from past studies were adapted to signify the context of medical academics.

**Methodology.** Reviews of the items by expert panels involved a few phases. In the first phase, supervisory committee members played a role in checking all the initial items. From the original 79 items, these were reduced to only 52 items. Later, after modification in the second phase, the scale was reviewed by a two-member panel of experts in work psychology and Rasch model, respectively. They were invited to join the expert panel and review the questionnaire through e-mail. The survey asked for Yes/No dichotomous answers, item quality and different views and comments through open-ended questions. Based on the comments in the first and the second phases, all the items were rephrased to ensure that the scale would truly evaluate intention to stay, and that no prediction questions (i.e., future intentions) would be measured. In the third phase, eleven experienced medical academics from public universities were identified and invited as panel of experts to review the 52-item scale. The selection of experts was based on their seniority, wide variety of experiences in holding administrative posts (e.g., Dean and Deputy Dean of Medical Faculty, Hospital Director), working as a medical academic at a public university and conducting instrument development studies especially using Rasch measurement model. They were invited through e-mail by using the snowball sampling method. Out of eleven candidates, nine experts from various Malaysian public universities of different grades and positions kindly gave their feedback for further improvement.

A questionnaire was developed to determine whether the panel of experts which consisted of medical academics agreed with the proposed set of items. The survey asked for Yes/No answers, item quality and different views and comments through open-ended questions. For content validity test, responses were examined using the basic Rasch dichotomous model \((Yes=1/No=0)\) answers. The Wright map (or variable map), which displays the distribution of item agreement on the right and allocation of experts on the left according to the item number, was analysed. Then, the scale
was sent for proofreading to check any spelling and grammatical errors before proceeding to another phase, which was the cognitive debriefing. The analysis using the basic Rasch dichotomous model was made with the help of the widely used software Winsteps version 3.72.3 (Linacre, 2011).

**Findings.** The decision made towards the 52 items can be classified into three groups. The most agreed-to items under the first group, these are located at -1.92 logits (SE 1.91) and the least agreed-to items are under the third group. They are item numbers 30 (*Friend support*), 32 (*Colleagues will stay*), 9 (*Feeling towards clinical practice*) and 31 (*Colleagues support*), located at the top of item distribution at +1.62 logits (SE .91) for item 30 and 32; and +0.76 logits (SE 0.99) for items 9 and 31.

Attributes were checked based on the PMC with the following rule of thumb: PMC=x, 0.4<x<0.8 (Aziz et al., 2008). These items were found to have fallen outside of this range: *colleagues support, career remain sustainable, role as educator, need to obtain a grant, allocation for career development and family support*. In Rasch analysis, further verification would be required by examining the OUTFIT measures for MNSQ value: MNSQ=y, 0.5<y<1.5. These items were found to have violated this parameter: *feel proud with work, feeling dealing with patients, allocation for career development, well-defined conversation, university’s decision making, family support, perform clinical tasks and skill to do administrative works*. Additional analysis on the Z-std value, where Z-std=z, -2<z<+2, showed that no item was beyond the set value. Hence, all items were considered to be acceptable for further analysis. Furthermore, the fit statistics revealed that there were 36 items which the experts had unanimously agreed. These corresponded to the most agreed-to items on the variable map under the first group category. This category was also considered to be no threat to the measurement.

Unidimensionality is apparently a crucial feature in determining the instrument that measures in one continuum (single direction and shape). An instrument which is vague in measuring the parameters that it is supposed to measure may produce a confusing result. Raw variance explained by measure as shown in Figure 3 was an achievement at 26.5% rather than 27.0%.

Rasch analysis requires a minimum of 40% raw variance explained by measures achievement and a benchmark level of a good unidimensionality instrument measure is 60% (Azrilah et al., 2013).

![Fig. 3. Table of standardized residual variance.](image)

However, the result showed low achievement and did not exceed the minimum requirement of 40%. The reason for not achieving the measurement was due to the interference of items that are known as noise. Noise level measured was 41.5%, which exceeded the control limit of 15%. Figure 4 displays the items that were contributed to the noise. Ten items had high standardized residual correlation values that exceeded 0.70. This showed that the respondents (or expert panels) were viewing those particular pairs of items as similar and confusing.

![Fig. 4. Largest standardized residual correlations.](image)

Besides unidimensionality, this study investigated further on person statistics (see Figure 5), whereby it confirmed the result of item statistics. The data analysis showed that only two expert panels had fallen outside the range, that is, PE8 (OUTFIT MNSQ value of .29) and PE6 (PMC value of .31). However, these values did not vary greatly from the values that had been set and INFIT Z-std value showed that they were within the acceptable range ((Z-std=z, -2<z<+2). The person statistics also revealed that the top three people with maximum measures were the expert panels who had agreed to all of the items. They were accepted in this analysis because they did not represent any threat to the measurement.
Counterchecking against the Guttman scalogram (see Figure 6) indicated that the items given were very agreeable to PE2, PE5 and PE7. The purpose of scalogram analysis was to observe person response distribution per item. It would also show a particular Person Misfit. The persons that were in disagreement over a number of the items were PE3 and PE4. Nevertheless, there were some areas of concern for PE8 item 32, PE1 for items 29 and 30, PE6 for items 9 and 19, and PE9 for items 13, 31 and 32. Theoretically, the scores should not be zero; nevertheless, the expert panels responded 0 to those items, which showed their disagreement.

There were several reasons for the disagreement among experts which involved Part A (Feelings about behaviours), Part B (Behavioural beliefs) and Part C (Normative beliefs) such as similarities of questions being asked, vague or unclear statement and improvement on the sentence structure.

**Discussion.** In this study, expert reviews play a role in content review and to ensure the constructed items are applicable to all members of the targeted population that are medical academics in Malaysian public universities. According to Wolfe and Smith (2007b), validity evidence relating to content aspect of validity can be generated from multiple ways, and one is through expert reviews. In order to ensure that content validity would not be compromised even if the items were to be discarded, the filter process is to be implemented by using Rasch measurement model. Where all the feedback from the expert panel is analyzed using this model. There were several analyses involved, among them are items fit, person misfit order, variable map, Principal Component Analysis or unidimensionality and Guttman scalogram. This study aimed to examine one aspect of validity, that is, content validity of five principal dimensions with 52 proposed items used to measure medical academics’ intention to stay in service at public universities. The current study presents a review from nine experienced medical academics from several public universities. As mentioned in the earlier explanation, all 52 items under five main dimensions were accepted for next stage, that is, the pilot test. However, based on result shown in Figure 3 and Figure 5, several items had to be discarded to ensure clarity in achieving the objective of the study. These items were item number 1, 3, 6, 8, 9, 10, 13, 19, 26, 28, 29, 30, 31, 32, 49 and 52. For example, item 1 (Feelings about behaviours dimension: feel proud with work) seemed to have a similar meaning with other items such as item number 8, 10, 26, 28, 49 and 52. Therefore, items which had been identified by the Rasch model as problematic would call for a more thorough examination to enhance their level of unidimensionality.

Two approaches can be applied based on the recommendations by Azrilah et al. (2013), namely, either to retain the items by rephrasing them for better clarity or to discard them based on filter method. In this study, researchers chose not to discard the items but rather opted to rephrase and change the sentence structure for better clarity and improved meaning in achieving the objective of the study. Moreover, additional items that were required for dimension with fewer than five items under Part C, that is, Normative Beliefs as this dimension only had four items. Even though Hinkin, Tracey, & Enz (1997) have suggested that a quality scale would comprise four to six items per dimension, the minimum number of items required for each dimension for this study was at least five. This was to ensure that there would be sufficient items for testing should any item would be discarded after the pilot test. Theoretically, studies have shown that both theories (TRA and TPB) that are combined to form IBM, are considered as well-established theories which explain the relation between attitude and intention that leads to performing a behaviour. Therefore, recent development in the literature has included new emerging variables extracted from other theories such as Self-Determination Theory, Social Capital Theory, Social Exchange Theory and so forth to these theories (Tangaraja, Mohd Rasdi, Ismail, and Abu Samah, 2015). Conversely, the interest of this present study was only in the concept of intention or intention-to-behaviour as had been explained in details by Fishbein and Ajzen (1975).
Conclusion. This article presents an analysis of a test to help improve the content validity of a scale to measure medical academics’ intention to stay in service at public universities through reviews from panel of experts. Reviews from panel of experts is one of the validity evidence relating to content aspect of validity. The content validity test using Rasch measurement model in the current study investigated the expert panels’ decision and acceptability of the quality of each item and dimension that could explain the intention to stay before the construct would be tested in the next stage, in which the construct validity of the scale would be further examined. Where any item to be dropped or modified can be identified by using specified analysis in ensuring that it achieves the objective of the study. In conclusion, the lack of fit in existing scale and the weaknesses of the previous scale indicate that there had been gaps in the current literature. This study offers some insights to the literature on quality scale measurement in the social sciences, with specific reference to the intention to stay scale which has been tailored to the needs of medical academics in higher education.

Limitations and future directions. Because of time constraints, the present study combined the definition of each dimension and item developed to be checked together by panel of experts. However, serious attempts were made through few phases of quality and content checking by expert panels in order to ensure that each dimension and item developed would be appropriate to test medical academics’ intention to stay in service at public universities. These attempts could help to improve the quality of the items being developed. For future study, another appropriate theory or model (or other appropriate theories and models) that could explain intention to stay in service should be included to strengthen the existing theoretical item hierarchy. The chosen theory or model (or theories or models) should emphasize the ability to represent the intention to stay in service. Combining several theories and models is perceived as the step that is able to strengthen the existing theories and models related to intention to stay in service. In addition, before the ITSS is implemented for pilot testing, it has to be verified first for comprehensiveness by conducting cognitive debriefing to the potential respondents. This is to ensure that the adapted items are easy to understand and are in line with the research objective. Finally, to ensure the highest quality measure some steps using Rasch analysis have to be taken multiple times in an iterative process.

Acknowledgments: This project is financially supported by Ministry of Higher Education and Universiti Putra Malaysia.

References


