VALIDATION OF TECHNICAL AND VOCATIONAL TEACHERS' COMPETENCY EVALUATION INSTRUMENT USING THE RASCH MODEL

Nor Fatimah A Aziz[1], Hishamuddin Ahmad[2], Irdayanti Mat Nashir[3]
[1]Bahagian Biasiswa & Pembiayaan, Kementerian Pendidikan Malaysia, Cyberjaya Selangor
[2] Faculty of Human Development, Universiti Pendidikan Sultan Idris, Tanjung Malim, Perak.
[3] Faculty of Technical and Vocational, Universiti Pendidikan Sultan Idris, Tanjung Malim, Perak.
*email : norfatimah_aziz@yahoo.com

Abstract

This pilot study was conducted with the aim to validate the instrument used in evaluating the competency of teachers in the field of technical and vocational education. This instrument consists of 45 items and is administered to 53 teachers from a selected vocational college. The Rasch Model with the help of the Winstep Version 3.72.3 software has been used in this study for the purpose of checking the functionality of the item and the validity of the instrument. An analysis has been made based on the suitability of items in measuring the construct, item and person reliability and separation index, polarity and residual correlation value. The Rasch analysis showed that the item reliability was valued at 0.92 while the person reliability valued at 0.96 with their item MNSQ between overfit (<0.6) and misfit (>1.4). Based on the findings, there are three items that were dropped because of failing to meet the inspection criteria. The finalized instrument consists of 42 items, in which it is suitable for evaluating the four constructs in the competency evauation of technical and vocational teachers in vocational colleges.

Keywords: competency, Rasch Model, vocational, teacher, evaluation.

INTRODUCTION

The year 2017 saw an increase in student enrollment rates for vocational education in Malaysia. A total of 6.3 percent (25,947 students) were registered in 2016 and in 2017 this percentage had increased to 7.2 percent (27,886 students). Most of these enrollments involved a total of 80 Vocational colleges with the percentage of 83 percent (23,035 students), compared to the enrollment into secondary schools that offer vocational programs (Education Performance and Delivery Unit, 2018). In order to ensure the quality of the student and fulfill the requirement of the industry, it is essential that competent teachers are needed.

Hence, teachers' competency is seen as a major contributor in ensuring the success of these students. In order to determine whether the educational institution will success or not, it all depends on the teacher itself as a medium to convey the knowledge and that a competent teacher will be able to optimize the student's potential. Hasnah and Jamaludin (2017) also agreed that teachers needed to have a high level of competency in carrying out their duties by possessing knowledge in subject matter, mastering the pedagogical skill, technology and communication skills, diversifying teaching resources and strategies, and having positive attitudes and personality.

To ensure the validity and reliability of the constructed instrument, a pilot study has been conducted at a vocational college. Once the data is obtained, an analysis has been made using the Rasch Model for the purpose of ensuring the reliability and validity of the instrument. Through this approach, researchers are able to examine the functionality of the item by diagnosing six different aspects, which is unidimensionality reliability and item and person separation index, the suitability of the item fit in measuring the construct, items polarity, standardized residual correlation value, and the validity scale that was used.

PURPOSE OF STUDY

This pilot study was constructed to test the validity of TVET teachers' competency instruments in vocational colleges. In this study, Rasch Model is used to validate the measurement functioning of this instrument.

METHODOLOGY

This pilot study is a survey study, using a quantitative approach. A set of questionnaire was distributed to a total of 53 teachers at a vocational college located in Selangor. The number of respondents that has been decided is sufficient in accordance with the opinion presented by Linacre (1994), in which to achieve the 99 percent confidence level, the most appropriate sample to use is 50 people with a minimum sample size of between 27-61 respondents.

Researchers have adapted an instrument from several instrument in order to evaluate and ensure the level of competency among TVET teachers. This instrument consists of 45 items that have been divided into four constructs. The instrument also has undergone content and face validity/logical validity process by getting comments and suggestions from the five designated experts. As soon as the instrument has been improved and refined in accordance with the comment and suggestion received from the expert, the instrument is then circulated for the purpose of the pilot study. After all the data has been collected, the data is then analyzed by using the Rasch model through Winstep Version 3.72.3 software.

FINDINGS

The administered pilot study was then analyzed with the use of the Winstep software which is based on the Rasch model approach. The verification or validity has been made on the functionality of the items in terms of (i) unidimensionality, (ii) reliability item and person and separation index, (iii) the suitability of the item fit in measuring the construct, (iv) polarity, (v) standardized residual correlation value, and (vi) the validity of the scale that has been done. The description of each item's functionality verification is as explained below:

Unidimensionality

A unidimensionality or dimension uniformity is an important aspect to be analyzed in order to ensure that the objective of the constructed instrument is achievable. Hishamuddin , Siti Eshah , and Mohd Razimi (2018); Hishamuddin , Siti Eshah, Mohd Razimi , Siti Rahaimah and Ismail Yusuf (2019) assessed unidimensionality assumption using exploratory factor analysis in Item response theory (IRT) but Rasch used Residual Principal Component Analysis (PCA) to ensure the unidimensionality of the developed instrument. The value of the raw variance explained by measures is 49.8 percent, in which it has exceeded the minimum value of 40 percent to meet the Rasch requirements (Fisher, 2007). Meanwhile, the value of unexplained variance in 1st contrast was 8.6 percent less than a ceiling value of 15% (Fisher, 2007; Azrilah, Mohd Saidfudin, & Azami, 2017). This proved that the developed instrument is able to measure in a unidimensionality with an acceptable level of interference.

Reliability and Item and Person, Separation Index

The obtained value of Cronbach's alpha (α) will represent the value of reliability. Based on the suggestion proposed by Bond and Fox (2015), for Rasch model, the Alpha value that is acceptable is between the ranges of 0.71-0.99. Table 1 shows the score interpretation for the Cronbach's alpha value. The privilege of using the Rasch model is that the reliability values are not only focused on the individual but also in

term of the item. There are two types of separation index, which are, person separation index and item separation index. Person separation index represents a rough calculation on the ability of the instrument in dividing each individual to several stages based on the construct that are going to be measured. Meanwhile, the aim of item separation index is to estimate individual abilities by separating item difficulties into several stages in the construct that are going to be measured (Wright & Stone, 1979).

Table 1. *Cronbach's Alpha Score*

Conbach's alpha Score (α)	Reliability
0.9 -1.0	Very good, effective at a high level of consistency
0.7 – 0.8	Good and acceptable
0.6 – 0.7	Acceptable
< 0.6	The Item needs to be refined
< 0.5	The Item needs to be dropped

The Cronbach's alpha value that has been recorded is 0.97 and this means that the developed instrument is relevant and can be used repeatedly. Individual reliability value indicates the probability of repetition in the response results when the same test is performed and the large value of the respondent isolation refers to the separation ability of the studied respondent classification (Azrilah, Mohd Saidfudin, & Azami, 2017). In the analysis of this study, as presented in Table 2, the value of person reliability recorded was 0.96 while the separation value of the respondents was 5.08. The high value of person reliability means that the sample for this study is enough to precisely locate the items on the latent variable.

Item reliability values indicate the adequacy of items to measure what they need to measure, while the item separation indicates the quality of the item where the item's difficulty level can be separated (Azrilah, Mohd Saidfudin, & Azami, 2017). The result of the analysis showed that the reliability of the item is 0.92 and the item separation is 3.42 as shown in Table 2 below. In accordance with Bond and Fox (2015), the item trusted index or good respondents is that when the value is close to the value of 1.0. In addition, Linacre (2019) has stated that a good separation value is over 2.0.

Table 2. *The Item and Person Separation Index*

		Measurement
Person	Mean	1.44
	S.D	.91
	Reliability	.96
	Separation Index	5.08
Item	Mean	.00
	S.D	.53
	Reliability	.92
	Separation Index	3.42

The Suitability of Item Fit in Measuring the Construct

The index value of the Outfit Means Square and Infit Means Square shows the fit of the item in measuring the construct. It will detect the Outfit or misfit item in which the MNSQ value will provide the ratio of observation compared to the expectation (Azrilah, Mohd Saidfudin, & Azami, 2017). In accordance to Bond and Fox (2015), the MNSQ's infit value and MNSQ's outfit value should be between the range of 0.6 to 1.4, in order to ensure that the developed item is appropriate and able to measure the construct. The

z-Std value will show a high value where it will exceeded the range of -2.0 <ZSTD +2.0 if the infit and the outfit value of MNSQ are outside of the specified range.

Based on Table 3, the MNSQ's infit values that are out of specified range is items P2, A38, and A43. Whereas, the MNSQ's outfit values that are out of range are items P2 and A38. The mentioned items have exceeded the value of 1.4 and are considered to be within the range where the item may be considered to be removed. According to Bond and Fox (2015), values that are greater than 1.4 indicated that the item is homogeneous when being put on a scale with other items, while if the value recorded are below the value of 0.6, it indicated that there is overlap between the construct and other items. Thus, researchers had removed P2 and A38 items due to both of the MNSQ values for infit and outfit recorded being outside the specified range, but the item A43 were not dropped because of MNSQ outfit value was in range and the value of z-Std did not exceed range -2.0 < ZSTD +2.0.

Table 3. *Entry Item*

Item	Measurement (logit)	Infit		Outfit	
		MNSQ	Z-STD	MNSQ	Z-STD
P2	1.34	**1.64**	1.64	**1.82**	1.82
A38	.04	**1.63**	1.63	**1.45**	1.45
A43	-.38	**1.52**	1.52	1.16	1.16
K12	.38	1.29	1.29	1.40	1.40
S32	-.60	1.22	1.22	1.34	1.34
A36	-.55	1.34	1.34	1.32	1.32

An analysis has also been made on the table of entry item in order to detect items whether they are within the same dimensions of the same force of measurement or not. Table 4 shows an item having the same value, which is item A44 and A45 with the measurement of -0.22. This shows that the respondent saw the item as measuring the same thing. Therefore, the researcher had decided to drop item A44 and retains item A45 because of the item has a MNSQ value, which is closer to one and the size of z-Std value is closer to zero.

Table 4. *Integrated Misfit Items*

Item	Construct	Measurement (logit)	Infit		Outfit	
			MNSQ	Z-STD	MNSQ	Z-STD
A44	Assessment	**-.22**	1.21	1.0	1.12	.6
A45	Assessment	**-.22**	1.20	.9	1.09	.5

Polarity

The Point Measure Correlation value analysis or PTMEA CORR is to detect the polarity of items, which focus on showing the item is moving in one direction and parallel with the measured constructs. It also describes how far the item is capable of achieving its goals. The PTMEA CORR value may indicate either positive or negative values depending on the state of the item itself. The PTMEA CORR will show a positive value if the item moves in the same direction and is parallel to the measured constructs, while the negative value is shown when developed items are not moving in the same direction and also not in line with the constructs to be measured (Bond & Fox, 2015).

As shown in Table 5 below, all the PTMEA CORR or PMC value that had been obtained from the analysis of the pilot study showed a positive value and none of it showed a negative value. This shows that the entire item is well developed, oriented and paralleled to the constructs that is going to be measured. If

there is a negative value recorded in the PMC reading, the item should be revised and subsequently reconstructed or dropped as the item cannot measure what should be measured (Linacre, 2019).

Table 5. *Point Measure Correlation Value (PMC)*

Entry No.	Item No.	PMC	Entry No.	Item No.	PMC	Entry No.	Item No.	PMC
2	P2	.59	25	S25	.56	30	S30	.67
38	A38	.53	8	P8	.60	15	K15	.66
43	A43	.50	11	P11	.68	34	S34	.45
12	K12	.61	3	P3	.57	28	S28	.70
32	S32	.59	33	S33	.56	22	K22	.68
36	A36	.41	41	A41	.55	20	K20	.74
40	A40	.38	6	P6	.69	17	K17	.66
14	K14	.67	1	P1	.67	19	K19	.80
5	P5	.59	42	A42	.61	24	S24	.58
7	P7	.64	29	S29	.56	13	K13	.76
44	A44	.44	9	P9	.62	35	A35	.60
45	A45	.46	39	A39	.53	26	S26	.67
31	S31	.53	27	S27	.68	10	P10	.69
23	S23	.59	16	K16	.70	21	K21	.76
4	P4	.59	37	A37	.64	18	K18	.68

Standard Residual Correlation Value

The Standard Residual Correlation test is performed in order to make verification whether the instrument that has been developed is free or unrestrained from the uncertainty in the objective or towards the intention of why the examination is conducted. This is to ensure that no items are overlapped with each other and are not singled out (Hashimah, Mohd Isa, & Shahlan, 2018). In accordance with Linacre (2019); Azrilah, Mohd Saidfudin, & Azami (2017), the value that is most suitable and good for the standard residual correlation is less than 0.70 and the item in which the value is above 0.70 are possessing the same characteristics and will cause confusion among the respondent. Linacre (2019) also explains that high correlation values indicated that the item is not singular, has the same characteristics or several dimensions are combined and shared.

Table 6 clearly shows that an item with correlation value greater than 0.70 is P2 and P4. This item needs to be removed or purified to make the question clearer. However for the correlation value of 0.75, item P2 was dropped because its MNSQ outfit and infit values are not within the range and indicated that the item was misfit.

Table 6. *Standard Residual Correlation*

Correlation	Item No	Item No
.75	P2	P4
.67	K19	K20
.64	S29	S34
.62	S33	S34
.59	A40	A43
.58	P3	P4
-0.60	P2	S33
-0.55	P3	S33

Scale Validity

Rasch analysis is capable of calibrating the scale to ensure that the cited data is valid for analysis and processing. A good scale is necessary to ensure that the categories are formed based on the response and in line with the directed scale rating (Lily Hanefarezan, Maimun Aqsha, Ashinida, & Mus'ab, 2018). According to Linacre (2002) the value of the differences in calibration structures should be greater than 1.4 and less than 5. The differences in the value that exceeded the value of five, the rating should be separated and the differences that are less than 1.4 should be incorporated.

Scaling calibration is an important aspect in obtaining data validity. Un-calibrated scales will cause the generated data cannot be used for analysis purposes. Therefore, scale validity should be conducted in order to interpret the collected data and thus enabling the data to be analyzed and generating accurate results (Azrilah, Mohd Saidfudin & Azami, 2017). Based on table 7, the difference of calibration structure for category label 3 and 4 was 0.11 (-0.32-(-0.43)), category label 6 and 7 was 0.53 (-0.24-0.29), category label 7 and 8 was 0.18 (0.29-0.47) and category label 8 and 9 was 1.36 (0.47-1.83) which is overlapping between categories because they are not in between the acceptance range of $1.4<x<5$. Referring to this calculation, the scale for category label 3 and 4, 6 and 7, 7 and 8, and 8 and 9 was proposed to combine. The result of the combined scale has formed a 5-point Likert scale, which will be used in the field studies to compare the scale in pilot test than 10-point scale.

Table 7. *Scale Rating of Calibration Structure for the Scale of 10*

Category Label	The Observed Average	Calibration Structure	The Category Measurement	Differences
1	-.82	None	(-4.62)	
2	-.64	-3.49	-2.22	3.49
3	-.42	-.32	-1.33	3.17
4	-.15	-.43	-.94	**0.11**
5	.18	-2.14	-.57	1.71
6	.54	-.24	-.10	1.9
7	.95	.29	.52	**0.53**
8	1.43	.47	1.39	**0.18**
9	2.07	1.83	3.01	**1.36**
10	2.93	4.03	(5.21)	2.2

DISCUSSION AND CONCLUSION

Based on the analysis derived by using the Rasch model, the process of discarding and refinement of the item has been conducted. There are three items that have been discarded because of the failure to meet the inspection criteria, which led to the final instrument to be used in the field study consists of 42 items. Table 8 shows an overall summary of items being discarded or maintained in this study.

Table 8. *Summary of the Item*

No	Construct	Maintained Item	Discarded Item	Total Item Maintained	Total Discarded Item
1	Knowledge	P1, P3, P4, P5, P6, P7, P8, P9, P10 and P11	P2	10	1
2	Skills	K12, K13, K14, K15, K16, K17, K18, K19,	-	11	-

		K20, K21and K22			
3	Attitude	S23, S24, S25, S26, S27, S28, S29, S30, S31, S32, S33 and S34	-	12	-
4	Assessment	A35, A36, A37, A39, A40, A41, A42, A43, and A45,	A38, A44	9	2
Total				42	3

In this studies, the instruments has gone through several process of validation using Rasch model. After deleting 3 item from the 45 item, the analysis showed that all the 42 item fit the model with unidimensionality when the values of the variance is 49.8 percent and the value of unexplained variance in the first contrast is 8.6 percent. Their MNSQ item is between the range of 0.6 to 1.4 with z-Std value is not exceed range -2.0 <ZSTD+2.0. No item in negative value in Point Measure Correlation Value (PMC) means the item is well developed. Measuring rating scale revealed that mixing the scale by merging them makes the scale more effective compared than the original 10 scales. Conbach alpha was valued at 0.97, all items and person's reliability index was 0.92 and 0.96 which are more than 0.8. Item separation index was 3.42 and person separation index was valued at 5.08 which is greater than 2.0.

In conclusion, validation is very important in developing an instrument. The validity and reliability of the instrument will determined whether the instrument is capable of measuring what is needed to be measured or not to be measured (Noraini, 2013; Ghazali & Sufean, 2018). Hence, with the validity of this instrument, the accuracy of the finding in assessing the competency of TVE teachers in vocational college can produced meaningful measurement.

REFERENCES

Azrilah, A., Mohd Saidfudin, M., & Azami, Z. (2017). *Asas model pengukuran rasch: Pembentukan skala & struktur pengukuran*. Selangor: Universiti Kebangsaan Malaysia.

Bond, T., & Fox, C. (2015). *Applying the rasch model fundamental measurement in the human Sciences*. New York: Routledge.

Fisher, W. (2007). Rating scale instrument quality criteria . Rasch Measurement Transaction 21(1), 1095.

Ghazali, D., & Sufean, H. (2018). *Metodologi penyelidikan dalam pendidikan*. Kuala Lumpur: Penerbit Universiti Malaya.

Hashimah, M., Mohd Isa, H., & Shahlan, S. (2018). Kesahan dan kebolehpercayaan instrumen indeks pemupukan kreativiti dalam pengajaran guru dengan elemen islam (I-CFTI) berdasarkan pendekatan model rasch. *Jurnal Pendidikan Malaysia*, 77-88.

Hasnah, I., & Jamaludin, B. (2017). Kompetensi guru bahasa melayu dalam menerapkan kemahiran berfikir aras tinggi dalam pengajaran dan pembelajaran. *Jurnal Pendidikan Bahasa Melayu*, 56-65.

Hishamuddin , A., Siti Eshah , M., & Mohd Razimi, H. (2018). Detecting item bias in an anatomy & physiology test for nursing students using item response theory. *International Journal of Academic Research in Progressive Education and Development*, 97–109.

Hishamuddin , A., Siti Eshah, M., Mohd Razimi , H., Siti Rahaimah , A., & Ismail Yusuf , P. (2019). Measuring the academic success of students with ASICS using polytomous item response theory. *International Journal of Advanced and Applied Sciences*, 123-129.

Lily Hanefarezan, A., Maimun Aqsha, L., Ashinida, A., & Mus'ab, S. (2018). Kesahan dan kebolehpercayaan instrumen strategi pembelajaran kolokasi bahasa arab : Analisis menggunakan model rasch. *Jurnal Pendidikan Malaysia*, 131-140.

Linacre, J. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7:4 p.328

Linacre, J. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1), 85–106

Linacre, J. (2019). *A user's guide to winstep® ministep* Rasch-Model Computer Program.

Noraini, I. (2013*). Penyelidikan dalam pendidikan*. Selangor: McGraw-Hill Education (Malaysia) Sdn. Bhd.

Unit Pelaksanaan dan Prestasi Pendidikan. (2018). *Laporan tahunan 2017 : Pelan pembangunan pendidikan malaysia 2013-2025*. Putrajaya: Kementerian Pendidikan Malaysia.
Wright, B., & Stone, M. (1979). *Best Test Design*. MESA Press Chicago.