International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 05, 2020
ISSN: 1475-7192

A Scale Calibration for Vocational Learning Styles Instrument

Mazlili Suhaini*, Adnan Ahmad and Normila Mohd Bohari

Abstract--- *Learning styles identify the behaviours and attitudes that decide an individual's preferred method of learning. An overabundance of learning styles literatures is more focused on learners in general and have been conceptually elaborated in the literature, however, few empirical data and few appropriate instruments have been carried out on the learning style of vocational students. This article sets out to analyse the development of The Learning Styles of Vocational Student instrument to routinely identify students' learning style in the vocational field. Hence, the main objective of this paper is to empirically evaluate the rating scale categories in the instrument that is used to evaluate the students' learning styles in the vocational field. Guided by quantitative design, the data was collected from 57 respondents from a vocational college in the North of Malaysia. A scale calibration that is based on the Rash Model analysis specifically analyse the effective rating scale categories. The findings reveal that the instruments function optimally with a four category Likert scale across all the four dimensions, rather than a five-category structure as has been originally intended. For that, it is suggested that the initial five-category Likert scale be modified for the actual study.*

Keywords--- *Rasch Model, Scale Calibration, Vocational, Learning Style.*

I. INTRODUCTION

Learning styles have been debated for years and there are various established instruments that have been developed to measure a student's learning style. Learning styles were evaluated in various ways based on different theoretical models of learning. Felder-Silverman, Honey and Mumford, Kolb, Dunn and Dunn, Entwistle and Myers-Briggs theories reflect the most common frameworks in the field of education. Each model has its own approach such as Dunn & Dunn [1] that focus on influence and stimulus, Riding & Rayner [2] models are more specific to cognitive skills development, Myers [3] focuses more on personality as a whole, while the Felder-Silverman Model focuses more on practical knowledge and skills. Felder & Silverman [4] state that this model is built specifically for analysing the engineering students learning style.

In order to identify the effective measurement scale in research instrument development, one of the models that has been suggested is the Rasch model. The Rasch measurement model was conceptualised by Georg Rasch, a Danish Mathematician in 1960, and is based on the item response theory (IRT), which describes how latent variables are calculated by a scale [5], [6]. IRT is a family of measurement models that are used to measure latent variables. It is a probabilistic model that uses logit as measuring units by transforming ordinal data into interval data where the data can be plotted to a linear scale.

Mazlili Suhaini, Postgraduate Student, School of Education, Universiti Teknologi Malaysia, Skudai, Malaysia.*

E-mail: mazlili.2385@gmail.com

Adnan Ahmad, Associate Professor, School of Education, Universiti Teknologi Malaysia, Skudai, Malaysia.

Normila Mohd Bohari, Postgraduate Student, School of Education, Universiti Teknologi Malaysia, Skudai, Malaysia.

DOI: 10.37200/IJPR/V24I5/PR201783
Received: 13 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

As stated by Bond & Fox [7], the Rasch measurement model converts qualitative observations data to linear measurements. In this model, they consider the respondents to be highly capable, has the potential to answer more questions correctly compared to respondents with a lower ability [7]–[9]. The difficulty of an item depends on the low probability that respondents will affirm the level of approval of the item. Besides that, Rasch converts raw data from scores to logits. This logit will be compared to a linear model for the probability of success. Logit represents the natural log in the form of interval [8], [10], [11].

The Rasch model has first been initiated in educational studies. In general, it has increasingly been used more in scale development studies [5]. Researchers accepted that Rasch provides enough measurement parameters with the ability to: (i) provide a linear scale by translating scores into a probabilistic model using logit as units of measurement, (ii) transform ordinal data into interval data allowing further statistical analysis; interval data are used in the calculation of various analysis to obtain unbiased and precise results, (iii) provide suggestions by its probabilistic model for missing data [7], [12] and [13]. This is because Rasch estimates the likely response of a person to an item by seeing the person's ability and the item's difficulty, (iv) assess the item's quality by identifying misfits or outliers that can be measured by three measures which are the point measure correlation, the infit and outfit mean square and the Z standard, and (v) provide different measures for item difficulty and person ability, which may be grouped or ranked by the item's difficulty or person's ability.

On the same subject, the Rasch model has been used to validate a scale to assess the students' learning styles instrument in the vocational field. However, this article only reports the validity of the rating scale via the calibration analysis of the Rasch scale. For that, the gap that requires the development of the measurement scale for the student's learning styles instrument is clarified for that reason. This is accompanied by the process or methodology and the analysis of data that have been performed using the Winstep software, version 3.72.3.

II. LEARNING STYLES

Learning styles are the easiest way for an individual to learn. [14]. Most likely students do not exclusively possess one style, they may have their own pattern in their learning preferences. A student has a wide range of interests and students with similar interest may have different levels of expertise, therefore, students should not be provided with the same learning service [15]. Nevertheless, a student's learning style is often taken easily as teachers believe that students' can grasp the teacher's lessons and assignments [16].

The way students learn and specifically their learning styles are given less attention particularly in the vocational fields. Felder [17] says that teachers would rather have general learning styles for one subject and in fact, they naturally teach with the standard learning style for most of the subjects. The clash between the most prominent of teaching styles or approaches and the student's learning styles could have a negative impact on the achievement, in particular that of the students.

An important research that characterised engineering students' learning preferences was proposed by Felder & Silverman [4]. Felder & Silverman [4] regarded learning style as a characteristic strength and preferences in the way one obtains information and processes it. Several studies had used this model and it was shown that engineering

DOI: 10.37200/IJPR/V24I5/PR201783
Received: 13 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

students were primarily inclined to the learning styles that were active, sensing, visual and sequential [4], [18]–[23]. However, there was a lack of empirical data on the matter of vocational students.

This article proposes a scale to determine the instrument of a student's learning style in the vocational field on the basis of the gap that has been discussed. The scale development procedure will be discussed briefly in the proceeding section. This is followed by an elaboration on the methodology of research and data analysis which report on the implementation of the Rasch model while evaluating the appropriateness and effectiveness of the rating scale.

III. MEASUREMENT SCALE

This instrument was adapted and modified to fit the four-dimensional with vocational elements found in the inventory Index of Learning Style (ILS) that had been developed by Felder & Soloman [24]. This instrument proposes 35 items in four dimensions as a measurement scale for assessing students' learning styles in the vocational field, as shown in Table 1. These proposed dimensions and items followed a systematic scale development process involving the pooling of items and a process of refining to filter redundant and unnecessary items. Then, a panel of experts approved the chosen dimensions and items by adhering to the item-level content validity index.

A five-point Likert scale was selected in the instrument with a range from 1 (strongly disagree), 2 (disagree), 3 (neither agree or disagree), 4 (agree), to 5 (strongly agree). Accordingly [25], a five or seven Likert scale produced a similar results, while a five scaling was the most frequent used in surveys [26]. For that, researchers decided to use a five-point Likert scale in this study.

Table 1: Instrument Items

Dimensions	**Number of items**
Active	9
Sensing	8
Visual	9
Sequential	9
Total	35

IV. METHODOLOGY

The Rasch model provides multiple empirical evidence such as testing items and person fit in measuring construct, reliability, separation index for item and person, detecting the polarity of an item, and the functionality of the rating scale structure. However, this study reports only the diagnosis of the scale measurement, the applicable remedies and the effect on the overall reliability.

A sample of 60 students was chosen for the pilot study and distributed to the Electrical Technology students in classes after obtaining the approval from the Education Planning and Research Division (EPRD), Department of Technical and Vocational Education (BPTV), College Director, and the lecturer's permission. Confidentiality and

DOI: 10.37200/IJPR/V24I5/PR201783
Received: 13 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

anonymity were ensured by avoiding any linkages between the research data that could reveal the participants' identity. Of the 60 instruments that were distributed, 57 instruments had been returned. The data were then analysed using Winsteps version 3.72.3, on the basic software of a Rasch measurement model.

The Rasch model performs the evaluation on the basis of a sample from the respondents' response to a range of measurement scale. In Rasch, each individual is classified by ability, whereas items are classified by difficulty. The categorisation is the result of the interaction between the person and the item's difficulty, which uses the log's odd values. Rasch converts responses into log odd units, based on the likelihood of success depending on the discrepancy between the person's ability and the item's difficulty. The log's odd units allow the person's ability and item's difficulty to be mapped in a log ruler. The mapping is based on two hypotheses: (i) the more capable (developed) a person is the more likely it is that he or she will approve all the items, and (ii) the easier the items are the more likely they are approved by all the respondents. Rasch model predicts the position of persons and items in a map based on these two assumptions. Rasch is also able to analyse the efficiency of the rating scale structure, which is the bottom line of this study [7], [12].

4.1. Data Analysis

Rasch scale calibration provides empirical evidence to detect whether the respondents understand the scaling labels and are able to differentiate the scale. As stated by Linacre [27], he pointed out four guidelines to diagnose a problem rating scale. Based on Table 2, the first indicator is the observed count which shows the respondents' answers to a given rating scale. A minimum of 10 observations is required in each rating category. When the category frequency is low, the structure's calibration is imprecisely estimated and is potentially unstable. The second indicator is the observed average which shows the pattern of the respondents. The observed average measure should be increased monotonically up the rating scale. Otherwise, the significance of the rating scale is unclear for that data set and it is questionable, and therefore any derived indicators are of doubtful use.

The third indicator is structure calibration which is the strength of the Rasch measurement model. The necessary degree of advance in structure calibration decreases with an increasing number of categories. The advance must be at least 1.4 logits between the structure calibration for a scale of three categories in order to be comparable to two dichotomies, while for a scale of five categories, the advances of at least 1.0 logits between the structure calibrations are needed in order for that scale to be comparable to four dichotomies. Structure calibration also needs to be less than 5.0 logits. If this becomes further apart, the information function decreases in the middle, suggesting that the scale is offering less information on respondents that seems to be better targeted by the scale. The fourth indicator is the probability curves. It will show the ordered categories before and after collapsing any categories. The threshold corresponds with the intersecting points between the curves in the probability curves.

The contravention of these indices suggests a collapse or combination of the rating scale that is involved. Bond & Fox [7] stated that the first guideline for collapsing the rating scale categories is either upward (for example, category 4 collapses into category 5) or downward (for example, category 4 collapses into category 3), which must be logically based on their labels. For instance, it is illogical or irrational to collapse agree and disagree, instead of agree and strongly agree. A second guideline is when there are suggestions on collapsing in both cases. The best way to choose

DOI: 10.37200/IJPR/V24I5/PR201783
Received: 13 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

which is better is to collapse by comparing across each categorisation of the reliability and validity indices for the variable. The higher the reliability for both persons and items are, the best categorisation to be collapsed.

Table 2: Indicators for well-functioning rating scale [27]

Indicators	**Descriptions**
Observed count	Minimum of 10 high and stable observed count
Observed average	Expected to be increase monotonically
Structure calibration	Expected difference between threshold: For a three categories scale, 1.4<X<5 For a five categories scale, 1.0<X<5
Probability curves	Each category is expected to have distinct peak and correspond with the threshold

4.2. Findings

Rasch offers multiple indices that empirically detect the scaling labels by the respondents as stated in Table 2. In this study, the rating scale diagnostics are shown in Table 3 and Figure 1. The first obvious problem in Table 3 is that Category 1 has only 2 (0%) observed count across all prompts. Linacre [27] points that the observed count should be a minimum of 10 observations which is the requirement in each rating category to be declared as a well-functioning category. A low frequency in observation is potentially unstable for the scale rating.

Referring to the second indicator, the observed average shows the pattern of the respondents. It shows that the values are increasing monotonically. Table 3 shows that the response pattern starts from +0.59 logit and moves one way towards +2.70 logit, and it shows that the pattern of the respondents' responses are normal. Structural calibration is the strength of the Rasch model measurement. In this study, all the deviation value between category are greater than 1.0 and less than 5.0, except category 2 and category 3. The difference is 0.34 (-2.07-[-1.73]), which is less than 1.0. A value below 1.0 is assigned to overlap between categories and could not be distinguished by the scales by the respondents.

In Figure 1, we can identify the problematic rating scale of category 1 as its probability curve is redundant and dim by category 1 and category 2. Compared to the other categories, it has no distinct peak. These problems with the rating scale impede interpretation that respondents find it easier to endorse category 2 than to endorse category 1. This is also a sign that category 1 is not functioning well, where the rating scale is not being used by the respondents in the way it has been intended. In this situation, Bond & Fox [7] suggest that the collapsing of the rating scale categories would improve variable construction and interpretation.

DOI: 10.37200/IJPR/V24I5/PR201783
Received: 13 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

4.2.1. Collapsing Categories

As has been explained previously, the initial labels of the rating scale are 1 (strongly disagree), 2 (disagree), 3 (neither agree or disagree), 4 (agree), and 5 (strongly agree). Following the guidelines of Bond & Fox [7], categories should collapse logically. Thus, based on the scale labelling, it is reasonable that the respondents may not be able to clearly distinguish between category 1 and category 2, compared to category 2 and category 3. To be more

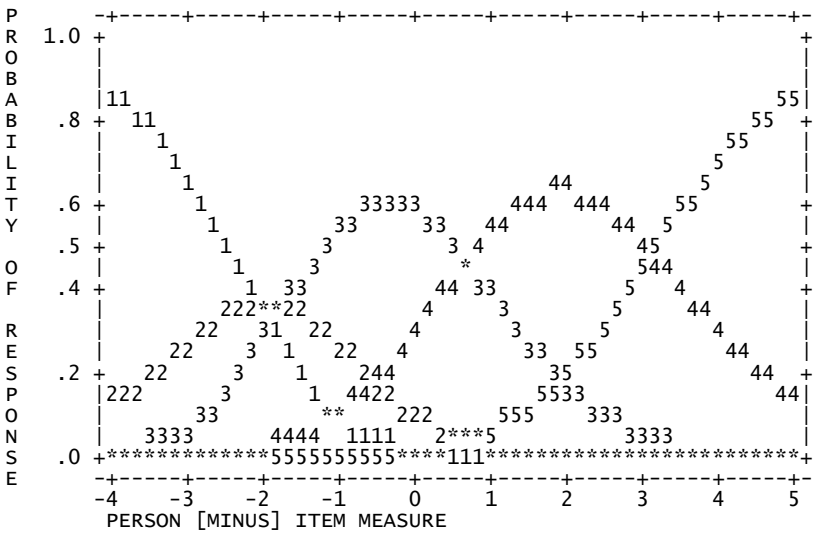

Figure 1: Probability curves for a problematic rating scale (before collapsing)

convincing, the researchers clarify which is the best remedy by comparing collapsing category 1 upward into category 2 and by collapsing category 2 upward into category 3.

In Table 4, the observed count has met the criteria value, where the minimum respondent in category 1 is more than 10 respondents, while the average observed count is steadily increasing. This increase is referred to by Bond & Fox [7] as a monotonic order which represents the well-functioning of each category. The difference between the categories in structure calibration is also above 1.4 and less than 5.0. Therefore, the results show that the use of four categories instead of five is more acceptable. It is proposed based on the comprehensive review from several respondents, that the scaling be relabelled into 1 (disagree), 2 (neither agree or disagree), 3 (agree), and 4 (strongly agree).

The probability curves in Figure 2 indicate there is no redundancies between categories. For that reason, it is more appropriate to use four scaling instead of five scaling. As is claimed by Lozano and his friends [26], the usage between four and seven categories is an ideal number of response category. Besides that, he also points out that a scale is good when it can be discriminated against by the respondents.

In Table 5, the observed count still has not met the criteria value, where the minimum respondents in category 1 is less than 10 respondents, even though the observed average is steadily increasing monotonically. The differences in structure calibration between categories is in the range of $1.0<X<5.0$, except category 2 and category 3 which is more than 5.

DOI: 10.37200/IJPR/V24I5/PR201783
Received: 13 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

Table 3: Diagnostic for problematic rating scale (before collapsing)

Category Label	**Observed Count**	**Observed %**	**Obsvd Avrge**	**Infit MNSQ**	**Outfit MNSQ**	**Structure Calibratn**
1	2	0	.59	1.07	1.09	NONE
2	28	1	.82	1.04	1.05	-2.07
3	430	22	1.25	.98	.97	-1.73
4	1069	54	1.92	1.06	1.09	.67
5	466	23	2.70	.97	.98	3.14

Table 4: Diagnostic for problematic rating scale (after collapsing category 1 into category 2)

Category Label	Observed Count	Observed %	Obsvd Avrge	Infit MNSQ	Outfit MNSQ	Structure Calibratn
1	30	2	.09	1.04	1.05	NONE
2	430	22	.54	.98	.97	-2.38
3	1069	54	1.21	1.06	1.09	-.04
4	466	23	1.99	.97	.98	2.43

Table 5: Diagnostic for problematic rating scale (after collapsing category 2 into category 3)

Category Label	**Observed Count**	**Observed %**	**Obsvd Avrge**	**Infit MNSQ**	**Outfit MNSQ**	**Structure Calibratn**
1	2	0	.94	1.00	.94	NONE
2	458	23	1.43	.99	.98	-4.31
3	1069	54	2.14	1.07	1.10	.93
4	466	23	2.96	.97	.97	3.38

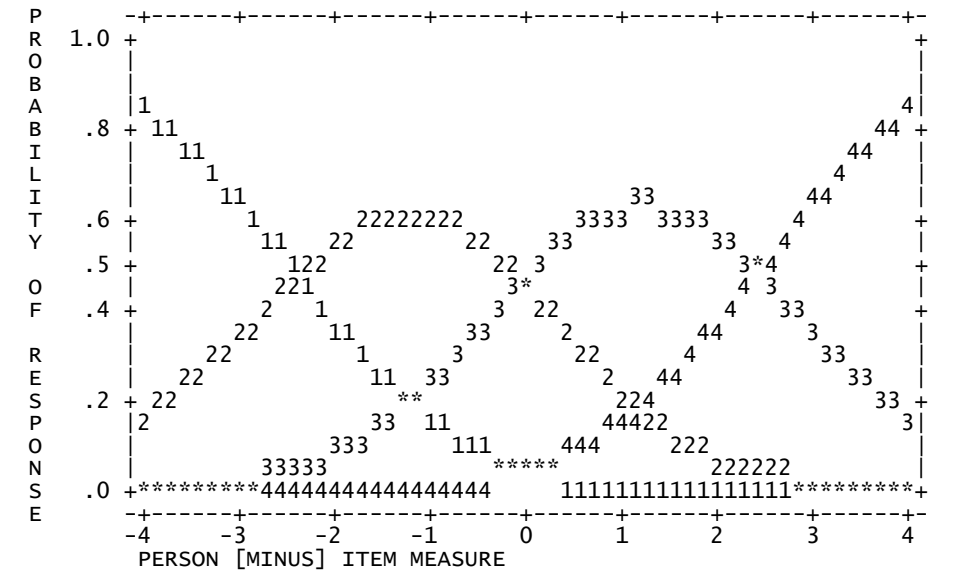

Figure 2: Probability curves for a problematic rating scale (after collapsing category 1 into category 2)

DOI: 10.37200/IJPR/V24I5/PR201783
Received: 13 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

Figure 3 shows category 2 and category 3 are not expected to have distinct peak and correspond with the threshold.

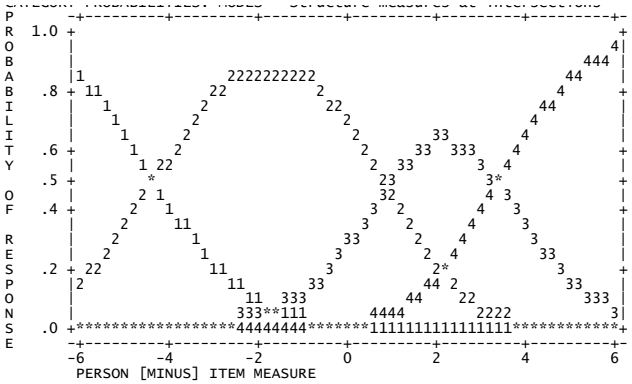

Figure 3: Probability curves for a problematic rating scale (after collapsing category 2 into category 3)

V. DISCUSSION

This is a basic rasch modeling concept. Diagnostic or analysis is performed in order to develop and asses in practice the meaning of a concept. There are no fixed rules that can be followed to which the significane of a test becomes invalid once the arbitary value is approved. Because of that, the results become meaningless. Therefore, it is decisive to remain scientific and empirical about the investigations and explore a handful of categorisations before setting on the preferred one.

Based on the diagnostic in the findings, collapsing each way resulted in monotonic observed average and categories structure calibration. Other than that, collapsing improved category definition in both cases, except for the observed count for collapsing category 2 into category 3. It is clear that collapsing category 1 into category 2 is the preferred one. This was confirmed by assesing the quality of the various reliability and validity indices for the variable. According to Wright & Masters [10], comparing the reliability and validity indices could be another way or methods rather than looking at the indicators of category diagnostic. Table 6 shows that collapsing category 1 into category 2 generates the higher reliability for both persons and items. These indicators have confirmed that the best option is the upward collapse of category 1 into category 2

Table 5: Comparison of three categories

Categories	Person separation	Item Separation
Before collapse	2.64	2.10
After collapse (category 1 into category 2)	2.83	2.17
After collapse (category 2 into category 3)	2.66	2.05

DOI: 10.37200/IJPR/V24I5/PR201783
Received: 13 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

VI. CONCLUSION

The current study provided analyses on diagnosis of the rating scale, the sign, remedy and its effects. Rasch also provides empirical evidence on the design of the rating scale through a summary of the category structure that is supported by the probability curves. Other than that, the precision of the collapsing scale can be measured in the value of reliability and validity indices, for the variables on person and item. This is to confirm which category is more consistent with the theory that has generated the items in the first place. Consequently, whilst the rating scale helps us to determine the best categories, Rasch's knowledge of reliability and validity indices tells us how to measure works as a whole. Therefore, this article proposed that the defined scale should be applied using a four point Likert scale.

ACKNOWLEDGEMENTS

The authors wish to thanks to the Malaysian Ministry for their sponsorship of this study. The authors also express our deepest gratitude to the students, teachers and lecturers who have generously shared their time and thoughtful attention.

REFERENCES

[1] R. Dunn and K. J. Dunn, "Learning Styles/Teaching Styles: Should They... Can They... Be Matched?," *Educ. Leadersh.*, vol. 36, no. 4, pp. 238–244, 1979.
[2] R. Riding and S. Rayner, *Cognitive Styles and Learning Strategies: Understanding Style Differences in Learning and Behaviour*. Routledge Taylor & Francis, 1998.
[3] I. B. Myers, *The Myers-Briggs type indicator : manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1962.
[4] R. M. Felder and L. K. Silverman, "Learning and Teaching Styles In Engineering Educationard," *Eur. Corros. Congr. EUROCORR 2015*, vol. 7, no. June, pp. 674–681, 1988.
[5] R. F. De Vellis, *Scale development: Theory and applications*, 2nd ed., vol. 26. Thousand Oaks, CA: Sage Publications., 2003.
[6] J. Singh, "Tackling measurement problems with item response theory : Principles, characteristics, and assessment, with an illustrative example," *J. Bus. Res.*, vol. 57, pp. 184–208, 2004.
[7] T. G. Bond and C. M. Fox, *Applying the rasch model: Fundamental measurement in the human sciences*, 3rd ed. Mahwah, NJ: L. Erlbaum., 2015.
[8] E. J. Smith, "Metric development and score reporting in Rasch measurement," *J. Appl. Meas.*, vol. 1, no. 3, pp. 303–326, 2000.
[9] G. Rasch, *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Education Research. Expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago, IL: The University of Chicago Press, 1960.
[10] B. D. Wright and G. N. Masters, *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA PRESS, 1982.
[11] J. M. Linacre, *A User's Guide to WINSTEPS MINISTEP*, 3.74.0. 2012.
[12] A. A. Aziz, M. S. Masodi, and A. Zaharim, *Asas model pengukuran Rasch : pembentukan skala & struktur*. Penerbit UKM, 2017.
[13] A. Tennant and P. G. Conaghan, "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 Rheum.*, vol. 57, no. 8, pp. 1358–1362, 2007.
[14] G. Djigic, S. Stojiljkovic, and A. Markovic, "Personality Traits and Learning Styles of Secondary School Students in Serbia," *BCES Conf. Books*, vol. 14, no. 1, pp. 127–134, 2016.
[15] L. Joseph and S. Abraham, "Instructional Design for Learning Path Identification in an E-Learning Environment Using Felder-Silverman Learning Styles Model," *2017 Int. Conf. Networks Adv. Comput. Technol. NetACT 2017*, no. July, pp. 215–220, 2017.
[16] M. Mohamad, Y. Yusof, and N. M. Hanafi, "Connecting Learning Styles and Cognitive Dimension in Building Construction Education," in *Proceeding of the International Conference on Social Science Research*, 2013,

DOI: 10.37200/IJPR/V24I5/PR201783
Received: 13 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020

vol. 2013, no. June 2013, pp. 919–928.
[17] R. M. Felder, "Reaching the Second Tier: Learning and Teaching Styles in College Science Education.," *Journal of College Science Teaching*, vol. 22, no. 5. pp. 286–90, 1993.
[18] N. Omar, M. M. Mohamad, and A. Nazura, "Dimension of Learning Styles and Students ' Academic Achievement," *Procedia - Soc. Behav. Sci.*, vol. 204, no. November 2014, pp. 172–182, 2015.
[19] M. Mohamad, Y. Heong, and N. Kiong, "Disparity of Learning Styles and Cognitive Abilities in Vocational Education," *Int. J. Soc. Hum. Sci. Eng*, vol. 8, no. 1, pp. 8–11, 2014.
[20] E. A. Holt, C. Chasek, M. Shaurette, and R. Cox, "The Learning Styles of Undergraduate Students in CM Bachelor's Degree Programs in the U.S.," *Int. J. Constr. Educ. Res.*, vol. 14, no. 1, pp. 4–21, 2018.
[21] N. Kourakos, L. Karaoglanoglou, D. Koullas, and E. Koukios, "Learning Styles as a Tool for the Education of Chemical Engineers," *EPH-International J. Educ. Res. (ISSN 2208-2204)*, vol. 1, no. 7, pp. 26–35, 2017.
[22] P. K. Tulsi, M. P. Poonia, and Anupriya, "Learning Styles and Achievement of Engineering Students," *IEEE Glob. Eng. Educ. Conf. EDUCON*, vol. 10-13-Apri, no. April, pp. 192–196, 2016.
[23] M. S. Zywno, "A Contribution to Validation of Score Meaning for Felder- Soloman ' s Index of Learning Styles," *Eng. Educ.*, pp. 1–16, 2003.
[24] R. M. Felder and B. A. Soloman, *Index of Learning Styles*. 1991.
[25] J. Dawes, "Do data characteristics change according to the number of scale points used?," *Int. J. Mark. Res.*, vol. 50, no. 1, pp. 61–78, 2008.
[26] L. M. Lozano, E. G.-C. Cueto, and J. Muniz, "Effect of the number of response categories on the reliability and validity of rating scales," *Methodology*, vol. 4, no. 2, pp. 73–79, 2008.
[27] J. M. Linacre, "Investigating rating scale category utility," *J. Outcome Meas.*, vol. 3, no. 2, pp. 103–122, 1999.

DOI: 10.37200/IJPR/V24I5/PR201783
Received: 13 Feb 2020 | Revised: 07 Mar 2020 | Accepted: 23 Mar 2020