International Journal of Evaluation and Research in Education (IJERE)
Vol. 12, No. 4, December 2023, pp. 2362~2373
ISSN: 2252-8822, DOI: 10.11591/ijere.v12i4.27703

Adaptation and evaluation of psychometric properties of HELAM using the Rasch model in the Indonesian context

Trikinasih Handayani[1], Moh Irma Sukarelawan[2], Mohd Effendi Ewan Mohd Matore[3], Much Fuad Saifuddin[1]

[1]Department of Biology Education, Faculty of Teacher Training and Education, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[2]Department of Physics Education, Faculty of Teacher Training and Education, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[3]Research Centre of Education Leadership and Policy, Faculty of Education, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Article Info

Article history:

Received Jun 7, 2023
Revised Oct 6, 2023
Accepted Oct 17, 2023

Keywords:

HELAM
LMS
Rasch model
Teacher professional education

ABSTRACT

Many schools have shifted their learning mode from face to face to fully online in order to reduce virus transmission and avoid learning loss, resulting in accelerated use of the learning management system (LMS). The implementation of online learning over the last 3 years in Indonesia needs to be evaluated. However, the available evaluation scales are in different cultural contexts. Therefore, the existing scales should be adapted. This study aims to adapt and report the psychometric properties of the Hexagonal E-Learning Assessment Model (HELAM) scale in the context of Indonesian culture. The adaptation followed the procedure suggested by Beaton. To achieve a measurement accuracy of up to 0.5 logit at a 99% confidence level, 326 teachers taking the Teacher Professional Education (TPE) program in position program were involved. The psychometric properties of the HELAM scale were analyzed using the Rasch model to improve accuracy and observe the quality of the HELAM scale. The results of the analysis show that the HELAM scale has good psychometric properties in the context of Indonesian culture at the instrument and item level. The elements of reliability and validity have met the good criteria. Thus, the adapted HELAM scale can be applied to the Indonesian context.

Corresponding Author:

Trikinasih Handayani
Department of Biology Education, Faculty of Teacher Training and Education, Universitas Ahmad Dahlan
Tamanan, Banguntapan, Bantul, Yogyakarta 55166, Indonesia
Email: trikinasih@pbio.uad.ac.id

1. INTRODUCTION

Schools at various levels of education have changed the mode of learning during the COVID-19 pandemic, including schools in Indonesia. Most schools have shifted their learning activities from face to face to fully or partially online [1], [2]. This is implemented to anticipate the learning loss [3]. Learning management systems (LMS) is one of the learning platforms chosen because of its easy management. LMS has integrated material management, monitoring of participants and teachers, as well as adjustment of the learning and teaching process [4]. The use of LMS in learning has also become one of the solutions to minimize the spread of the virus in Indonesia and in the world. The COVID-19 pandemic has accelerated the use of LMS which is one of ideal learning platforms for teachers in various countries, including those in Indonesia.

The Indonesian Government has established teacher as a professional occupation in 2005, in which every teacher is required to take a teacher certification program. This program has also been implemented in various countries such as the United States [5]–[7], Brazil [8], Canada [9], China [10], Hong Kong [11],

Turkey [12], and Georgia [13]. This certification program develops role-oriented and competency-based teacher professionals [14]–[16], which leads to continuous professional development [17]–[19].

Various forms of certification programs have been developed [19]. The Ministry of Education and Culture of the Republic of Indonesia has conducted certification programs in various forms, including portfolio (2006–2008), Teacher Professional Education program (2009–2017) [20], and Teacher Professional Education (TPE) program in position (2018–2021) [21]. The Teacher Professional Education Program in Position is conducted in a blended learning mode [22], especially during the COVID-19 pandemic [23].

Although the evaluation of teacher education programs has been carried out by several researchers [24], [25], evaluations of blended learning activities in this program have not been done much. Therefore, it is necessary to evaluate the blended learning program of the teacher education program. Evaluations can involve participants as active actors in the use of LMS. One instrument that can be used in the blended learning evaluation process is the Hexagonal E-Learning Assessment Model (HELAM) instrument developed by Ozkan and Koseler [26].

The HELAM scale can be an alternative instrument in learning evaluation in LMS. User perceptions of LMS during the COVID-19 pandemic in Indonesia need to be examined. While developing new instruments requires a long process. Therefore, an adaptation process is needed. HELAM is a systematic and comprehensive scale of evaluation of the application of LMS in Indonesia because the dimensions used in HELAM combine social and technical aspects. Given that the HELAM scale was developed in different cultural contexts, an adaptation process is needed to be used in the Indonesian context [10], [27].

Adaptation of HELAM to different cultural contexts, including in Indonesia, is key to ensuring its validity, enforceability and relevance, and avoiding cultural biases in research results and data validity. The adaptation process also ensures consistent understanding and response from respondents with diverse cultural backgrounds, maintains openness to cultural diversity, and allows research results to be more applicable and relevant in various cultural contexts. Therefore, the adaptation of HELAM becomes an important stage in producing quality and relevant research in various cultural contexts. The HELAM scale has been used in Indonesia before [28]. However, there are limited research that reports on the quality of the HELAM scale in the Indonesian cultural context.

The study on the quality of the HELAM scale using factor analysis on graduate and post-graduate students at Brunel University, United Kingdom showed that 65 items spread over six dimensions have satisfactory quality, namely the quality of system, service, content, learner perspective, instructor attitudes, and supportive issues [29]. The analysis of the HELAM scale using the factor analysis approach has a number of weaknesses [29]. Factor analysis based on Classical Test Theory (CTT) only describes the quality of the scale at the instrument level, not the level of individual items. Reliability is often limited to Cronbach's alpha. Classical test theory has item dependent and person dependent properties. Another problem that CTT cannot solve is to look at the functionality of the Likert rating scale used in HELAM. Thus, an alternative analysis is needed to evaluate the HELAM scale in a particular case study, in this case the cultural context in an area.

The Rasch model is an alternative analysis that can address the limitations of CTT. In the Rasch model, the raw data obtained are transformed into interval data [30]. Furthermore, Rasch analysis produces a smaller standard error and allows estimation of reliability and validity that is independent of item-dependent and person-dependent properties [31], [32]. The results of the analysis of the Rasch model are able to explain the quality of the scale to the level of individual items [33], [34]. Rasch can be used to see the consistency of the existing response patterns. Rasch analysis provides a very effective alternative for exploring psychometric properties and accounting for response bias [35].

Further, the Rasch model is able to evaluate the functionality of the Likert rating scale used in the scale [36]. More importantly, the Rasch model is a psychometric technique that can improve the accuracy of making instruments, monitor the quality of instruments, and assess the performance of respondents [37]. Therefore, this study aims to evaluate the psychometric nature of the HELAM scale comprehensively in the Indonesian cultural context using Rasch modeling. This research has filled the gap in evaluating the psychometric nature of the HELAM scale for teacher certification programs in Indonesia.

2. RESEARCH METHOD

The implementation of the adapted version of the HELAM scale has been carried out to evaluate the psychometric properties in the Indonesian context. The respondents involved were 326 teachers taking the TPE program in position program at Ahmad Dahlan University, Indonesia. The respondents were from 14 provinces in Indonesia. The location was chosen based on the amount of data needed to meet the sample size. A minimum sample size of 50 respondents can be used to meet the adequacy of data stability at an accuracy of 1 Logit with a 99% confidence level [38]. Meanwhile, the use of sample size in the range of 250 can increase the measurement accuracy up to 0.5 logit at a 99% confidence level [39]. Therefore, the sample size

used is expected to have an impact on data stability and a good model in the Rasch analysis [40]. The characteristics of the respondents are summarized in Table 1.

Table 1. Characteristics of respondents

Variable	Frequency	Percentage (%)
Gender		
Female	258	79.1%
Male	68	20.9%
Age		
≤ 30	55	16.9%
31-35	106	32.5%
36-40	104	31.9%
≥ 41	61	18.7%
Mean	36.2	
SD	6.0	

The adapted version of the HELAM scale has been evaluated for its suitability in the Indonesian context by three language experts through adaptation procedures [41], [42]. The HELAM scale consists of 73 items which are divided into two parts, the first part in the form of demographic data to capture respondent data in general and the second part in the form of learners' LMS experiences. The HELAM scale consists of six dimensions, namely: Learner's Perspective (LP), Instructor Attitudes (IA), System Quality (SyQ), Information Content Quality (ICQ), Service Quality (SeQ), and Supportive Issues (SI) [29]. The second part employs a 4-point Likert scale, from 1 (Strongly disagree) to 4 (Strongly agree). The confidentiality of responses from participants is kept and the participants were given a right to withdraw their responses if they feel uncomfortable. All data is collected online by using Google Forms. The data collection was conducted for about four months using a convenience sampling technique.

The Rasch model was used to evaluate the psychometric properties of the HELAM scale. This model is considered as an extension or alternative of CTT [43]. The Rasch model was chosen because it allows the process of transforming data from ordinal to interval to meet HELAM's assumptions as a measuring tool. A good measuring instrument must have a scale with the same interval distance as a physical measuring instrument in general. For data analysis, we used Winsteps 6.4.1 software [44].

Before evaluating the psychometric properties of the items in HELAM, an initial screening process was carried out on the person and rating scale used. Extreme person data and person that have a Pt. Mea. Corr negative (misfit) were omitted from the analysis. Pt. Mea. Corr negative or zero indicates that the data has an orientation opposite to the prediction of the model. After screening the person, it was continued with the screening of the rating scale used. The rating scale functionality was visually screened based on the rating scale probability curve. Evaluation of the psychometric properties of the HELAM scale began with analyzing Unidimensionality, Reliability, Item fit, and then Differential Item Functioning (DIF). Reliability was reviewed based on separation and reliability on items, persons, and Cronbach's alpha. The Wright map supported construct validity. Item fit used Outfit MNSQ, Outfit ZSTD and Pt criteria. Mea. Corr. Finally, Differential Item Functioning (DIF) was employed to see the bias of items in HELAM towards gender.

3. RESULTS AND DISCUSSION

3.1. Initial screening

Initial screening was carried out to obtain good data quality. Initial screening was conducted on the person and rating scale used. In the screening, 82 (25%) persons were removed from the analysis because of outliers and misfits. The results of the screening rating scale are shown in Table 2.

Table 2 shows that the observed mean value increased, from -0.56 to 3.06. Similarly, Andrich Threshold has increased from None to 3.18. The observed count indicates an increase. The use of the rating scale also shows good functionality of the Likert scale point 1 (Strongly Disagree), 3 (Agree), and 4 (Strongly Agree). However, each index on the Likert scale rating 2 (Disagree) shows that the functionality is not optimal. The functionality of Likert scale rating 2 (Disagree) is visualized in Figure 1. The figure shows that the Likert scale rating 2 (Disagree) does not have an individual peak while the other Likert scale ratings show good functionality. This can be seen from the presence of each peak on Likert scale rating 1 (Strongly Disagree), 3 (Agree), and 4 (Strongly Agree).

Table 2. Rating scale function on HELAM

Category label	Count	Observed average	Infit MNSQ	Outfit MNSQ	Andrich threshold
1. (Strongly disagree)	665	-0. 56	1.13	1.30	None
2. (Disagree)	0	0	0.00	0.00	Null
3. (Agree)	12187	0.77	0.98	0.92	-3.18
4. (Strongly agree)	3008	3.06	0.94	0.86	3.18

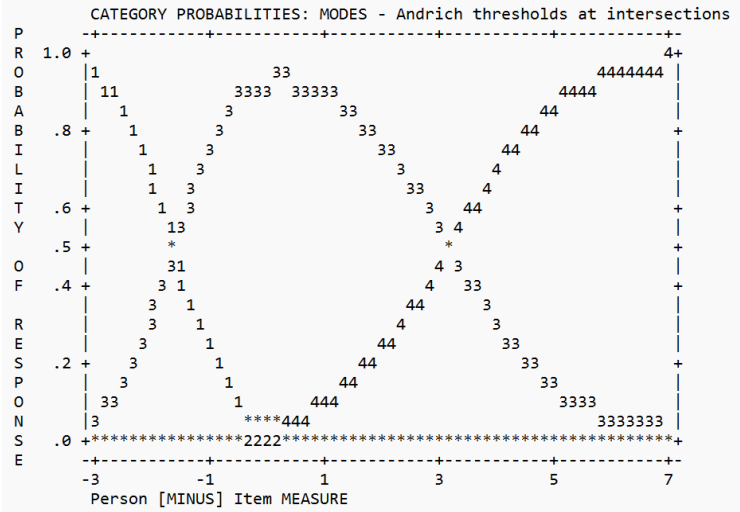


Figure 1. Probability curves of HELAM Likert scale

3.2. Unidimensionality

Unidimensionality of HELAM was determined using Principal Component Analysis of the residual (PCAR). The results of the HELAM unidimensionality evaluation were reviewed based on Raw variance explained by measures and Unexplained variance in 1st contrast. The score of raw variances explained by measure is 31.6% (the expected value is 34.0%) and the Unexplained variance in 1st contrast score is 4.0%. Meanwhile, the Eigen value in Unexplained variance in 1st contrast is 3.78.

3.3. Statistical summary

After conducting the initial screening, an evaluation of the reliability of the HELAM scale was carried out. Reliability was reviewed based on item and person perspective. The results of the reliability analysis based on the Rasch parameters are shown in Table 3.

Table 3 indicates that the HELAM scale in the Indonesian context has good reliability. The reliability of the item and person is 0.97 and 0.94, respectively. Meanwhile, the reliability of the instrument is indicated by the Cronbach's Alpha index, which is 0.95. The index of separation on items and persons shows good reliability, at scores of 5.37 and 4.12. Statistically, the average and standard deviation of items is 0.00 logit and 0.82 logit, while the average and standard deviation of the person is 1.15 logit and 1.34 logit.

Table 3. Statistical summary of HELAM in Indonesian context

	N	Mean	SD	Separation	Reliability	Cronbach's alpha
Item	65	0.00	0.82	5.37	0.97	0.95
Person	244	1.15	1.34	4.12	0.94	

3.4. Wright map

The distribution of item difficulty in HELAM was evaluated visually using the Wright map. Figure 2 visualizes the distribution of item difficulty in logit size. Based on Figure 2, the logit bar stretches from -3 logit to 5 logit. The average logit person is higher than the average logit item. This shows that the average LMS acceptance rate among teachers is higher than the average item difficulty level. Logit items are in the range of -2.66 logit to 1.85 logit, and person logit is from -1.47 logit to 4.63 logit. The item most easily approved by respondents is the SeQ2 item and the item most difficult to agree on is the SyQ4 item.

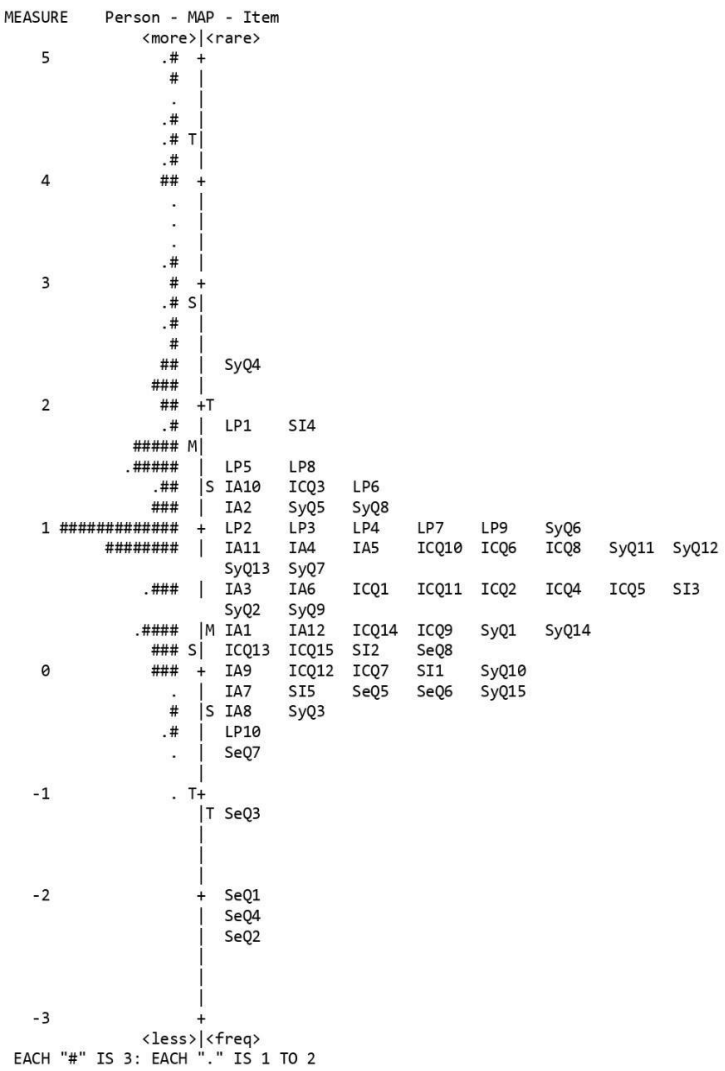


Figure 2. Wright Map of HELAM

3.5. Item fit

The quality of the items in HELAM was reviewed based on their suitability to the MNSQ infit value and the MNSQ outfit. Item matches are summarized in Table 4. The table shows that 19 items on the HELAM scale need to be revised or removed from the HELAM scale. There were five items (SyQ4, SeQ1, SeQ2, SeQ3, SeQ4) spread over the System Quality and Service Quality dimensions have outlier locations. Another 15 items were identified as misfit. In addition, three items (LP1, LP5, LP10) on the Learner's Perspective dimension, three items on the System Quality dimension (SyQ1, SyQ4, SyQ12), five items (ICQ1, ICQ2, ICQ3, ICQ4, ICQ9) on the Information Content Quality dimension, and each of the two items on the Service Quality dimension (SeQ5, SeQ6). The Supportive Issues dimension (SI4, SI5) has an MNSQ outfit index outside the acceptance range (0.50–1.50). Four items (LP1, SyQ4, ICQ3, SI4) of the 15 misfit items had a Pt value. Mea. Corr outside the acceptance range (0.40–0.85). Item match is then reinforced and supported visually through Figure 3.

Table 4. Item fit of HELAM in Indonesian context

Item	Mea. (logit)	S.E.	Diff. level	Item fit index			DIF index		
				Outfit MNSQ	Outfit ZSTD	Pt. Mea. Corr.	Cont.	t	Prob.
LP1[b]	1.39	0.09	Very difficult	3.45	9.90	0.12	-0.48	-2.28	0.06
LP2	0.53	0.12	Difficult	1.29	1.40	0.42	-0.02	-0.07	0.81
LP3	0.42	0.13	Difficult	0.95	-0.19	0.40	-0.60	-2.14	0.19
LP4	0.47	0.13	Difficult	0.93	-0.29	0.49	-0.20	-0.67	0.88
LP5[b]	1.09	0.10	Very difficult	1.66	3.38	0.40	-0.10	-0.43	0.69
LP6	0.87	0.11	Very difficult	1.20	1.11	0.47	0.06	0.21	0.90
LP7	0.42	0.13	Difficult	1.02	0.16	0.49	-0.17	-0.57	0.01
LP8	1.07	0.10	Very difficult	1.10	0.63	0.44	-0.29	-1.27	0.36
LP9	0.42	0.13	Difficult	0.83	-0.80	0.42	-0.36	-1.21	0.41
LP10[bc]	-0.95	0.16	Very easy	2.01	5.09	0.40	1.25	3.37	0.03
IA1	-0.05	0.15	Easy	0.61	-2.10	0.57	0.26	0.70	0.05
IA2	0.61	0.12	Difficult	1.21	1.08	0.45	0.56	1.70	0.07
IA3	0.10	0.14	Difficult	0.86	-0.64	0.53	0.31	0.87	0.66
IA4	0.30	0.14	Difficult	0.88	-0.54	0.52	-0.03	-0.08	0.28
IA5	0.34	0.13	Difficult	0.81	-0.93	0.57	-0.29	-0.94	0.98
IA6	0.08	0.15	Difficult	0.97	-0.08	0.53	0.69	1.87	0.81
IA7	-0.59	0.16	Easy	0.98	-0.07	0.62	-0.15	-0.40	0.88
IA8	-0.66	0.16	Easy	0.95	-0.24	0.63	-0.25	-0.66	0.68
IA9	-0.31	0.16	Easy	1.21	1.04	0.60	0.05	0.14	0.87
IA10	0.93	0.11	Very difficult	0.90	-0.50	0.56	-0.13	-0.51	1.00
IA11	0.24	0.14	Difficult	1.13	0.65	0.55	-0.60	-2.00	0.58
IA12	0.02	0.15	Difficult	1.02	0.19	0.57	-0.29	-0.84	0.81
SyQ1[b]	0.02	0.15	Difficult	0.38	-3.86	0.66	-0.17	-0.49	0.44
SyQ2	0.12	0.14	Difficult	0.58	-2.30	0.62	-0.15	-0.44	0.17
SyQ3	-0.66	0.16	Easy	0.99	0.00	0.62	0.03	0.09	0.71
SyQ4[ab]	1.85	0.08	Very difficult	2.90	9.74	0.20	0.33	1.59	0.30
SyQ5	0.71	0.12	Difficult	0.74	-1.46	0.50	-0.05	-0.17	0.79
SyQ6	0.49	0.13	Difficult	0.85	-0.72	0.53	0.00	0.00	0.28
SyQ7	0.24	0.14	Difficult	0.69	-1.57	0.53	-0.11	-0.32	0.45
SyQ8	0.64	0.12	Difficult	1.29	1.44	0.40	0.10	0.34	0.46
SyQ9	0.10	0.14	Difficult	1.25	1.18	0.60	-0.04	-0.13	0.61
SyQ10	-0.33	0.16	Easy	0.58	-2.43	0.66	-0.11	-0.28	0.31
SyQ11	0.30	0.14	Difficult	0.53	-2.66	0.54	0.31	0.89	0.69
SyQ12[b]	0.32	0.13	Difficult	1.72	2.93	0.43	0.00	0.00	0.23
SyQ13	0.39	0.13	Difficult	0.90	-0.44	0.40	0.78	2.19	0.27
SyQ14	-0.03	0.15	Easy	0.62	-2.02	0.53	0.29	0.78	0.56
SyQ15	-0.48	0.16	Easy	0.80	-1.07	0.56	0.71	1.87	0.56
ICQ1[b]	0.12	0.14	Difficult	0.44	-3.32	0.60	0.47	1.29	0.07
ICQ2[b]	0.10	0.14	Difficult	0.42	-3.49	0.50	0.19	0.52	0.01
ICQ3[b]	0.81	0.11	Very difficult	1.77	3.47	0.19	0.40	1.34	0.06
ICQ4[b]	0.18	0.14	Difficult	0.39	-3.76	0.55	0.04	0.11	0.25
ICQ5	0.12	0.14	Difficult	0.89	-0.47	0.54	0.07	0.20	0.63
ICQ6	0.28	0.14	Difficult	0.53	-2.69	0.56	-0.16	-0.51	0.53
ICQ7	-0.46	0.16	Easy	0.88	-0.57	0.56	-0.13	-0.34	0.38
ICQ8	0.24	0.14	Difficult	0.59	-2.22	0.58	-0.60	-2.00	0.63
ICQ9[b]	-0.09	0.15	Easy	0.34	-4.29	0.68	0.20	0.55	0.16
ICQ10	0.32	0.13	Difficult	0.52	-2.78	0.53	-0.59	-1.99	0.48
ICQ11	0.16	0.14	Difficult	0.55	-2.51	0.60	-0.32	-0.97	0.97
ICQ12	-0.33	0.16	Easy	0.86	-0.65	0.61	0.18	0.47	0.82
ICQ13	-0.26	0.16	Easy	0.61	-2.16	0.62	0.00	0.00	0.80
ICQ14	-0.05	0.15	Easy	0.51	-2.77	0.62	0.40	1.07	0.61
ICQ15	-0.28	0.16	Easy	0.50	-3.36	0.69	0.53	1.39	0.14
SeQ1[a]	-2.42	0.15	Very easy	0.87	-1.06	0.55	-0.28	-0.78	0.29
SeQ2[a]	-2.66	0.15	Very easy	0.70	-2.33	0.56	-0.19	-0.54	0.80
SeQ3[a]	-1.64	0.15	Very easy	1.20	1.64	0.54	-0.07	-0.20	0.82
SeQ4[a]	-2.45	0.15	Very easy	0.79	-1.75	0.54	-0.56	-1.57	0.03
SeQ5[b]	-0.54	0.16	Easy	1.54	2.57	0.53	0.05	0.13	0.31
SeQ6[b]	-0.51	0.16	Easy	1.57	2.64	0.51	-0.04	-0.10	0.92
SeQ7	-1.15	0.16	Very easy	1.36	2.33	0.58	0.29	0.76	0.51
SeQ8	-0.28	0.16	Easy	0.77	-1.16	0.61	0.24	0.64	0.84
SI1	-0.41	0.16	Easy	0.50	-3.13	0.68	0.06	0.17	0.40
SI2	-0.16	0.15	Easy	0.50	-3.25	0.63	0.11	0.31	0.78
SI3	0.10	0.14	Difficult	0.72	-1.39	0.56	0.45	1.22	0.30
SI4[b]	1.45	0.09	Very difficult	3.74	9.90	0.18	-0.04	-0.17	0.85
SI5[b]	-0.64	0.16	Easy	2.01	4.44	0.50	0.06	0.16	0.51

Remarks: [a]=item outlier; [b]=item misfit; [c]=item DIF; Mea.=Measure; S.E=Standard Error; Diff. level=Difficulty level; Cont.=Contrast; Prob.=Probability

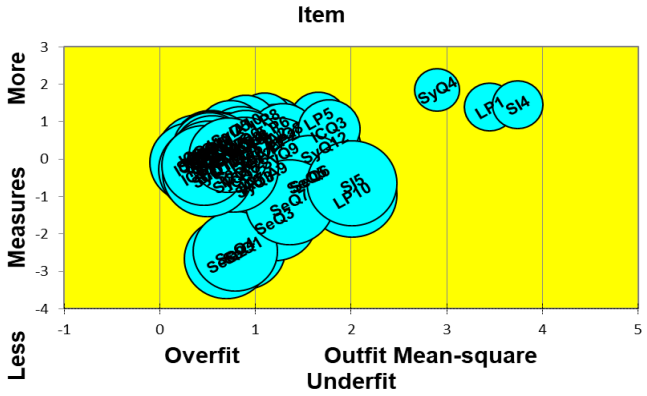


Figure 3. Bubble chart of item match based on Fit statistic type outfit (unweighted) and Mean-square (Chi-squared/d.f.)

3.6. Differential item functioning (DIF)

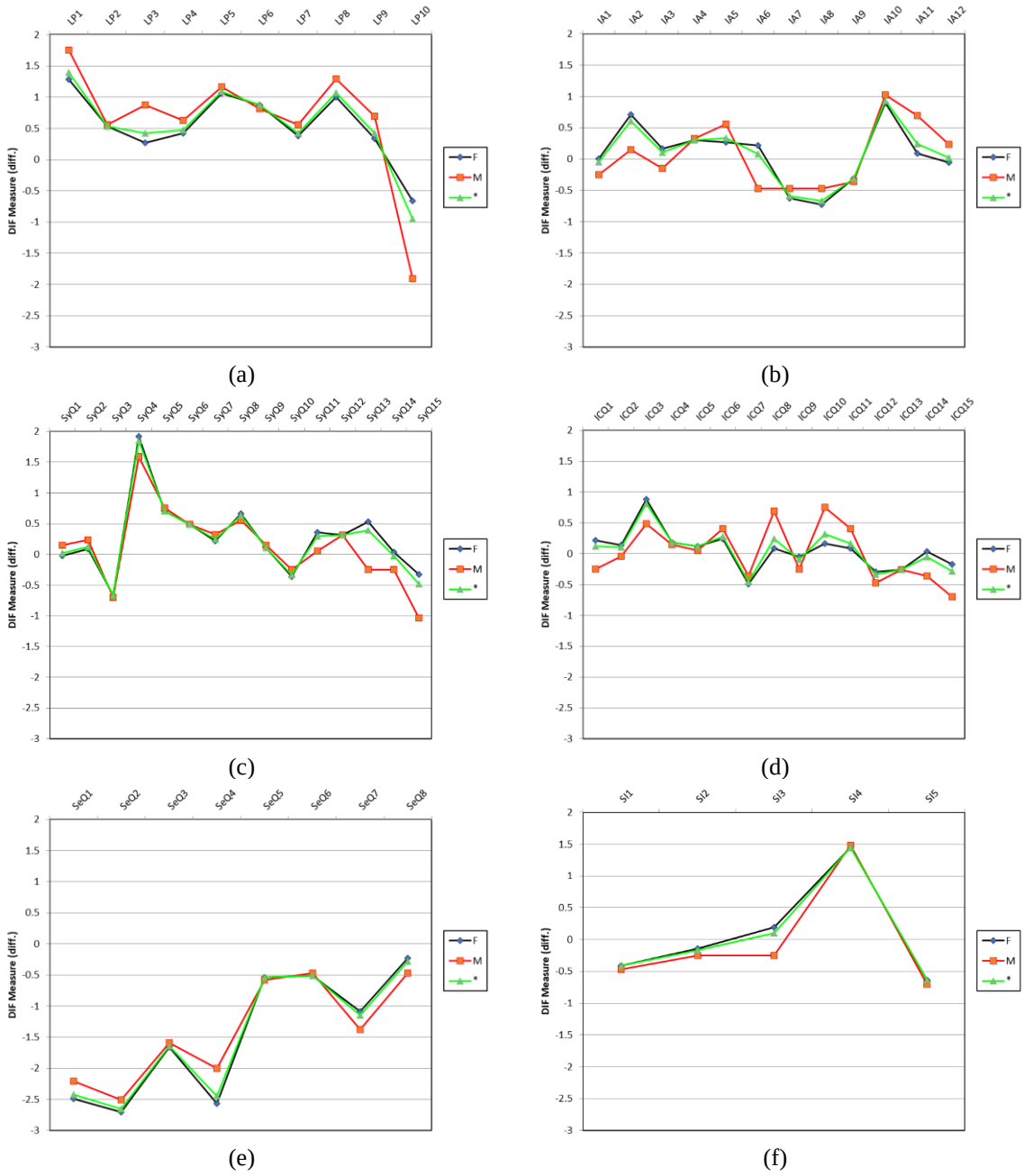
Item bias in HELAM was evaluated through Differential Item Functioning (DIF). DIF Contrast Index, Rasch-Welch, and Mantel Haenszel probability. The results of the DIF analysis of 65 items are shown in Table 4 and visualized in Figure 4.

Based on Table 4, 15 items are indicated to have a DIF on gender. There were four items (LP1, LP3, LP7, LP10) have DIF on the Learner's Perspective dimension as shown in Figure 4 (a). In addition, four items (IA1, IA2, IA6, and IA11) in the Instructor Attitudes dimension as seen in Figure 4 (b) are indicated to be infected with DIF. Meanwhile, on the System Quality dimension presented in Figure 4 (c), SyQ13 and SyQ15 items experience DIF. In the Information Content Quality dimension shown in Figure 4 (d), items ICQ2, ICQ8, ICQ10, and ICQ15 show the presence of DIF. SeQ4 items on the Service Quality dimension displayed in Figure 4 (e) are also indicated to have DIF. In the Supportive Issues dimension as seen in Figure 4 (f), DIF is not indicated based on gender.

The psychometric properties of the HELAM scale were analyzed using the Winstep 4.6.1 software [44]. After screening the person, it was found that 75% of the respondents were used in the analysis because they had a fit response. In addition to screening people, screening was carried out on the 4-point Likert rating scale used. The first psychometric property reported is the Likert rating scale functionality. This rating scale analysis is very important to see whether or not the 4 rating used in the adapted version of the HELAM scale can be understood well by the respondents [45], [46]. The results of the rating scale analysis carried out show that the 4-point Likert rating scale used in HELAM in the Indonesian context is less functional. None of the respondents chose "Disagree". This shows that in the next HELAM implementation, the recommended Likert scale rating is 3-point. The rating scale functionality is visually strengthened in Figure 1. The three ratings shown in Figure 3 show that the probability of 3 ratings has a response probability value of more than 0.5 and indicate that the respondents understand. This has not been reported in the original version which was developed by Ozkan and Koseler [26], [29].

The second psychometric property is unidimensionality. Unidimensionality is used to see whether HELAM measures a single ability or not [47], [48]. Unidimensionality is determined using Principal Component Analysis of the residuals [49]. The raw variance score is close to the expected value in the Rasch model of 34.0%. The value of Unexplained variance in 1st contrast is 4.0%, equivalent to an Eigen value of 3.78, which is greater than criterion 2 [40]. This indicates that there is no measurement noise, but there are items that come from different variables. The results of further investigations prove that the suspected items are not from other variables because the loading value of all items does not exceed 0.60. So, it can be concluded that the adapted HELAM scale measures a single ability.

Furthermore, the reliability assessment of the HELAM scale yields highly encouraging results. A summary statistic is employed to evaluate its reliability, with the findings demonstrating excellent reliability [50]. The person reliability value of 0.96 indicates remarkable consistency among respondents in their responses to each item within HELAM, underscoring the reliability of their feedback. Concurrently, the item reliability score of 0.96 signifies the high quality of the items contained within HELAM, bolstering the overall reliability of the scale [51]. This robust internal consistency is reinforced by Cronbach's alpha value of 0.97, signifying the exceptional alignment between respondents and the scale's items [46]. This reinforces the consistency of the HELAM scale, aligning with the findings reported by Ozkan and Koseler when utilizing a classical test theory approach [29].



Note(s): F=Female, M=Male

Figure 4. DIF plot item in HELAM based on gender in (a) Learner's perspective, (b) Instructor attitudes, (c) System quality, (d) Information content quality, (e) Service quality, and (f) Supportive issues

The separation person index value exceeding 2 [44] demonstrates the HELAM scale's remarkable sensitivity in discerning respondents' abilities. HELAM can proficiently classify respondents' abilities into up to four distinct levels. Simultaneously, the separation item index reveals that the items within HELAM can be categorized into five levels of diversity. This extensive five-level diversity range significantly contributes to a comprehensive understanding of the item difficulties, spanning from very easy to very difficult, with a minimum range of three being necessary to provide respondents with a fair opportunity to answer correctly [52], [53].

An analysis of the distribution of persons and items in the Wright map was carried out to evaluate the construct validity and sensitivity of the HELAM scale. The Wright map was used to map the level of person ability and respondent's ability hierarchically in the same logit scale [54]. Items with the lowest level of difficulty are located on the lower right, and items with the highest level of difficulty are located on the

upper right. Persons with low abilities are placed on the bottom-left side and people with high abilities are placed on the top-left side. The logit ruler stretches from -3 logit to 5 logit. Based on Figure 2, the distribution of the difficulty level of items is quite good because the items are evenly distributed from the easy to the most difficult. However, some items need to be revised. For example, items SeQ1, SeQ2, SeQ3, and SeQ4 need to be modified to increase the level of difficulty of the item so that it is more functional in capturing information along the level of teacher ability. In addition, items that have the same level of difficulty can be modified to have a higher level of difficulty to be more functional in capturing information from high-ability teachers. An even distribution of items at each difficulty level of the item will increase the sensitivity and reliability of the instrument [55] and support construct validity.

The next psychometric properties are items fit. Items that fit will make a significant contribution to the definition of constructs in general [56]. Item suitability was evaluated based on the MNSQ outfit criteria, ZSTD outfit and Pt. Mea Corr. The ideal MNSQ value is 1 and a value of 0.50–1.50 is a good range showing a productive measure [44], [51], [57]. From the 65 items, 19 items were identified as being misfit to the Rasch model because the MNSQ value is out of range. Nineteen items need to be revised for future studies because the value of Pt. Mea. Corr. shows a positive value even though it is outside the acceptance range, 0.40–0.80 [46], [51]. Pt. Mea. Corr. on the acceptance range indicates that all items function in the same direction to predict the latent nature of the HELAM scale [58]. This contributes to a good level of construct validity [32].

A good scale must provide the same probability in answering correctly between groups of subjects. The probability of responding to a statement should not be influenced by subject attributes, such as gender [59]. A scale that gives the probability of answering differently to the subject will cause bias. The item will function differently, and the information obtained will be inaccurate. Measurement bias will cause the test to be invariant or unequal between groups [60]. Bias will have an impact on decreasing the validity of the measurement scale. Item bias in HELAM is evaluated through Differential item functioning.

Differential item functioning analysis was conducted to see the trend of items in HELAM based on gender attributes. The items are free of bias if the DIF contrast value (acceptance value between -0.5 to 0.5), Rasch-Welch t (acceptance value between -2 to 2), and Mantel Haenszel Probability (acceptance value >0.05) are within the assessment range. An item is affected by DIF if all three indexes are outside the receiving range simultaneously. From the 15 items indicated to be infected with DIF, only LP10 items were infected with DIF because the DIF contrast value was outside the range of ± 0.5, the Rasch-Welch t value was outside the range of ± 2, and the probability value was <0.05. When male and female respondents with the same level of ability give different responses, then the item functions differently [61].

Figure 4 (a) visualizes that LP10 items experience gender bias, while Figures 4 (b)–(f) do not show any gender bias. Based on Figure 4 (a), further evaluation of item bias can be carried out. Statements in item LP10 tend to favor males more than females. Meanwhile, this situation is inversely proportional to the other nine items in the Learner's Perspective dimension. Therefore, items that are infected with DIF need to be discarded or maintained by making revisions. This means that 64 adapted items in the Indonesian cultural context can be assessed in the same way for male and female respondents [61]. This information complements psychometric properties that have not been reported in a previous study [29].

4. CONCLUSION

The Rasch model is a subset of modern test theory that can be used to evaluate the psychometric properties of measurement scales. Linguists have recommended the use of the HELAM scale empirically in the Indonesian context. Based on the evaluation of psychometric properties, the HELAM scale has good quality in terms of reliability and validity. The reported psychometric properties of the HELAM scale have supported its use in different cultural contexts in the future. However, some items indicate a misfit and are affected by DIF based on gender. Therefore, they need to be removed from the scale or revised. Empirically, the 3-point Likert rating scale is appropriate for the Indonesian cultural context.

This research has a significant contribution to information on HELAM's psychometric properties and provides new information regarding HELAM's psychometric properties in other cultural contexts, especially in the Indonesian cultural context. However, since collecting new data is not possible in this study, further research is needed to revise and evaluate the psychometric properties of items that are misfit, affected by DIF, and increase item difficulty levels to fill in gaps in item location at a higher level of difficulty. It is recommended that the functionality of 4-point Likert rating scale used in other cultural contexts should be evaluated. Further, modifying the use of a Likert rating scale of more than 4-point should be considered.

ACKNOWLEDGEMENTS

The authors want to express gratitude to Alpendi and Wahyu Nanda Eka Saputra for their contributions to this final analysis, revision, and discussion manuscript. Finally, the authors also like to express gratitude to Universitas Ahmad Dahlan for providing financial support through a professorship candidate research grant No. 98/2023.

REFERENCES

[1] M. Daumiller, R. Rinas, J. Hein, S. Janke, O. Dickhäuser, and M. Dresel, "Shifting from face-to-face to online teaching during COVID-19: The role of university faculty achievement goals for attitudes towards this sudden change, and their relevance for burnout/engagement and student evaluations of teaching quality," *Computers in Human Behavior*, vol. 118, p. 106677, 2021, doi: https://doi.org/10.1016/j.chb.2020.106677.

[2] A. El-Soussi, "The shift from face-to-face to online teaching due to COVID-19: Its impact on higher education faculty's professional identity," *International Journal of Educational Research Open*, vol. 3, p. 100139, 2022.

[3] A. E. Hallin, H. Danielsson, T. Nordström, and L. Fälth, "No learning loss in Sweden during the pandemic: Evidence from primary school reading assessments," *International Journal of Educational Research*, vol. 114, p. 102011, 2022, doi: https://doi.org/10.1016/j.ijer.2022.102011.

[4] B. Sezer and R. Yilmaz, "Learning management system acceptance scale (LMSAS): A validity and reliability study," *Australasian Journal of Educational Technology*, vol. 35, no. 3, pp. 15–30, 2019.

[5] C. E. Feistritzer, "National overview of alternative teacher certification," *Education and Urban Society*, vol. 26, no. 1, pp. 18–28, 1993, doi: https://doi.org/10.1177/0013124593026001003.

[6] J. E. Koppich, D. C. Humphrey, and H. J. Hough, "Making use of what teachers know and can do: Policy, practice, and national board certification," *Education Policy Analysis Archives*, vol. 15, no. 7, 2007, doi: https://doi.org/10.14507/epaa.v15n7.2007.

[7] K. Tobin, "Control of teacher certification in the United States," *Peabody Journal of Education*, vol. 87, no. 4, pp. 485–499, 2012, doi: https://doi.org/10.1080/0161956X.2012.705150.

[8] A. M. Bof, "Distance learning for teacher training in Brazil," *The International Review of Research in Open and Distributed Learning*, vol. 5, no. 1, pp. 1–14, 2004, doi: https://doi.org/10.19173/irrodl.v5i1.172.

[9] M. A. Coulter and P. C. Abney, "A study of burnout in international and country of origin teachers," *International Review of Education*, vol. 55, no. 1, pp. 105–121, 2009, doi: https://doi.org/10.1007/s11159-008-9116-x.

[10] X. Shi and P. A. Englert, "Reform of teacher education in China," *Journal of Education for Teaching*, vol. 34, no. 4, pp. 347–359, 2008, doi: https://doi.org/10.1080/02607470802401537.

[11] K. C. Lai and D. Grossman, "Alternate routes in initial teacher education: A critical review of the research and policy implications for Hong Kong," *Journal of Education for Teaching*, vol. 34, no. 4, pp. 261–275, 2008, doi: https://doi.org/10.1080/02607470802401370.

[12] A. T. Gökçe, "Alternatively certified elementary school teachers in Turkey," *Procedia - Social and Behavioral Sciences*, vol. 2, no. 2, pp. 1064–1074, 2010, doi: https://doi.org/10.1016/j.sbspro.2010.03.150.

[13] M. N. Kobakhidze and M. Akiba, "Teacher Certification Examinations in Georgia: Outcomes and Policy Implications," in *Teacher Reforms Around the World: Implementations and Outcomes (International Perspectives on Education and Society)*, vol. 19, Emerald Group Publishing Limited, Bingley, 2013, pp. 25–51.

[14] R. A. Childs, M. Ross, and A. P. Jaciw, "Initial teacher certification testing: Preservice teachers' experiences and perceptions," *Canadian Journal of Education*, vol. 27, no. 4, pp. 455–475, 2002, doi: https://doi.org/10.2307/1602245.

[15] N. Cooc, "Teaching students with special needs: International trends in school capacity and the need for teacher professional development," *Teaching and Teacher Education*, vol. 83, pp. 27–41, 2019, doi: https://doi.org/10.1016/j.tate.2019.03.021.

[16] D. Potolea and S. Toma, "The dynamic and multidimensional structure of the teachers professional development," *Procedia - Social and Behavioral Sciences*, vol. 180, pp. 113–118, 2015, doi: https://doi.org/10.1016/j.sbspro.2015.02.093.

[17] H. Aiken, C. Varghese, S. Pedonti, M. Bratsch-Hines, and L. Vernon-Feagans, "Targeted Reading Intervention teacher certification: An approach to building and sustaining teacher expertise in rural schools," *Literacy Research and Instruction*, vol. 59, no. 4, pp. 346–369, 2020, doi: https://doi.org/10.1080/19388071.2020.1777230.

[18] S. Fatimah, Y. Tiarina, F. Fitrawati, and A. S. Mira, "English teachers' and lecturers' perceptions of reflective practice through video recording at the teacher certification program," *Studies in English Language Education*, vol. 8, no. 2, pp. 670–689, 2021, doi: https://doi.org/10.24815/siele.v8i2.18931.

[19] A. L. Rose and J. A. Sughrue, "Professional development needs of novice alternatively certified teachers in Florida: teachers' perspectives," *Teacher Development*, vol. 25, no. 4, pp. 449–464, 2021, doi: https://doi.org/10.1080/13664530.2021.1947887.

[20] M. Nurhattati, A. D. Buchdadi, and C. F. Yusuf, "Teacher certification in Indonesia: An education policy analysis," *Universal Journal of Educational Research*, vol. 8, no. 5, pp. 1719–1730, 2020, doi: https://doi.org/10.13189/ujer.2020.080508.

[21] M. Situmorang, S. Gultom, A. Mansyur, S. Gultom, and W. Ritonga, "Implementation of Learning Innovations to Improve Teacher Competence in Professional Certificate Programs for In-Service Teacher," *International Journal of Instruction*, vol. 15, no. 2, pp. 675–696, 2022, doi: https://doi.org/10.29333/iji.2022.15237a.

[22] M. A. Ashraf, S. M. Tsegay, and Y. Meijia, "Blended learning for diverse classrooms: Qualitative experimental study with in-service teachers," *Sage Open*, vol. 11, no. 3, p. 21582440211030624, 2021, doi: https://doi.org/10.1177/21582440211030623.

[23] P. Manghirmalani Mishra, R. Saboowala, and N. Gandhi, "Attitude Prediction of In-service Teachers Towards Blended Learning Using Machine Learning During COVID-19 Pandemic," in *Intelligent Systems Design and Applications: 21st International Conference on Intelligent Systems Design and Applications (ISDA 2021)*, 2022, pp. 1129–1141.

[24] A. Daud, N. Novitri, and M. Hardian, "Evaluation of the Teacher Professional Education Program (PPG): English Students' Perspective," *AL-ISHLAH Jurnal Pendidikan*, vol. 12, no. 2, pp. 185–200, 2020, doi: https://doi.org/10.35445/alishlah.v12i2.286.

[25] F. Fahrurrozi, "Pengaruh kualitas LMS dan prilaku belajar terhadap kualitas belajar mahasiswa PPG," *Jurnal Muara Pendidikan*, vol. 7, no. 1, pp. 131–136, 2022, doi: https://doi.org/10.52060/mp.v7i1.756.

[26] S. Ozkan and R. Koseler, "Multi-dimensional evaluation of E-learning systems in the higher education context: An empirical investigation of a computer literacy course," in *2009 39th IEEE Frontiers in Education Conference*, 2009, pp. 1–6. doi: https://doi.org/10.1109/FIE.2009.5350590.

[27] A. Zaidi *et al.*, "Cultural adaptation and psychometric validation of the Pregnancy Experience Scale–Brief version (PES-Brief) in Pakistani women with antenatal anxiety symptoms," *SSM-Mental Health*, vol. 2, 2022, doi: 10.1016/j.ssmmh.2021.100055.

[28] G. L. Pritalia, S. D. Budiyanto, L. T. Dewi, and S. Kusrohmaniah, "Critical Factor of E-Learning Component using HELAM and AHP," in *MATEC Web of Conferences*, 2018, vol. 218, p. 03020. doi: https://doi.org/10.1051/matecconf/201821803020.
[29] S. Ozkan and R. Koseler, "Multi-dimensional students' evaluation of e-learning systems in the higher education context: An empirical investigation," *Computers & Education*, vol. 53, no. 4, pp. 1285–1296, 2009, doi: https://doi.org/10.1016/j.compedu.2009.06.011.
[30] S. Vogt, S. Petri, R. Dengler, H.-J. Heinze, and S. Vielhaber, "Dyspnea in amyotrophic lateral sclerosis: Rasch-based development and validation of a patient-reported outcome (DALS-15)," *Journal of Pain and Symptom Management*, vol. 56, no. 5, pp. 736–745, 2018, doi: https://doi.org/10.1016/j.jpainsymman.2018.08.009.
[31] A. Hergesell, "Using Rasch analysis for scale development and refinement in tourism: theory and illustration," *Journal of Business Research*, vol. 142, pp. 551–561, 2022, doi: https://doi.org/10.1016/j.jbusres.2021.12.063.
[32] H. L. Lim and Y. P. Poo, "Diagnostic Test to Assess Misconceptions on Photosynthesis and Plant Respiration: Is It Valid and Reliable?" *Jurnal Pendidikan IPA Indonesia*, vol. 10, no. 2, pp. 241–252, 2021, doi: https://doi.org/10.15294/jpii.v10i2.26944.
[33] M. Y. L. Chiu, H. T. Wong, and W. W. N. Ho, "A comparative study of confirmatory factor analysis and Rasch Analysis as item reduction strategies for SAMHSA recovery inventory for Chinese (SAMHSA-RIC)," *European Journal of Psychiatry*, vol. 34, no. 2, pp. 74–81, 2020, doi: https://doi.org/10.1016/j.ejpsy.2020.02.002.
[34] A. G. Merkin *et al.*, "New avenue for the geriatric depression scale: Rasch transformation enhances reliability of assessment," *Journal of Affective Disorders*, vol. 264, pp. 7–14, 2020, doi: https://doi.org/10.1016/j.jad.2019.11.100.
[35] K. D. Bradley, M. R. Peabody, K. S. Akers, and N. Knutson, "Rating Scales in Survey Research: Using the Rasch model to illustrate the middle category measurement flaw," *Survey Practice*, vol. 8, no. 1, pp. 1–11, 2015, doi: 10.29115/SP-2015-0001.
[36] C. Van Zile-Tamsen, "Using Rasch analysis to inform rating scale development," *Research in Higher Education*, vol. 58, no. 8, pp. 922–933, 2017, doi: https://doi.org/10.1007/s11162-017-9448-0.
[37] W. J. Boone, "Rasch analysis for instrument development: Why, when, and how?" *CBE: Life Sciences Education*, vol. 15, no. 4, pp. 1–7, 2016, doi: https://doi.org/10.1187/cbe.16-04-0148.
[38] C. Jong, T. E. Hodges, K. D. Royal, and R. M. Welder, "Instruments to Measure Elementary Preservice Teachers' Conceptions: An Application of the Rasch Rating Scale Model," *Educational Research Quarterly*, vol. 39, no. 1, pp. 21–48, 2015.
[39] J. Linacre, "Sample size and item calibration stability," *Rasch Measurement Transactions*, vol. 7, no. 4, p. 328, 1994.
[40] W. Ling Lee, K. Chinna, and B. Sumintono, "Psychometrics assessment of HeartQoL questionnaire: A Rasch analysis," *European Journal of Preventive Cardiology*, vol. 28, no. 12, pp. e1–e5, 2021, doi: https://doi.org/10.1177/2047487320902322.
[41] J. Muñiz, P. Elosua, J.-L. Padilla, and R. K. Hambleton, "Test adaptation standards for cross-lingual assessment.," in *Educational Measurement: From Foundations to Future*, C. S. Wells, M. Faulkner-Bond, and M. Hambleton, Eds. The Guilford Press, 2016, pp. 291–304.
[42] D. E. Beaton, C. Bombardier, F. Guillemin, and M. B. Ferraz, "Guidelines for the process of cross-cultural adaptation of self-report measures," *Spine*, vol. 25, no. 24, pp. 3186–3191, 2000, doi: 10.1097/00007632-200012150-00014.
[43] N. E. Ismail, N. S. Jimam, M. L. Dapar, and S. Ahmad, "Validation and reliability of healthcare workers' knowledge, attitude, and practice instrument for uncomplicated malaria by Rasch measurement model," *Frontiers in Pharmacology*, vol. 10, p. 1521, 2020, doi: https://doi.org/10.3389/fphar.2019.01521.
[44] J. M. Linacre, *A User's guide to Winsteps-ministep: Rasch-model computer programs. Program Manual 3.68.0*. Chicago, IL, 2009.
[45] D. R. Axon, J. M. Augustine, T. Warholak, and J. K. Lee, "Improving rating scales: applying Rasch analysis to student pharmacists' attitudes towards herbal medications," *Currents in Pharmacy Teaching and Learning*, vol. 11, no. 7, pp. 658–663, 2019, doi: https://doi.org/10.1016/j.cptl.2019.03.015.
[46] B. Sumintono and W. Widhiarso, *Rasch model application for social sciences research*. Trim Komunikata Publishing House (in Indonesian), 2014.
[47] P. Baghaei, "Development and psychometric evaluation of a multidimensional scale of willingness to communicate in a foreign language," *European Journal of Psychology of Education*, vol. 28, no. 3, pp. 1087–1103, 2013, doi: https://doi.org/10.1007/s10212-012-0157-y.
[48] M. Tabatabaee-Yazdi, K. Motallebzadeh, H. Ashraf, and P. Baghaei, "Development and Validation of a Teacher Success Questionnaire Using the Rasch Model," *International Journal of Instruction*, vol. 11, no. 2, pp. 129–144, 2018, doi: https://doi.org/10.12973/iji.2018.11210a.
[49] M. I. Sukarelawan, J. Jumadi, H. Kuswanto, S. Soeharto, and F. N. Hikmah, "Rasch Analysis to Evaluate the Psychometric Properties of Junior Metacognitive Awareness Inventory in the Indonesian Context," *Jurnal Pendidikan IPA Indonesia*, vol. 10, no. 4, pp. 486–495, 2021, doi: https://doi.org/10.15294/jpii.v10i4.27114.
[50] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychological Assessment*, vol. 6, no. 4, p. 284, 1994, doi: https://doi.org/10.1037/1040-3590.6.4.284.
[51] T. G. Bond and C. M. Fox, *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press, 2013.
[52] D. Adams, M. H. J. Tan, B. Sumintono, and S. P. Oh, "Blended Learning Engagement in Public and Private Higher Education Institutions: A Differential Item Functioning Analysis of Students' Backgrounds," *Malaysian Journal of Learning and Instruction*, vol. 17, no. 1, pp. 133–158, 2020, doi: https://doi.org/10.32890/mjli2020.17.1.6.
[53] R. Fisher, A. Perényi, and N. Birdthistle, "The positive relationship between flipped and blended learning and student engagement, performance and satisfaction," *Active Learning in Higher Education*, vol. 22, no. 2, pp. 97–113, 2021, doi: https://doi.org/10.1177/1469787418801702.
[54] M. Kamaruddin and M. E. E. Mohd Matore, "Development and Validation of Psychometric Properties of the 10 IB Learner Profile Instrument (10IBLP-I): A Combination of the Rasch and Classical Measurement Model," *International Journal of Environmental Research and Public Health*, vol. 18, no. 12, p. 6455, 2021, doi: 10.3390/ijerph18126455.
[55] C. Pozo Muñoz and B. Bretones Nieto, "Spanish version of the flourishing scale (FS) on the parents of children with cancer: A validation through Rasch analysis," *Frontiers in Psychology*, vol. 10, no. 1–8, 2019, doi: https://doi.org/10.3389/fpsyg.2019.00035.
[56] W. Rahayu, M. D. K. Putra, D. Iriyadi, Y. Rahmawati, and R. B. Koul, "A Rasch and factor analysis of an Indonesian version of the Student Perception of Opportunity Competence Development (SPOCD) questionnaire," *Cogent Education*, vol. 7, no. 1, p. 1721633, 2020, doi: https://doi.org/10.1080/2331186X.2020.1721633.
[57] D. Andrich and I. Marais, *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer, 2019.

[58] F. Liu, Z. Zhang, B. Lin, Z. Ping, and Y. Mei, "Assessing the psychometric properties of the Chinese return-to-work self-efficacy questionnaire using Rasch model analysis," *Health and Quality of Life Outcomes*, vol. 20, no. 1, pp. 1–10, 2022, doi: https://doi.org/10.1186/s12955-022-01929-7.
[59] M. Ardiyaningrum, L. Badriah, and S. A. Widodo, "Differential item function of gender in the mathematics elementary school tryout test," *Journal of Physics: Conference Series*, vol. 1315, no. 1, p. 012036, 2019, doi: https://doi.org/10.1088/1742-6596/1315/1/012036.
[60] K. Sadeghi and Z. Abolfazli Khonbi, "An overview of differential item functioning in multistage computer adaptive testing using three-parameter logistic item response theory," *Language Testing in Asia*, vol. 7, no. 1, pp. 1–16, 2017, doi: https://doi.org/10.1186/s40468-017-0038-z.
[61] K. Saghafi, S. M. R. Amirian, and M. E. Shirvan, "Differential Item Functioning Analysis of Persian Adaptation of Foreign Language Classroom Anxiety Scale Against Gender," *Human Arenas*, vol. 5, pp. 1–15, 2021, doi: https://doi.org/10.1007/s42087-020-00172-0.

BIOGRAPHIES OF AUTHORS

Trikinasih Handayani is an Associate Professor at University Ahmad Dahlan, Yogyakarta, Indonesia. She was appointed lecturer in the university in 2000 and went on to pursue her graduate studies in education Science at the Universitas Negeri Yogyakarta. She is passionate about improving the quality of teaching and learning of students and their development in schools and in higher education settings. Her research interest in teacher education, biology education, 21st century teaching and learning. She can be contacted at email: trikinasih@pbio.uad.ac.id.

Mohd Effendi @ Ewan Mohd Matore is an Associate Professor of psychometric and educational measurement at the Universiti Kebangsaan Malaysia (UKM) in Bangi, Malaysia. He received his bachelor's degree in Mechanical Engineering (Manufacturing), and master's degree in Technical Vocational and Education Training (TVET) from Universiti Tun Hussein Onn Malaysia (UTHM), Johore in 2005 and 2007, respectively. He obtained his Ph.D. degree in Psychometrics and Evaluation from Universiti Sains Malaysia (USM) in 2015. He has published more than ten high-impact journals and more than hundreds of academic articles in the research areas of measurement and evaluation, instrument development, and items psychometric assessment. His research interests include psychometric analysis, index development and intelligence assessment. He can be reached via email at effendi@ukm.edu.my.

Moh Irma Sukarelawan is Asst. Prof. at Universitas Ahmad Dahlan (Department of Physics Education, Faculty of Teacher Training and Education), Yogyakarta, Indonesia. He obtained his Doctoral degree in the Department of Educational Science, Graduate School, Universitas Negeri Yogyakarta (UNY) in 2023. His research focuses on physics education, misconception, metacognition, and Rasch model. He can be contacted at email: irma.sukarelawan@pfis.uad.ac.id.

Much Fuad Saifuddin is Asst Prof. at Universitas Ahmad Dahlan (Department of Biology Education, Faculty of Teacher Training and Education), Yogyakarta, Indonesia. His research interests, include E-Learning and technology Utilization in learning. His expertise encompasses Blended learning and multimedia technology in biology learning. He can be contacted at email fuad.saifuddin@pbio.uad.ac.id.