Journal of Physics:
Conference Series

PAPER • **OPEN ACCESS**

Using Rasch model for the development and validation of energy literacy assessment instrument for prospective physics teachers

To cite this article: M Yusup 2021 *J. Phys.: Conf. Ser.* **1876** 012056

Journal of Physics: Conference Series **1876** (2021) 012056 doi:10.1088/1742-6596/1876/1/012056

Using Rasch model for the development and validation of energy literacy assessment instrument for prospective physics teachers

M Yusup

Physics Education Department, Faculty of Teacher Training and Education, Universitas Sriwijaya, Jalan Palembang-Prabumulih KM 32 Indralaya, Ogan Ilir, South Sumatera, Indonesia

m_yusup@fkip.unsri.ac.id

Abstract. As the world facing energy crisis, energy literacy should be an important part in preparing prospective physics teachers. This research address the lack of currently available instrument to assess prospective physics teachers' energy literacy. This paper aims to provide a validity argument of the ELA. Two pilot tests were carried out in three universities in Indonesia. In field-test I, the instrument was given to 112 students. The partial-credit Rasch model was applied to the data to examine items and test properties. The instrument was then revised and given to 123 students in field-test II. Through the processes of removing and/or modifying misfit items based on Rasch analyses, a set of 33 items that has a unidimensional construct of energy literacy was empirically established. Instrument validation results successfully provided a validity argument as they met inferences of scoring, generalization, explanation, and extrapolation. Further instructions on how to use results obtained from the instrument are also provided.

1. Introduction

The world is facing the threat of being in a global energy crisis. On the one hand, energy demand of humans as individuals or socially increased. Almost all human activities are very dependent on energy. On the other hand, the use of fossil energy sources causes many problems, both in terms of their diminishing availability and their impact on the environment [1]. To face the threat of an energy crisis, in addition to efforts to find new and renewable alternative energy, it is also important to cultivate citizens' energy literacy through education.

Energy is an important concept in physics so that citizens can make thoughtful decisions related to important social issues such as energy production and use and climate change [2], [3]. Therefore, science education has an important role to prepare young adults now to become future decision-makers regarding energy [4]. At this point, prospective physics teachers have an important and strategic role. However, to measure how literate they are in energy issues, an assessment instrument is needed.

Research on the development of instruments and measurement of energy literacy has been carried out. Based upon the age group of participants, studies that have been carried out vary from children of elementary school [5]–[7] to middle and high school [8]–[18]. However, research on energy literacy of

the prospective teacher has not been conducted so far. This paper aims to present the development and validation of an instrument to measure energy literacy for prospective physics teachers (named ELA).

The development of instruments using Rasch analysis is an iteration process. Revisions to items or rubrics are based on the results of Rasch's analysis, including item fit statistics, item category structure, differential item functioning (DIF), person-item maps (Wright maps), and dimensionality. The fit of the items with the model is determined by looking at the mean square residual (Mnsq) and standardized mean square residual (Zstd). Both Mnsq and Zstd indicate the difference between what is observed and what is expected by the Rasch model [19]. Mnsq is the residual square, while Zstd is the normalized t score of the residual. There are two Mnsq and Zstd for all persons for each item. Thus, there will be four fit statistics, infit statistics (mnsq infit and zstd infit), and outfit statistics (mnsq outfit and Zstd outfit).

Rasch's statistical perspective views person and item as the same parameter. Thus the fit criteria are also the same. From a substantive perspective, people and items are different. To determine the fit statistics of an item, the following steps are performed: 1) checking whether there are point-measure correlations whose values are negative. If there is a negative one, several possible causes: a. error in the answer key (rubrics), if this error occurs then the rubric needs to be fixed; (b) error entering data into MS Excel; (2) checking outfit first before seeing the infit; (3) checking the mean square (mnsq) first before seeing z standardized (Zstd), and (4) checking the underfit first before seeing overfit or negative, because underfit threatens validity more than overfit.

This paper aims to provide a validity argument of the ELA following argument-based approach to validation [20], [21]. In this study, the range of values from 0.70 to 1.30 for mnsq outfit and mnsq infit was determined as a criterion for an item said to be fit the Rasch model (item fit). If an item does not fit the model (misfit), then the item is considered to be repaired or removed. The final decision on whether an item is repaired or removed was taken by considering the overall results of Rasch's analysis as aforementioned.

2. Method

The instrument was developed through five steps adapted from Kuo, Wu, Jen, & Hsu [22]. First, developing an assessment framework: analyzing competencies and components of energy literacy and organizing them. Items developed following Yusup's et al. framework [23]. Second, designing items: designing assessment instruments to measure the literacy of all components and the level of complexity in the framework. Items were in the Indonesian language. Third, developing scoring rubrics: developing an output space based upon an assessment framework for scoring guides for each item. Fourth, conducting pilot testing and field tests: collecting validity evidence to support the theoretical basis of the construct. And fifth, applying the Rasch model: using the Rasch measurement model to link the score data with the energy literacy component specified in the assessment framework.

Before the ELA was field-tested, it was given in the written form (booklet) to 10 participants in pilot tests. Following the written test, the participants then were interviewed. The interviewees were given the answer sheets they had been working on. Questions are given to explore how their thinking process when answering each item so that it can produce answers as stated in the answer sheet. The results of the pilot test were considered to revise the items, specifically in wording.

2.1. Participants

During the development of ELA, this study involved prospective physics teachers from different levels/semester and institutions. The participants were from three state universities in two provinces in Indonesia; from the first year to the third year of their study. The participants were chosen based upon consideration of differences in geography and cultural groups. In field-test 1, ELA was given to 62 students. ELA was then revised and given to 123 students in field-test 2. The participants involved in each stage of field-testing are different people.

2.2. Data analyses

Rasch partial credit model was used to analyze data, with the help of Winsteps software [24]. The results of the Winsteps operation are used to provide evidence of validity. The evidence includes the category fit, dimensionality, item fit, reliability, and separation for items and person, Wright map, and differential item functioning (DIF).

Data were analyzed by evaluating empirical and theoretical evidence for the next steps, whether the item is removed or repaired. The results of the analysis of pilot test data form the basis of action on problematic items for further testing in field-test 1. Likewise, the results of field test 1 data analysis form the basis of actions on problematic items for field-testing 2 as a final version of ELA. To provide a validity argument, the final field test data were analyzed to provide evidence of scoring, generalization, explanation, and extrapolation inference.

3. Results

The final version of ELA consists of 33 items originating from revised items in the field-test 1. There are five themes included in ELA: light bulbs (BL), photovoltaic (PV), air conditioning (AC), energy conservation (KE), and energy teaching (PE). Figure 1 presents the ELA parameters. The person's mean measure was 0.39 logit, upper the item mean difficulty (which set at 0,0 logit). It means that the ELA was quite easy for the students.

3.1. Evidence for validity argument

3.1.1 Scoring Inference. The score category of the items where the average abilities are disorder is shown in Table 1. The problematic items in the score category are Item BL1, PV2, AC1, AC4, and PE1. All mnsq infit values and mnsq outfits are not problematic because none are greater than 2. Items BL1, PV2, AC1, and AC4 each have four score categories (0, 1, 2, 3). The category 1 score of Item BL1 was disordered. The average ability for score category 1 is lower than the average ability for score category 0. The difference in ability in the two categories is 0.19 logit. Category score 2 in PV2 was also disordered, with a difference of 0.16 with average ability in the score category 1. Item AC4 was disordered in category score 1 and category score 3. The difference in the average ability on score 1 and score 3 was 0.23 and 0.53 respectively from each of the lower score categories. Category score 1 of Item AC4 was disordered in the average ability with a difference of 0.09 logits from the average ability in the score category 0. Category score 2 of Item PE1 was also disordered with a difference of 0.09 logits from the average ability score category 1.

Table 1. Disordered steps (score category) of some items (shown with an asterisk (*))

Item	Score category	Frequency	Average measure	*Infit mnsq*	*Outfit mnsq*	*Threshold*
BL1	0	13	0,33	1,34	1,17	*None*
	1	58	0,13*	0,74	0,79	-1,61
	2	22	0,76	0,61	0,53	1,22
	3	27	1,08	0,87	0,87	0,38
PV2	0	17	-0,06	0,70	0,76	*None*
	1	28	0,67	1,23	1,13	-0,68
	2	5	0,51*	1,21	0,84	1,86
	3	25	1,10	0,95	0,90	1,19
AC1	0	14	0,28	1,31	1,44	*None*
	1	35	0,05*	0,79	0,80	-1,64
	2	68	0,76	0,73	0,77	-1,00
	3	5	0,23*	1,55	1,15	2,64
AC4	0	2	0,18	1,26	1,28	*None*
	1	30	0,09*	1,00	1,06	-2,11
	2	55	0,58	0,90	0,97	0,35
	3	35	0,72	1,10	1,07	1,76
PE1	0	67	0,31	1,01	1,01	*None*
	1	28	0,86	0,82	0,69	-0,32
	2	9	0,77*	1,20	1,33	0,32

Table 2 shows all mnsq infit values and mnsq outfits after collapsing the disorder score categories. The score category 1 in Item BL1 was still disordered, but the difference in average ability with a lower score category is no more than 0.5 logit. Five of the final version of ELA items, that are Item BL1, PV2, AC1, AC4, and PE1 were disordered in the average monotonic of testee ability. For a small number of samples, statistical differences in the ability of less than 0.5 logits are still statistically acceptable [25]. This disorder can be caused by "accidents" in the sample. Also, the more complex and open-ended an item, the more difficult it is to anticipate various possible responses (answers) and to develop fair and explicit scoring criteria that can be applied to all responses [26]. It can be concluded that the score categories in the rubric for all ELA items function as expected.

Table 2. Score category after collapsing score 1 and score 2 for *Item* BL1 and PV2.

Item	Score category	Frequency	Average measure	*Infit mnsq*	*Outfit mnsq*	*Threshold*	
BL1	0	13	0,35	1,16	1,17	*None*	0
	1	80	0,32*	0,87	0,87	-1,66	1+2
	2	22	1,06	0,85	0,84	1,66	3
PV2	0	17	-0,04	0,76	0,75	*None*	0
	1	33	0,66	0,74	0,70	-0,66	1+2
	2	25	1,08	0,85	0,85	0,66	3

3.1.2 Generalization inferences. The analysis was conducted to obtain evidence of generalisability, namely separation index and person reliability; and item separation and reliability. Reliability analysis was carried out to determine the stability and internal consistency of ELA. In field-testing 2 no testee obtained extreme scores (maximum or all true score or minimum or all false score). Figure 1 shows the final version of the ELA person separation index was 2.10 and the person's reliability was 0.81. Cronbach alpha value was 0.85. These parameters are in a good category. Separation and item reliability were 6.39 and 0.98, respectively. It is natural to find a greater separation number for items than persons because the number of items is smaller than persons [27].

The separation index and person reliability imply that ELA is sensitive enough to distinguish between high- and low-ability testees. Separation and item reliability imply that the sample person is sufficient to confirm the hierarchy of the item difficulty level. The results of the separation and reliability analyses above provide evidence of the generalisability of the ELA items.

3.1.3 Explanation inference. Table 3 shows the item statistics of ELA. The largest outfit and mean square infit values are Item KE1 (mnsq outfit = 1.15 and mnsq infit = 1.18). While the smallest outfit value and mean square infit is Item SE9 (mnsq outfit = 0.70 and mnsq infit = 0.80). The largest Zstd value is item KE1 (outfit Zstd = 1.3 and infit Zstd = 1.7) and the smallest Zstd is in Item SE1 (outfit Zstd = -1.4 and infit Zstd = -1.3). All items in Table 3 have mnsq outfit and infit values within the criteria range of 0.70 to 1.3. Likewise, there are no items that have a Zstd value of more than 2.0 or less than -2.0. These statistics show that all ELA items fit the Rasch model.

Fit is the core of Rasch measurement [28]. After going through field testing stages twice, all items in the final version of ELA are fit with the Rasch model which shows that this instrument is trustworthy [19], [29].

SUMMARY OF 123 MEASURED PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	43.2	30.0	.34	.27	1.00	.0	.99	.0
P.SD	10.6	2.7	.66	.02	.24	.9	.27	1.0
S.SD	10.6	2.7	.67	.02	.24	.9	.27	1.0
MAX.	67.0	33.0	1.74	.35	1.83	2.6	2.02	3.2
MIN.	19.0	21.0	-1.39	.24	.42	-2.6	.45	-2.1

REAL RMSE .29 TRUE SD .60 SEPARATION 2.10 PERSON RELIABILITY .81
MODEL RMSE .27 TRUE SD .60 SEPARATION 2.21 PERSON RELIABILITY .83
S.E. OF PERSON MEAN = .06

PERSON RAW SCORE-TO-MEASURE CORRELATION = .96
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .85 SEM = 4.13

SUMMARY OF 33 MEASURED *ITEM*

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	160.9	111.7	.00	.15	1.00	.1	.99	.0
P.SD	66.7	14.2	1.03	.04	.09	.8	.10	.8
S.SD	67.8	14.5	1.05	.04	.09	.8	.11	.8
MAX.	268.0	123.0	2.21	.23	1.18	1.6	1.15	1.3
MIN.	30.0	71.0	-1.93	.11	.80	-1.3	.70	-1.4

REAL RMSE .16 TRUE SD 1.02 SEPARATION 6.39 *ITEM* RELIABILITY .98
MODEL RMSE .16 TRUE SD 1.02 SEPARATION 6.51 *ITEM* RELIABILITY .98
S.E. OF *ITEM* MEAN = .18

ITEM RAW SCORE-TO-MEASURE CORRELATION = -.59

Figure 1. Summary of person and item measures

Table 3. Item statistics of ELA

Item	Measure	*Infit*		*Outfit*		*Ptmeasur-al*	
		Mnsq	*Zstd*	*Mnsq*	*Zstd*	*Corr.*	*Exp.*
SE9	1,80	0,80	-1,3	0,70	-1,3	0,54	0,36
PE1	1,71	1,06	0,4	1,04	0,3	0,34	0,36
BL7	1,36	0,99	0,0	0,95	-0,2	0,52	0,50
PE2	1,31	1,09	1,1	1,12	1,0	0,16	0,29
KE3	1,29	1,01	0,1	0,97	-0,2	0,49	0,45
SE7	1,04	1,10	0,8	1,10	0,8	0,40	0,48
SE1	0,94	0,86	-1,3	0,86	-1,4	0,54	0,39
AC2	0,93	1,09	0,6	1,13	0,7	0,51	0,57
AC1	0,81	1,10	0,8	1,10	0,8	0,32	0,42
BL5	0,70	0,89	-0,9	0,85	-1,1	0,60	0,53
PV2	0,52	0,92	-0,6	0,90	-0,6	0,60	0,53
SE8	0,29	0,96	-0,4	0,94	-0,4	0,52	0,49
BL1	0,23	0,93	-0,6	0,89	-0,9	0,52	0,49
KE6	0,16	1,11	1,1	1,11	1,0	0,34	0,44
AC3	0,09	0,90	-1,0	0,89	-1,0	0,49	0,38
PE2A	0,03	0,98	-0,1	0,98	-0,1	0,39	0,37
PE3A	-0,06	1,16	1,4	1,11	0,8	0,36	0,47
SE3	-0,21	0,98	-0,2	0,98	-0,2	0,38	0,35
BL3	-0,27	0,99	0,0	1,06	0,4	0,39	0,39
SE4	-0,43	0,95	-0,4	0,96	-0,3	0,44	0,38
BL2	-0,44	1,11	0,8	1,06	0,3	0,47	0,52
PV3	-0,55	1,17	1,7	1,13	1,1	0,07	0,28
PV5	-0,55	0,97	-0,2	0,94	-0,5	0,42	0,37
AC4	-0,64	1,06	0,6	1,08	0,7	0,35	0,42
SE5	-0,64	0,99	-0,1	1,00	0,0	0,36	0,34
KE4	-0,69	1,04	0,4	1,04	0,4	0,21	0,28
PE3	-0,70	1,00	0,1	0,94	-0,3	0,46	0,47
BL6	-0,93	0,98	-0,1	0,89	-0,6	0,31	0,26
KE1	-1,06	1,18	1,7	1,15	1,3	0,29	0,43
PE3B	-1,24	0,88	-1,2	0,86	-1,3	0,50	0,33
PE4	-1,41	0,94	-0,5	0,95	-0,4	0,38	0,29
KE7	-1,61	0,96	-0,1	0,94	-0,2	0,31	0,27
KE2	-1,80	1,01	0,2	1,02	0,3	0,29	0.31

Note. Mnsq = mean square; Zstd = Z standardized; Ptmeasur-al = point-measure biserial, Corr. = correlation, Exp. = expected.

Unidimensionality is the basis of the Rasch model. ELA items are designed with unidimensional assumptions. This dimensionality analysis was carried out to see whether ELA is truly unidimensional. Figure 2 shows that the unexplained variance in 1st contrast is 2,7749. The number informs that the ELA item is indicated to have a secondary dimension equivalent to three items. The items indicated to cause the secondary dimensions were Item AC4, KE5, and AC3. However, an examination of the contents of these items concluded that these items were still included in the energy literacy domain.

	Eigenvalue	Observed	Expected		
Total raw variance in observations	=	53.2072	100.0%		100.0%
Raw variance explained by measures	=	20.2072	38.0%		38.5%
Raw variance explained by persons	=	8.4402	15.9%		16.1%
Raw Variance explained by *items*	=	11.7670	22.1%		22.4%
Raw unexplained variance (total)	=	33.0000	62.0%	100.0%	61.5%
Unexplned variance in 1st contrast	=	2.7749	5.2%	8.4%	
Unexplned variance in 2nd contrast	=	2.2909	4.3%	6.9%	
Unexplned variance in 3rd contrast	=	2.0359	3.8%	6.2%	
Unexplned variance in 4th contrast	=	1.7845	3.4%	5.4%	
Unexplned variance in 5th contrast	=	1.5806	3.0%	4.8%	

Figure 2. Standardized residual variance in eigenvalue units.

3.1.4 Extrapolation Inference. As previously noted that ELA has an item separation index of 2.10. The average person ability of 0.39 indicates that the ability of the sample is above the level of difficulty of the item. Figure 3 shows the item-person map (Wright map). On the Wright map there appear to be four gaps. The first gap is between Item PE1 and Item BL7 and the same level. The largest distance from the gap is 0.41 logit. The second gap is between Item PE2 and those in the same level as Item SE7, which the gap is 0.25 logit. The third gap is between Item AC4 and the same level as Item BL6 and KE1. The largest distance from the gap is 0.41 logit. The fourth gap is between Item PE4 and KE7, which is 0.20 logit. All of the intervals on the Wright map are less than 0.50 logit so that they are within an acceptable range.

Figure 3 also shows several items are redundant. Redundancy of item difficulty spans that are above 1 logit (Item BL7, KE3, and PE2), between 0 - 1 logit (Item AC2, SE1, and SE7; Items AC3, BL1, and KE6; PE2A and PE3A), below 0 logit (Item BL2, PV3, PV5, and SE4). Other redundant items are Item AC4, KE4, PE3, and SE5.

The redundant items measure different levels of thinking and knowledge domains. For example, in the last-mentioned pairs, Item AC4 measures at the level of "Knowledge use: problem solving" and "Knowledge related actions". Item KE4 measures at the level "Self-system: testing emotional responses" and "Knowledge related to actions". Item PE3 measure at the level of "Self-system: testing the importance and" PCK ". Meanwhile, Item SE5 measures at the level of "Analysis: specifying" and "System knowledge". The Wright map in Figure 3 also shows the highest position of the item (Item SE9) parallel to the top position of the person. This informs that ELA items can target people well to measure their energy literacy. This means that the ELA item can measure all levels of energy literacy in prospective physics teacher students.

One of the advantages of developing instruments using the Rasch model is that it can be used more widely by other parties because of the independent nature of the person and item [30]. The final version of the Wright ELA map still leaves some gaps between items. To improve the instrument, if the gap more than 0.5 logit [31], [32], test developer can fill in the gap by adding new items that are more difficult than the items underneath [33]. Of the two intersections on the Wright Map in Figure 3, only the gap between Item PV3 and SE9 is more than 0.5 logit. Redundant items can be retained for some reason, what is important is to be aware of the items that are redundant [33].

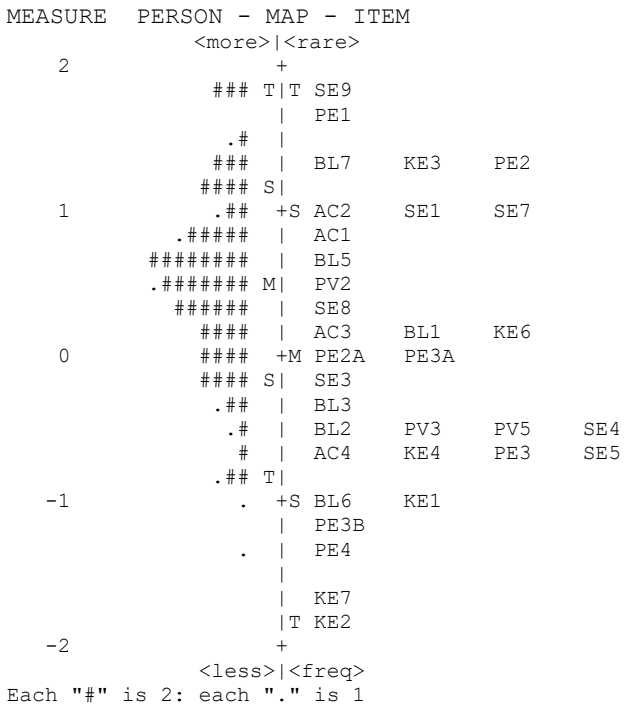

Figure 3. Wright maps for ELA

3.2. Using the ELA

The scores obtained from the administration of ELA to the prospective physics teachers describe the "level" of their competence. Categorizing competencies into certain levels was done by determining boundaries between categories. The category boundary that can be determined from data collected using ELA is the boundary between category 1 and category 2, as well as between category 2 and category 3. In other words, competencies measured using ELA are categorized into three levels of energy literacy. Table 4 shows the conversion of raw scores to Rasch scale scores within the mean of 500 as produced by Winsteps software. The Rasch scale obtain then match to the Wright map in Figure 4 to see the level of energy literacy.

4. Conclusion

The final version of ELA consisted of 33 items along with the scoring rubric that has tested using the Rasch model. The test results show all items meet the criteria of fit with the Rasch model and have evidence to support the instrument validity argument. The rubric developed also has a score category that is fit for the level of student ability. ELA can measure the energy literacy of prospective physics teacher students at each ability level. ELA can also be used without having to analyze the Rasch model using certain software. ELA users can convert the test results in the form of raw scores into a logit scale using the conversion table provided. The location of the testee's energy literacy level on a logit scale can be found on the Wright map. Furthermore, the location on the Wright map can be used to determine the energy literacy competency of the testee.

Table 4. Conversion between raw score and rasch model for ELA

Score	Measure	S.E	Score	Measure	S.E	Score	Measure	S.E
4	-220E	186	30	395	28	56	560	25
5	-91	106	31	403	27	57	566	25
6	-9	79	32	410	27	58	572	25
7	43	67	33	417	27	59	579	26
8	82	59	34	424	26	60	586	26
9	114	54	35	431	26	61	593	26
10	142	50	36	438	26	62	600	27
11	166	47	37	445	26	63	607	28
12	187	45	38	451	25	64	615	28
13	206	43	39	458	25	65	623	29
14	223	41	40	464	25	66	632	30
15	239	39	41	470	25	67	641	31
16	254	38	42	476	25	68	650	32
17	268	36	43	482	25	69	661	33
18	280	35	44	488	25	70	672	34
19	293	34	45	494	24	71	684	36
20	304	33	46	500	24	72	698	38
21	315	33	47	506	24	73	713	40
22	325	32	48	512	24	74	729	43
23	335	31	49	518	24	75	751	47
24	345	31	50	524	24	76	776	53
25	354	30	51	530	24	77	808	61
26	363	29	52	536	24	78	852	74
27	371	29	53	542	24	79	027	103
28	379	28	54	548	25	80	1051E	184
29	387	28	55	554	25	81	1135E	192

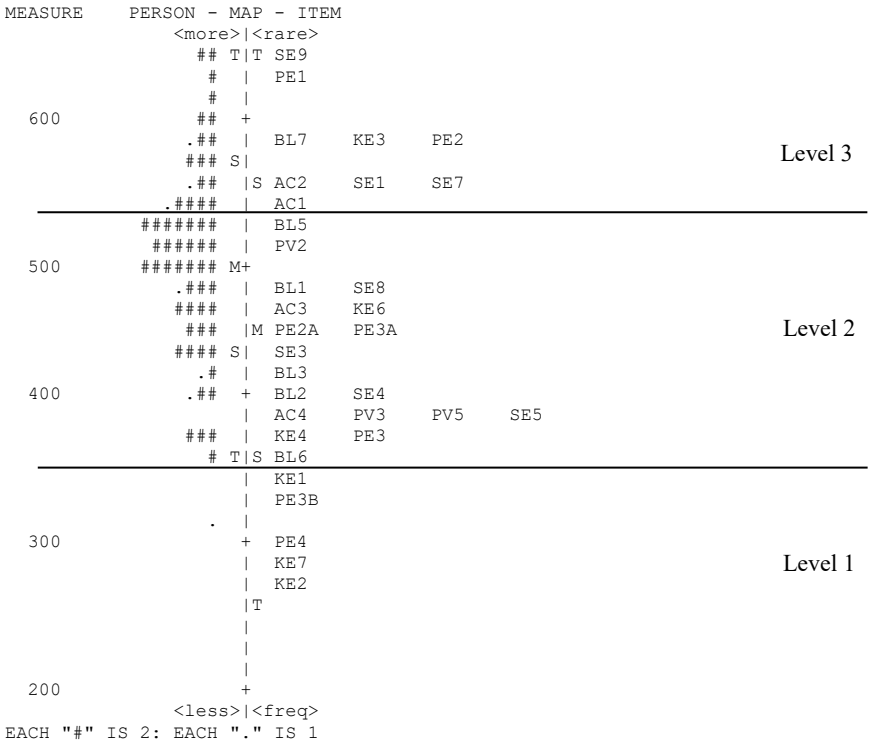

Figure 4. Wright map (the same as in Figure 4) with levels of energy literacy. The description of competencies refers to the competencies within items in each level.

References

[1] C. Dwyer, "The Relationship between Energy Literacy and Environmental Sustainability," *Low Carbon Econ.*, vol. 02, no. 03, pp. 123–137, 2011.
[2] R. Duit, H. Niedderer, and H. Schecker, "Teaching physics," in *Handbook of research on science education*, S. K. Abell and N. G. Lederman, Eds. New York: Routledge, 2010.
[3] L. Mohan, J. Chen, and C. W. Anderson, "Developing a multi-year learning progression for carbon cycling in socio-ecological systems," *J. Res. Sci. Teach.*, vol. 46, no. 6, pp. 675–698, 2009.
[4] R. P. Lee, "Misconceptions and biases in German students' perception of multiple energy sources: implications for science education," *Int. J. Sci. Educ.*, vol. 38, no. 6, pp. 1036–1056, 2016.
[5] I. Aguirre-Bielschowsky, R. Lawson, J. Stephenson, and S. Todd, "Energy literacy and agency of New Zealand children," *Environ. Educ. Res.*, vol. 23, no. 6, pp. 832–854, 2017.
[6] I. P. Cheong, M. Johari, H. Said, and D. F. Treagust, "What Do You Know about Alternative Energy? Development and Use of a Diagnostic Instrument for Upper Secondary School Science," *Int. J. Sci. Educ.*, vol. 37, no. 2, pp. 210–236, Jan. 2015.
[7] M. J. Fell and L. F. Chiu, "Children, parents and home energy use: Exploring motivations and limits to energy demand reduction," *Energy Policy*, vol. 65, pp. 351–358, 2013.
[8] K.-L. Chen, S.-Y. Liu, and P.-H. Chen, "Assessing multidimensional energy literacy of secondary students using contextualized assessment," *Int. J. Enviroment Sci. Educ.*, vol. 10, no. 2, pp. 201–218, 2015.
[9] S.-J. Chen, Y.-C. Chou, H.-Y. Yen, and Y.-L. Chao, "Investigating and structural modeling energy literacy of high school students in Taiwan," *Energy Effic.*, vol. 8, no. 4, pp. 791–808, Jul. 2015.
[10] O. L. Liu, K. Ryoo, M. C. Linn, E. Sato, and V. Svihla, "Measuring knowledge integration learning of energy topics: A two-year longitudinal study," *Int. J. Sci. Educ.*, vol. 37, no. 7, pp. 1044–1066, 2015.
[11] P. Davis, "The attitude and knowledge of Tasmanian secondary students towards energy conservation and the environment," *Res. Sci. Educ.*, vol. 15, pp. 68–75, 1985.
[12] J. E. DeWaters and S. E. Powers, "Energy literacy of secondary students in New York State (USA): A measure of knowledge, affect, and behavior," *Energy Policy*, vol. 39, no. 3, pp. 1699–1710, Mar. 2011.
[13] J. DeWaters, B. Qaqish, M. Graham, and S. Powers, "Designing an Energy Literacy Questionnaire for Middle and High School Youth," *J. Environ. Educ.*, vol. 44, no. 1, pp. 56–78, Jan. 2013.
[14] P. Halder *et al.*, "International survey on bioenergy knowledge, perceptions, and attitudes among young citizens," *Bioenergy Res.*, vol. 5, no. 1, pp. 247–261, 2012.
[15] C. C. Holden and L. H. Barrow, "Validation of the test of energy concept and values for high school," *J. Res. Sci. Educ.*, vol. 21, no. 2, pp. 187–196, 1984.
[16] D. J. Kuhn, "Study of the attitudes of secondary school students toward energy-related issues," *Sci. Educ.*, vol. 63, no. 5, pp. 609–620, 1979.
[17] Y.-F. Lay, C.-H. Khoo, D. F. Treagust, and A. L. Chandrasegaran, "Assessing secondary school students' understanding of the relevance of energy in their daily lives," *Int. J. Sci. Educ.*, vol. 8, no. 1, pp. 199–215, 2013.
[18] L. S. Lee, Y. F. Lee, J. W. Altschuld, and Y. J. Pan, "Energy literacy: Evaluating knowledge, affect, and behavior of students in Taiwan," *Energy Policy*, vol. 76, pp. 98–106, 2015.
[19] S. Wei, X. Liu, Z. Wang, and X. Wang, "Using Rasch measurement to develop a computer modeling-based instrument to assess students' conceptual understanding of matter," *J. Chem. Educ.*, vol. 89, no. 3, pp. 335–345, 2012.
[20] M. T. Kane, "An argument-based approach to validity," *Psychological Bulletin*, vol. 112, no. 3, pp. 527–535, 1992.

[21] M. T. Kane, "Validating the interpretation and uses of test scores," *J. Educ. Meas.*, vol. 50, no. 1, pp. 74–83, 2013.
[22] C.-Y. Kuo, H.-K. Wu, T.-H. Jen, and Y.-S. Hsu, "Development and validation of a multimedia-based assessment of scientific inquiry abilities," *Int. J. Sci. Educ.*, vol. 37, no. 14, pp. 2326–2357, 2015.
[23] M. Yusup, A. Setiawan, N. Y. Rustaman, and I. Kaniawati, "Developing a framework for the assessment of pre-service physics teachers' energy literacy," *J. Phys. Conf. Ser.*, vol. 877, p. 012014, Jul. 2017.
[24] J. M. Linacre, *A user's guide to Winsteps® Rasch-model computer program*. Beaverton, Oregon: Winsteps.com, 2016.
[25] J. M. Linacre, "Disorder category and threshold," 2017. [Online]. Available: http://raschforum.boards.net/thread/669/disorder-category-threshold. [Accessed: 15-Mar-2017].
[26] D. B. Swanson, G. R. Norman, and R. L. Linn, "Performance-based assessment: Lessons from the health professions," *Educ. Res.*, vol. 24, no. 5, pp. 5–11, 1995.
[27] K. E. Green and C. G. Frantom, "Survey development and validation with the Rasch model," in *International Conference on Questionnaire Development, Evaluation and Testing*, 2002.
[28] T. G. Bond and C. M. Fox, *Applying the Rasch model: Fundamental measurement in the human sciences*, 3rd. Ed. New York: Routledge, 2015.
[29] P. Baghaei, "The Rasch model as a construct validation tool," *Rasch Meas. Trans.*, vol. 22, no. 1, pp. 1145–1146, 2008.
[30] R. K. Hambleton, "Emergence of item response modeling in instrument development and data analysis," *Med. Care*, vol. 38, no. 9 Suppl, pp. II60-65, 2000.
[31] J. M. Linacre, "When does a gap between measures matter?," *Rasch Meas. Trans.*, vol. 18, no. 3, p. 993, 2004.
[32] J. Lai and D. T. Eton., "Clinically meaningful gaps," *Rasch Meas. Trans.*, vol. 15, no. 4, p. 850, 2002.
[33] W. J. Boone, J. R. Staver, and M. S. Yale, *Rasch analysis in the human sciences*. Dordrecht: Springer, 2014.