Development and Validation of Students' Competency Instrument on Science Process Skills

Ellyza Karim[a,*], Syakima Ilyana Ibrahim[b], Hanani Harun Rasit[c] and Bambang Sumintono[d]

[a,b,c]Teacher Educational Institute, Technical Educational Campus, Bandar Enstek, Negeri Sembilan, Malaysia
[d]Universitas Islam Internasional Indonesia

[a] ellyza.karim@ipgm.edu.my, [b] syakimailyana@ipgm.edu.my, [c] hanani.harun@ipgm.edu.my, [d]bambang.sumintono@uiii.ac.id

*corresponding author

Abstract: Science process skills are fundamental to science learning and consequently promote scientific literacy among students. Scientists use these skills to construct knowledge during investigation. Thus, science process skills should be acquired not only by scientists, but by all individuals. This study reviews the validity and reliability of an instrument development for science process skills using interval rating scale as answer choice ranging from poor to excellent. This instrument was developed based on the perception of 300 respondents who were the primary school leavers, aged thirteen. As the instrument was newly developed, a pilot study was conducted to determine the empirical proof on the validity and reliability of the test. A total of 68 items were constructed in the test using verified indicators based on the experts and literature reviews. Rasch Measurement Model was employed to analyse the data. Overall, the person reliability is found at 0.97 while the item reliability is 0.99. The range of Point Measure Correlation (PMC) is positive between 0.33 to 0.67 for all items. All items are accepted as the outfit mean square (MNSQ) has the range between 0.59 and 1.55 while the infit MNSQ is between 0.74 to 1.56 indicating a good measure of latent variables for item fit. Based on the item map, the findings suggest that, for experimenting skill, only 11% of students could design scientific steps independently. Meanwhile, for using space and time relationship skill, respondents reported that the item 'determination for the object position with time' was the most excellent item. The scale used in this instrument was also found to adequately measure the latent constructs of science process skills.

Keywords: science process skills, construct validity, reliability, scale calibration, unidimensionality

1. Introduction

The standard based Primary School Standard Curriculum (KSSR) in science aims to upgrade scientific literacy in order for better scientific understanding among school children in Malaysia. The acquisition of Science Process Skills enables children to be involved in science learning more effectively. This study presents development and validation of an instrument to capture students' competency in Science Process Skills (SPS) which focuses on primary school experiences using Likert scale items. This new instrument is developed to be used as a heuristic device to capture the current performance of students' science process skills in order to facilitate their experiences in science learning. To make sure the instrument is valid and reliable, Rasch measurement model is used to assess this instrument's capacity to emulate the properties of fundamental measurement. Rasch measurement is a psychometric technique that was developed to improve the precision of instrument validity and reliability. Apart from that, Rasch measurement model allowed researchers to fulfil the item response theory concept in order to produce robust analysis. In this study, the data collected is for the purpose of the pilot study. Further action will be taken to improve the instrument based on the pilot study findings. Tracking the proficiency in SPS is crucial to bolstering a deeper understanding of scientific concepts among school children especially in Malaysia.

Q. Zhang (ed.), *Proceedings of the Pacific-Rim Objective Measurement Symposium (PROMS 2023)*,
Atlantis Highlights in Social Sciences, Education and Humanities 23,
https://doi.org/10.2991/978-94-6463-494-5_20

2. Literature Review

For decades, many scholars have argued that scientific competence can be developed through the operation of science processes [6, 10, 12, 22]. They purported that inculcating the SPS in the way of scientist work is critical to the disposition of the creative and innovative talents in a younger generation. SPS helps students understand science by 'doing' or 'experiencing' the scientific thinking processes on their own, using the tools they need. Different researchers underlined different skills upon their concerns.

Curriculum Development Centre has listed twelve SPS which are divided into two categories: basic and integrated [9]. The basic SPS are observing, inferring, predicting, classifying, measuring, using space and time relationships and communicating. While the integrated SPS are formulating hypotheses, defining operationally, identifying and controlling variables, interpreting data and experimenting. Other skills highlighted in other articles are raising question [16] and formulating model [27].

To date, most of the SPS tests were designed using multiple-choice format as depicted in Table 1. Since the 1980s, researchers have the inclination to use multiple-choice questions to evaluate SPS knowledge. The most popular is the Test of Integrated Process Skills (TIPS) [10]. TIPS was then further modified as TIPS II [6] which are considered as the primary source of other SPS instrument developers. Based on the review of the relevant literature pertaining to measurement on SPS, it is noted that psychometric test with the Likert scale-rated items is rare. The normal practice is to measure SPS via cognitive domain which is represented as pen and paper response to multiple choice questions, as exemplified in Table 1. Most researchers tend to study students' knowledge about SPS by analysing their performance marks.

Table 1 Developed instruments for assessing SPS.

Researcher	Origin	Year	Skills studied	Instrument format
Dillashaw & Okey [10]	USA	1980	5	Multiple-choice
Burns, Okey & Wise [6]	USA	1985	5	Multiple-choice
Smith & Welliver [26]	USA	1990	13	Multiple-choice
Beaumont-Walters & Soyibo [7]	Jamaica	2001	5	Practical Structure
Abu Hassan & Rohana [1]	Malaysia	2003	6	Structure Multiple-choice
Kazeni [17]	South Africa	2005	5	Multiple-choice
Temiz, Tasar & Tan [27]	Turkey	2006	12	Multiple-choice
Edy Hafizan & Lilia [11]	Malaysia	2010	5	Multiple-choice
Ong Eng Tek et al. [22]	Malaysia	2012	12	Multiple-choice
Ong Eng Tek & Mohd Al-Junaidi [23]	Malaysia	2013	12	Multiple-choice
Ong Eng Tek et al. [24]	Malaysia	2015	7	Multiple-choice
Ellyza [14]	Malaysia	2020	12	Likert scale
Nazahiyah Mustafa, Ahmad Zamri Khairani & Nor Asniza Ishak [21]	Malaysia	2021	12	Multiple-choice Structure
Chin & Ellyza [8]	Malaysia	2023	12	Multiple-choice

Previous researchers tend to analyse their instruments by measuring item difficulty index, item discrimination and student's score percentage according to the Classical Test Theory (CTT) category.

Even the reliability index is analysed according to Cronbach alpha measurement. With the development of Item Response Theory (IRT) today, it is better to analyse the data using the latest techniques. This is because there is a weakness in CTT where it is more oriented to the analysis of items. This means that CTT cannot make predictions about the performance of its respondents individually when answering an instrument due to the nature of group dependency. Therefore, IRT is an alternative theory to improve the weaknesses found in CTT. This is based on the characteristics of the item that does not depend on the group, for example the score that describes the abilities of an individual does not depend on the test results as a whole. Specifically, when the researcher uses the Rasch measurement model, then the analysis can be explained to the characteristics of items and individual respondents, does not require parallel tests to assess the reliability of items-individuals and can also provide accurate measurements on each person's ability score [17].

Test of Integrated Process Skills (TIPS) with a mean discrimination index of 0.39 while the mean item difficulty index was 52 percent [10]. Only such analysis is presented by past studies. Therefore, today's researchers should improve the way of analysis because there are new techniques to analyse data that are much more relevant. TIPS was further improved to TIPS II [6] and became very popular until it was referred to and used by many SPS researchers, for example by lecturers from USM, Malaysia in 1998; from UMS, Malaysia in 2004 and lecturers from Carbondale University, America in 2012. American TIPS and TIPS II were also translated into other languages such as Turkish. TIPS II has a discrimination index in the range of 0.11-0.64 with a mean of 0.35 while the difficulty index was 0.15-0.87 with a mean of 0.53. Still analysed according to CTT and in the form of a multiple-choice test with four answer options. TIPS and TIPS II contain 36 questions covering various fields of science and are not detailed to specific fields. The sample consisted of seventh through twelfth grade students around the United States and tested the five SPS of identifying variables, operationally defining, hypothesising, experimenting, and analysing graphical data. The limitation of this study is that it only focuses on five SPS, so the researcher tried to produce an instrument capable of studying all twelve SPS.

Based on the paragraph above, TIPS and TIPS II published in the 80s are seen as the reference of many famous researchers. However, these researchers only analysed in the form of discrimination index and difficulty index, which is one of the CTT techniques. It was found that the researcher at that time had not taken the initiative to analyse the item more deeply. While at that time there was a factor analysis capable of analysing instruments in a IRT way. Factor analysis was published for the first time by Spearman in 1904 with only one factor. Then in 1915, Spearman's protégé, Carey, published a factor analysis that examined more than one factor. In 1930, Thurstone produced the latest factor analysis concept with technical innovations. This proves that IRT techniques have been around for a long time. However, recent researchers such as Turkish researchers still follow the style of previous researchers. They were found to still analyse the items in the form of discrimination index and difficulty index and analyse the reliability of the instrument using Cronbach's alpha value. Not only factor analysis, but the Rasch measurement model has been introduced since 1960 by Georg Rasch, a Mathematician from Denmark. So, it is appropriate that the previous researchers used variety of instruments to measure KPS using the latest data analysis tools such as factor analysis, Rasch measurement model or more recent ones which are Structural Equation Model (SEM).

Undoubtedly, multiple choice tests are also psychometric assessments just like Likert scale tests. However, the analysis of multiple-choice tests based on binary sets is similar to the selection of yes or no answers presented in descriptive form. Unlike the Likert scale, which is very rare in SPS studies, it can be analysed either descriptively or inferentially, depending on the researcher's objectives. Another strength of the Likert scale is that the data types can be translated into ordinal or

interval scale forms, unlike binary sets which are translated as nominal scales. Therefore, the development of this instrument is eligible for IRT analysis by specialising in Rasch measurement model analysis. Data collected nominally can only be analysed in the form of frequency. Interval data can be used to study correlation as needed for the purpose of criterion validity.

Albeit the robustness of such instruments, scholars argue that measuring SPS should not be focusing on the performance part only but might as well investigate the competency itself. The items were built randomly within topics of the revamped and implemented Malaysian primary science standard curriculum known as KSSR (*Kurikulum Standard Sekolah Rendah*, commonly abbreviated as KSSR; Malay) which was executed in 2011. Thus, the target respondents were the UPSR leavers aged 13 years old. Primary School Achievement Test, also known as *Ujian Pencapaian Sekolah Rendah* (commonly abbreviated as UPSR; Malay), refers to a standardised test for primary school finalists all around Malaysia before they enter secondary school.

Analysis of items using the application of the Rasch model of measurement was performed to check the psychometrics properties instead of other validation phases on instrument development. However, this article only presents findings performed by Rasch analysis in terms of validation and reliability. Reliability is the consistency of the results of assessment test over time. Validity refers to what a test wants to measure or the purpose of the test.

3. Methodology

This paper intends to present the procedure involved in developing and validating the instrument according to IRT analysis. The study of the validity and reliability of the instrument is very important to maintain the accuracy of the instrument from defects. The development of the instruments was conducted in three phases.

The first phase is the item construction. The items were developed consisting of twelve constructs based on seven basic SPS and five integrated SPS used in Malaysian Primary Science Standard Curriculum implemented in 2011 and reviewed in 2017. A total of 68 items delineated the twelve constructs based on the indicators proposed by past researchers [15, 16, 31]. Table 2 shows the indicators used to characterise the items according to the twelve constructs.

Table 2 Indicators of the twelve constructs of SPS

Construct	Indicator
Observing	use the sense of sight
	use the sense of hearing
	use the sense of smell
	use the sense of touch
	use the sense of taste
	state the similarities and differences on the things observed
	identify changes that occur
	order the sequence of events correctly
	use suitable equipment for sensory recognition
	observe quantitatively
Classifying	group objects based on certain characteristics
	identify the characteristics of different objects
	identify features of the same object
	classify the same object in different ways
Measuring and using numbers	calculate the quantity
	measure data precisely
	use the correct measuring tool
	give the correct unit on the measurement reading
	identify patterns from data tables

Inferring	use information from observations to draw preliminary conclusions
	identify inference weaknesses
	make multiple interpretations from one observation
	test the accuracy of the inference through additional observations
Predicting	use evidence from past or present experiences to predict future events
	determine possible findings from data patterns
	be careful with findings without solid evidence
	express the level of confidence in the accuracy of one's own predictions
Communicating	present ideas orally
	present ideas in writing
	use models to represent information
	use scientific terms when sharing results
	explain the meaning of the symbol
	use graphics to represent information
	record data based on investigation
	use scientific knowledge when constructing the question to be asked
Using space and time relationships	determine the rate of change of an event
	determine the position of the object with time
	describe the change in direction with time
	describe changes in shape with time
	describe the change in size with time
	arrange events in time
	explain the relationship between time and the distance of a moving object
Interpreting data	answer questions from the data collected
	draw conclusions based on the data collected
	analyze the results from the data collected
	identify patterns in the data obtained
	identify aspects that cause the investigation to be unfair
	express the relationship between information
Defining operationally	give meaning to terms based on own experience
	give meaning to terms based on findings
	give meaning to terms through the description of what has been observed and implemented
Controlling variables	identify what to measure on the response variable
	identify what to set on the constant variable
	identify what needs to be changed on the manipulated variable
	determine three types of variables
Making a hypothesis	make a relationship between the manipulated variable and the response variable
	use existing knowledge to make explanations
	know that there is more than one explanation to explain an event
	realize that the description given is only a suggestion
Experimenting	state the problem based on a known problem
	create a hypothesis based on a known problem
	determine the appropriate method, materials and apparatus as planned
	write experimental instructions so that others can repeat the activity just by following the instructions
	design procedures scientifically
	perform experiments to test hypotheses
	collect data honestly

	draw conclusions based on data interpretation
	report experimental results

The 5-point Likert rating scale was adapted in ascending order of perceived ability, ranging from score 1 (poor) to 5 (excellent), as depicted in Figure 1. The first scale is not zero due to the assumption that the respondent has at least a low ability on science process skills, as opposed to zero ability.

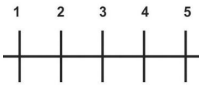

Figure 1 Likert rating scale

The second phase is the validity phase. This includes face validity and content validity. Three reviewers were appointed to check the face validity in terms of wording, science terminology, technical aspects and suitability of items with regards to students' level. Then a pre-test was administered to five UPSR candidates to detect any problem within the items. Minor changes were made upon face validity findings. Then, the items were presented to thirteen experts for their consensus on content validity. Fuzzy Delphi Method (FDM) was used to verify their consensus. A few items were modified according to experts' suggestions to produce better items [13].

Finally, the third phase was characterised by psychometric analysis where the results of reliability as well as validity were analysed using Rasch model analysis. Initially, data were prepared using electronic worksheet, MS Excel, for keyed in. Subsequently, the data was collected and analysed using Rasch analysis for two facet models via computer applications, WINSTEPS version 3.73. The quality of this instrument is determined based on several psychometric attributes such as reliability, separation, unidimensionality and fit statistics. Scale calibration is also done to investigate and confirm that the 5-category rating scale was a better model for SPS competency measurement. Besides that, information on the test items such as the strengths and weaknesses could be found in the item map analysis.

For the pilot study purpose, a set of 68-items test was distributed to a total of 300 Form One students regardless of their cognitive levels, drawn from eight schools, selected randomly all over Malaysia. The test was administered for the duration of one hour and ten minutes. The detailed breakdown of the respondents who participated in this study is given in Table 3.

Table 3 Breakdown of respondents in the piloting of SPS instrument (n=300)

Demographic Factor	**Factor**	**Frequency**	**Percentage (%)**
Gender	Boy	147	49.0
	Girl	153	51.0
Race	Malay	248	82.7
	Indian	27	9.0
	Chinese	23	7.7
	Others	2	0.7
Science UPSR Grade	A	96	32.0
	B	137	45.7
	C	49	16.3
	D	15	5.0
	E	3	1.0

4. Results and Discussion

4.1 Item polarity

In Rasch model analysis, Point-Measure Correlation (PMC) is used to identify item polarity. It is carried out to test whether all the items are moving in one direction with the construct [2]. All PMC measurements for each item in this instrument display a positive index between the range of 0.33-0.67. Positive items indicate that the coding of those items is working in the right direction [30]. The working parameter for an acceptable PMC value must be between: $0.4<x<0.8$ [25].

4.2 Reliability and separation index

Rasch analysis produces person separation index and items. To determine the reliability of the SPS instrument and to what extent the adequacy of the separation index of the SPS instrument, Table 4 shows the statistics generated by Rasch analysis of SPS indicates how Rasch model conforms to item and person separation indexes and reliability. Table 4 proved that a high value of the separation index will increase the value of the reliability index [28].

The separation index shows the number of ability strata identified for the group of persons and items. Separation of items and individuals means the extent of items will spread in the latent trait or the ability of individuals [4]. Separation index that is higher than 2.0 to be acceptable for both person and item [20]. This shows that the items in the test are spreading fairly with individual abilities in logits.

Table 4 Reliability analysis and separation of SPS index

Construct	ID item	Item measure		Person measure	
		Reliability	Separation	Reliability	Separation
Observing	PERHATI1-10	0.98	6.62	0.96	5.03
Classifying	NGELAS11-14	0.99	9.28	0.96	5.03
Measuring and using numbers	UKUR15-19	0.95	4.60	0.96	5.03
Inferring	INFER20-23	0.97	5.41	0.96	5.03
Predicting	RML24-27	0.99	8.22	0.96	5.03
Communicating	KOM28-35	0.99	9.01	0.96	5.03
Using space and time relationship	RUANG36-42	0.99	7.99	0.96	5.03
Interpreting data	TFSIR43-48	0.99	8.42	0.96	5.03
Defining operationally	DSO49-51	0.00	0.00	0.96	5.03
Controlling variables	PU52-55	0.91	3.10	0.96	5.03
Making a hypothesis	HIPO56-59	0.98	7.99	0.96	5.03
Experimenting	EKS60-68	0.99	9.34	0.96	5.03
Overall	**1-68**	**0.99**	**11.70**	**0.96**	**5.03**

The value of the separation index specifies the isolation of the item difficulty level. For overall item separation, the items' strata are distributed up to eleven levels. All the SPS constructs have good item separation indexes except for defining operational construct [19]. This result may be due to a lack of item quantity, where there are only three items for operational construct. By referring to other constructs, the minimum number of items should be four. To solve this problem, an item will be added to this construct for the actual research.

Throughout the analysis, the results suggest that there is a persistent in person separation in which there are five strata for the person. Person isolation index indicates the number of strata capabilities identified in the sample group. The five strata align with the grading system in the UPSR which also comprises of five grades namely A, B, C, D and E. As for reliability analysis, overall reliability for items in Table 4 shows a very good index which is 0.99 while reliability for person is 0.96. The value above 0.8 indicates a strong reliability [4].

4.3 Fit statistics

The statistics generated by Rasch analysis estimate the degree of items suitability that measures latent variables, assuring the item-fit of the instrument is within an acceptable range. The infit and outfit MNSQ of each item and respondent should be in the range of 0.50 to 1.50 for Likert scale [5].

The results show that infit MNSQ for the items are between the range of 0.74-1.56 while outfit MNSQ are 0.59-1.55 with the least standard error. This implies that all the items in the suggested range can be retained. If the MNSQ values are accepted then the Z-standard can be omitted [3, 4]. Hence, the Z-standard value is ignored.

The minimum and maximum infit and outfit MNSQ are in the good range as depicted in Table 5. The minimum and maximum person infit MNSQ are 0.53-1.50 while outfit MNSQ are 0.58-1.50. In all, 300 respondents were found fit with the Rasch Model.

Table 5 Summary statistics on item and person

	Item	**Person**
Measure		
Mean	0.00	1.41
S.D.	0.99	0.95
Max	2.56	5.63
Min	-2.24	-0.82
Infit MNSQ		
Mean	1.01	1.01
S.D.	0.16	0.25
Max	1.56	1.50
Min	0.74	0.53
Outfit MNSQ		
Mean	0.99	0.99
S.D.	0.17	0.21
Max	1.55	1.50
Min	0.59	0.58
S.E.	0.12	0.05
Alpha Cronbach	0.97	

4.4 Unidimensionality

As for correctness of measurement, unidimensionality of the items was also examined. Unidimensionality test identifies and measures the extent to which an item is measuring what it should be measuring, using the principal component analysis (PCA). Rasch analysis applies the PCA of the residuals, which indicates how much variance the instrument is measuring what it is supposed to measure.

Throughout the analysis, it is found that the raw variance explained by measures is 53.3% while the unexplained variance in first contrast is 2.9%. The Eigen value for the instrument is 3.7 which follows the rule of less than 5.0 [4], meaning that there is no clear existence of the second dimension. For the twelve constructs shown on Table 6, all the Eigen values are between 1.2-2.0. The ratio of variance explained with the first principal component variance for both results upon SPS construct are above the rule of 3:1 which is 8:1. Therefore, the results indicate good measurement to measure SPS competency in unidimensionality context.

Table 6 Eigen value of SPS instrument

Construct	**Eigen value**
Classifying	1.2
Predicting	1.3
Making hypothesis	1.3
Inferring	1.4
Interpreting data	1.4
Measuring and using numbers	1.5
Using space and time relationship	1.5
Controlling variables	1.6
Communicating	1.8
Experimenting	1.9
Observing	2.0
Defining operationally	2.0
Overall	**3.7**

4.5 Scale calibration

Scale calibration works to identify whether the rating scale categories have performed as intended. Rasch analysis enables such calibration, which is also known as thresholds, to investigate and confirm their assumption for improving measurement. If more categories are provided in the test but not used as proposed by respondents, then more categories do not necessarily implying more information could be collected.

Figure 2 shows the peak shapes of the probability curves at each category and are not shadowing each other. While Table 7 shows the observed average and Andrich threshold values discovered, consistently increased with the rating scale category. Outfit MNSQ was less than two. The outfit statistic measures explained the variance. A score higher than two implies greater noise for the unexplained variance. This pattern confirms that it is not necessary to amend the rating scale. It suggests that the instrument adequately measures the latent constructs of SPS.

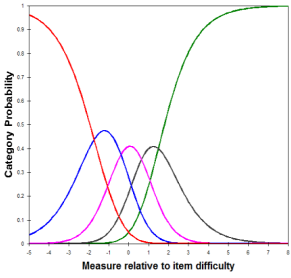

Figure 2 Model Probability Characteristic Curves

However, the deduction product of some category thresholds in Table 7 is not in between 1.4 to 5.0 logits. With regard to the thresholds, each threshold or step up the scale, should be at least 1.4 logits greater than the last, to show appropriate distinction between categories [32]. However, intervals of more than 5 logits indicate that there is a gap in the measurement of the trait. As most calculations in this instrument produce less than 1.4 logits, the solution taken was to label all the categories. In order to interpret the responses, each rate must be reported by the meaning it adheres to. For instance, category 1 as poor ability, 2 as fair ability and so on. This may produce a more clear and meaningful understanding among respondents.

Table 7 Summary of five categories structure

Category		**Observed**		**Observed Average**	**Sample Expect**	**MNSQ**		**Andrich threshold**	**Category Measure**
Label	Score	Count	%			**infit**	**outfit**		
1	1	482	3	-0.60	-0.86	1.26	1.23	NONE	(-3.00)
2	2	1677	12	-0.22	0.15	0.94	0.95	-1.74	-1.21
3	3	2971	21	0.53	0.52	1.01	0.99	-0.39	0.07
4	4	3854	27	1.25	1.53	1.02	0.93	0.64	1.24
5	5	5228	37	2.38	2.35	0.94	0.96	1.49	(2.83)

$$S_{1\text{-}2} = 0.00 - (-1.74) = 1.74 \; (> 1.4)$$
$$S_{2\text{-}3} = -0.39 - (-1.74) = 1.35 \; (< 1.4)$$
$$S_{3\text{-}4} = 0.64 - (-0.39) = 1.03 \; (< 1.4)$$
$$S_{4\text{-}5} = 1.49 - (-0.64) = 0.85 \; (< 1.4)$$

Even though the results do not meet the complete prerequisite, the results represent a close enough approximation [18]. In fact, the calculation in Table 8 while the five categories were collapsed into four, three and two categories show that the best results in separation and reliability fall in the fifth category rating scale. As a conclusion, the five-category is retained but with the prior labels revised to poor ability (1), fair ability (2), moderate ability (3), good ability (4) and excellent ability (5).

Table 8 Summary of scale collapsing effects.

Rating scale	**Separation**		**Reliability**		**Infit MNSQ**	**Outfit MNSQ**
	Person	**Item**	**Person**	**Item**		
5-Category	5.01	10.70	0.96	0.99	1.04	1.06
4-Category	3.99	8.82	0.94	0.99	1.03	1.05
3-Category	2.79	5.89	0.89	0.97	1.04	1.08
Dichotomy	2.11	3.71	0.82	0.93	1.05	1.06

4.6 Item Map

Figure 3 shows the Wright map that provides the location of all students (left) and items (right) on the logit ruler. The highest students indicate the most competent students while the highest item indicates the most difficult item to comply with, represented by item EKS64 under experimenting skill (designing procedures scientifically). The lowest item located is also the easiest skill to achieve represented by item RUANG37 (using space-time relationship ability) which shows the ability to describe location with time. However, the appearance of 17 off target items with zero respondent shows that these items are easily achieved.

There is a group of students located above the maximum item logit; +2.56 logit indicating all these students excel in all the skills measured. Moreover, there is another group of person free items among these excellent students. This situation indicates that the item difficulty does not challenge

these respondents. The person mean is higher than the item mean. This illustrates that the items can be accomplished easily or in other words, most of the respondents have good ability in SPS.

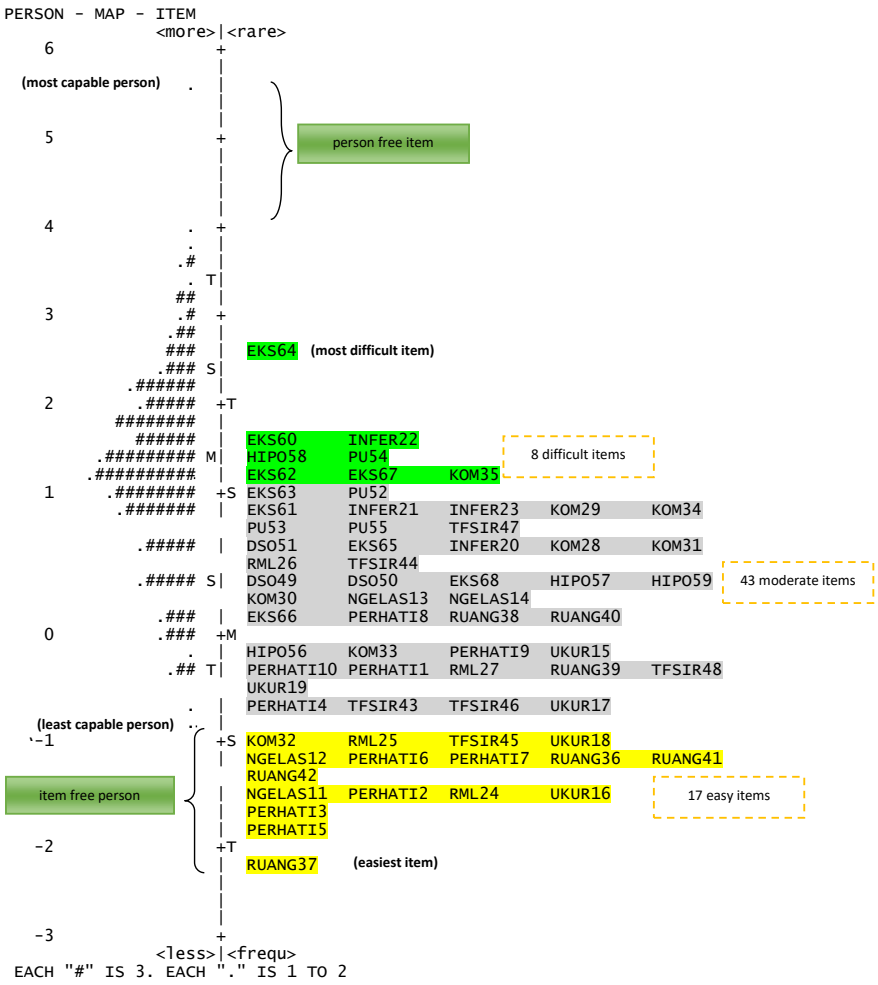

Figure 3 Mapping Difficulty Item - Respondent Ability in SPS

A group of 119 students in the center, located between the first and second term of standard deviation, attains moderate ability for their level of SPS. These students face difficulty with eight items as the items are located above their ability measure. They are capable upon most of the items. The appearance of item free person illustrates that these items do not measure the corresponding sample skills. These 17 items are considered as easy items and need amendments to increase difficulty.

Guttman the psychometrician advocating that, "the more able the person, the more likely a success on any item [29]. The more difficult the item, the less likely a success for any person". The

part that appears to be the person free item and item free person are poor signs indicating that the items cannot measure person's ability in that skill. The pattern of the Wright map also illustrates most of the items are easy when it clustered at a lower location on the right logit ruler. Whereas the clustered pattern on the left of the logit ruler shows majority of the respondents can easily answer most items. The results led to an initial conclusion that a researcher should increase the item difficulty to precisely capture the students' science process skill that could both illustrate and differentiate the ability of excellent students.

5. Conclusion and suggestion

Throughout the analysis, it is gathered that assessing and interpreting the competency level of SPS among students are undoubtedly important. With the premise that gauging students' level of SPS at the preliminary stage would diminish potentially arising problems dealing with laboratory tasks, early identification is apt. The ability of the instrument to measure the initial skills and capacity of primary school leavers during their early entrance into secondary school would allow secondary school teachers to deliberately plan the relevant remedial activities before the students barely move to the higher level of schooling. Thus, developing an SPS instrument is a must to assist science educators in gauging their students' SPS level.

There is a need to develop and validate a Malaysian-based SPS instrument. Rasch provides empirical evidence on instrument development study. This study shows the necessity to raise the items' difficulty on the SPS instrument, develop enough items for each construct, and label the five categories rating scale. Then the instrument is ready to be administered to the next respondents for their true stage of the instrument development process.

On reflection, the students' weakness in designing the procedural scientific process revealed through this study paved an understanding that current strategies should be deliberately revised to raise students' attainment and performance in SPS. The use of the text and exercise books must be extended through the creative designing of lessons such as via a project-based approach. Students' habit of copying the procedural steps like referring to a cookbook recipe should be prevented but they start to transform into self-design procedure whenever they are doing a science investigation or experiment. Avoid repeating the current mistake in the way students conduct their science tasks. This action may trigger the awareness of producing more designable scientifically competent students in the future.

The study has revealed eye-opener findings. Students' weakness in designing the scientific procedural processes fortifies the understanding of the need to deliberately revise the current strategies in inculcating the SPS skills in schools. Textbooks, despite their importance, should not be merely considered as the sole reference in assisting students' learning. Designing learning demands more of the teacher's determination to be creative in inculcating and instilling scientific skills in students. Teachers have the autonomy to choose their own teaching approaches and resources. This effort is aligned with the global shift in employability needs and the new transformation urged by the government to produce the skilled workforce in the STEM career-based context.

Based on the findings, the instrument tested is capable of measuring students' proficiency in SPS based on their schooling experience. In addition, the Malaysia-based instrument can also identify students' SPS ability based on local science curriculum experience. Further in measuring SPS competence through student curriculum experience, the instrument also provides added value to educators in evaluating the ability of the science curriculum being implemented. Therefore, this instrument is seen as very appropriate to be used in identifying students' competence in SPS as well as helping teachers improve their practices in providing the best SPS practical experience to students.

6. Acknowledgement

The authors convey their gratitude to Prof Kamisah Osman, National University of Malaysia for providing guidance throughout the research.

7. References

[1] Abu Hassan Kassim & Rohana Hussin. (2003). *Tahap Penguasaan Kemahiran Proses Sains dan Hubungannya dengan Pencapaian Kimia di Kalangan Pelajar Tingkatan Empat Daerah Johor Bahru.* Retrieved from Universiti Teknologi Malaysia website: http://eprints.utm.my/2345/.

[2] Azrilah Abdul Aziz, Mohd Saifudin Masodi & Azami Zaharim. (2013). *Asas model pengukuran Rasch: Pembentukan skala dan struktur pengukuran.* Bangi: UKM Press.

[3] Bambang Sumintono & Wahyu Widhiarso. (2014). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial.* Cimahi: Trim Komunikata Publishing House.

[4] Bond, T. G. & Fox, C. M. (2007). A*pplying the Rasch model: Fundamental measurement in the human sciences.* London: Lawrence Erlbaum Associates.

[5] Boone, W. J., Satver, J. R. & Yale, M. S. (2014). R*asch analysis in the human sciences*. Dordrecht: Springer.

[6] Burns, J. C., Okey, J. R. & Wise, K. C. (1985). Development of an integrated process skill test: TIPS II. *Journal of Research in Science Teaching, 22*(2): 169-177.

[7] Beaumont-Walters, Y. & Soyibo, K. (2001). An analysis of high school students' performance on five integrated science process skills. *Research in Science & Technological Education*, *19*(2). Retrieved from http://dx.doi.org/10.1080/ 02635140120087687.

[8] Chin, K.Y. & Ellyza Karim. (2023). *Hubungan antara Tahap Penguasaan Kemahiran Proses Sains dan Tahap Pencapaian Sains Murid Tahun Enam.* Research Project, Teacher Educational Institute, Technical Educational Campus.

[9] Curriculum Development Centre. (2021). *Dokumen standard kurikulum dan pentaksiran sains tahun enam.* Putrajaya: MOE.

[10] Dillashaw, F. G. and Okey, J. R. (1980). Test of integrated science process skills for secondary students. *Science Education. 64*, 601-608.

[11] Edy Hafizan Mohd Shahali & Lilia Halim. (2010). Development and validation of a test of integrated science process skills. *Procedia Social and Behavioral Sciences, 9*, 142-146. doi:10.1016/j.sbspro.2010.12.127

[12] Edy Hafizan Mohd Shahali, Lilia Halim & Subahan Meerah. (2010). Perception, conceptual knowledge and competency level of integrated science process skill towards planning a professional enhancement programme. *Sains Malaysiana*, *41*(7), 921-930.

[13] Ellyza Karim, Jamil Ahmad & Kamisah Osman. (2017). Fuzzy Delphi method for content validation of integrated science process skills instrument. *International Journal of Academic Research in Business and Social Sciences, 7*(6), 773-785. doi: 10.6007/IJARBSS/v7-i6/3037

[14] Ellyza Karim. (2020). *Pembinaan dan Pengesahan Instrumen Kemahiran Proses Sains.* Doctoral dissertation, National University of Malaysia.

[15] Harlen, W. & Elstgeest, J. (1992). *UNESCO sourcebook for science in the primary school: A workshop approach to teacher education.* Paris: UNESCO Publishing. Retrieved from http://unesdoc.unesco.org

[16] Harlen, W. (2006). *Teaching, learning and assessing science 5-12*. London: SAGE Publications Ltd.

[17] Juliana Mohd Janjang. (2017). *Kesahan, kebolehpercayaan instrumen Reflective Thinking Questionnaire (RTQ) dan pengukuran tahap pemikiran reflektif dalam kalangan guru pelatih IPG.* Master thesis, National University of Malaysia.

[18] Kazeni, M. M. M. (2005). Development and validation of a test of integrated science process skills for the further education and training learners. Master dissertation, Faculty of Natural and Agricultural Sciences, University of Pretoria, South Africa.

[19] Linacre, J. L. (2002). Optimizing rating scale category effectiveness. *J Appl Meas*. *3*(1), 85-106.

[20] Linacre (2005). *A user's guide to WINSTEPS and MINISTEPS: Rasch-model computer programs.* Chicago: MESA.

[21] Linacre (2007). *A user's guide to WINSTEPS and MINISTEPS: Rasch-model computer programs.* Chicago: MESA.

[22] Nazahiyah Mustafa, Ahmad Zamri Khairani & Nor Asniza Ishak (2021). Calibration of the science process skills among Malaysian elementary students: A Rasch model analysis. *International Journal of Evaluation and Research in Education. 10*(4), 1344-1351.

[23] Ong, E. T., Wong, Y. T., Sopia Mad Yassin, Sadiah Baharom, Asmayati Yahya & Zahid Md. Said. (2012). Malaysian based science process skills inventory: development, validation and utilisation. *CREAM – Current Research in Malaysia*. *1*(1), 125-149.

[24] Ong, E. T. & Mohd. Al-Junaidi Mohamad. (2013). Pembinaan dan penentusahan instrumen kemahiran proses sains untuk sekolah menengah. *Jurnal Teknologi (Social Sciences). 66*(1), 7–20.

[25] Ong, E. T., Norjuhana Mesmen, Siti Eshah Mokshein, Sabri Mohd Salleh, Nik Azmah Nik Yusuff, & Koon, P. Y. (2015). Basic science process skills test for primary schools: Item development and validation**.** *Jurnal Pendidikan Sains & Matematik Malaysia. 5*(1).

[26] Saunders, M. N. K., Lewis, P. & Thornhill, A. (1997). *Research methods for business students.* London: Pitman.

[27] Smith, K. A & Welliver, P. W. (1990). The development of a science process assessment for fourth-grade students*. Journal of Research in Science Teaching*. *27*(8), 727-738.

[28] Temiz, B. K, Tasar, M. F & Tan, M. (2006). 6 and validation of a multiple format test of science process skills. *International Education Journal. 7*(7), 1007-1027.

[29] Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

[30] Wright, B. D. & Stone, M. H. (1999). *Measurement essentials. 2nd ed*. Delaware: Wide Range.

[31] Wright, B. D. & Stone, M. H. (2004). *Making measures.* Chicago: The Phaneron Press.

[32] Yeoh, P. C. & Gan, C. M. (2004). Emphases of Malaysian primary science curriculum. In. Yap K. C., Goh N. K., Toh K. A., & Bak H. K. *Teaching primary science*. Jurong: Pearson Prentice Hall.

[33] Zile-Tamsen, C. V. (2017). Using Rasch analysis to inform rating scale development. *Res High Educ.* doi: 10.1007/s11162-017-9448-0.