International Journal of Instruction
e-ISSN: 1308-1470 ● www.e-iji.net

October 2022 ● Vol.15, No.4
p-ISSN: 1694-609X
pp. 855-870

Article submission code: 20211204160911

Received: 04/12/2021
Revision: 28/05/2022

Accepted: 21/06/2022
OnlineFirst: 31/08/2022

Assessment Development in Measuring Students' Context-rich Problem-solving Ability in Physics

Giovanni Pelobillo
University of Mindanao, Philippines, *giovanni_pelobillo@umindanao.edu.ph*

This study developed 17-item physics problem-solving ability to measure the students' performance in solving context-rich problems. In this process, 8 students were interviewed on how they solve physics problems. Qualitative analysis showed 37 latent attributes that describe their ability. Attributes were surveyed to 370 students to establish psychometrics via rating scale-graded response model (RS-GRM) and to ensure psychometric assumptions through Mokken's scale analysis (MSA). Analyses showed that there are 17 valid items or set of outcomes in designing assessment tasks that define what is to be learned as a problem solver. It represents a unidimensional construct (SRMSR=0.043) that provides more information to person ability that ranges from -2.0 to +2.0 values (SE2=0.81). The measurement information generated useful information in describing the students' problem-solving ability. It also demonstrated a 3-ordered response expectations that could represent the solver's actual reasoning as a a cognitive structure. This study noted that the ordered response requires further study and exploration.

Keywords: context-rich problem, problem-solving ability, mokken's scale analysis, rating scale-graded response model, assessment development

INTRODUCTION

Studies on assessing problem-solving have emerged since the conception of giving quality information to both students and teachers. Physics teachers highly value the ways of evaluating the students' ability to solve problems (Adams & Wieman, 2007). Its effect is relevant in terms of attitude (Prince, 2004), informing the targeted outcomes towards curriculum improvements, and identification between novice, and expert problem solvers (Larkin et al., 1980). However, most supplementary books used by the teachers may seem problematic because of their traditional approach as it draws repetition (Hollingworth & McLoughlin, 2001) and exercise. Its oversimplified nature has little significance to real-life problems. Meaning, assessing problem-solving requires the utilization of real-life contexts.

The inclusion of real-life scenarios in word problems is called context-rich. It enables the students to construct reasoning and ideas which makes them more quality and less mathematical (Antonenko *et al.*, 2011). Understanding how students solve context-rich

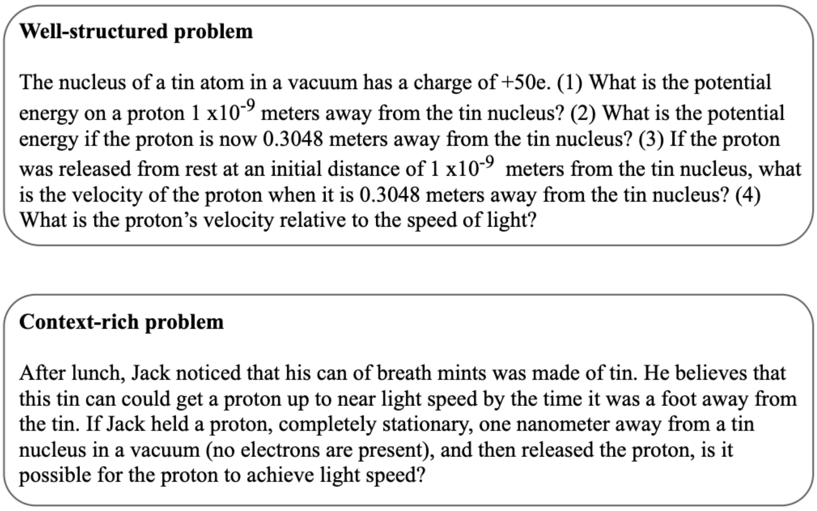
Citation: Pelobillo, G. (2022). Assessment development in measuring students' context-rich problem-solving ability in physics. *International Journal of Instruction*, *15*(4), 855-870. https://doi.org/10.29333/iji.2022.15446a

problems acknowledges the use of attributes perceived as assessment items rather than criteria. Several Scholars (Heller et al., 1992; Huffman, 1997; Murthy, 2007; Ogilvie, 2007) developed scoring criteria and rubrics to examine assessment categories in problem-solving. This strategy would be difficult to use because there could be underlying items in the criteria. Assessment items are the latent attributes to the student's problem-solving ability. It represents their thinking and learning in problem-solving which are interpreted as learning outcomes (Jonassen, 1997) of the construct or the students' context-rich problem-solving ability.

In problem-solving assessment, the utilization of criteria and rubric scoring provides internal consistency of categories with total score measures in classical test theory (CTT). Each person has a true score of a construct if there were no errors in measurement. It is viewed that as the true score increases, the item responses of the same construct should also increase (Cappelleri *et al.*, 2014). This testing is commonly used in scoring student performance that their true scores are based on the number of correct item responses. Meaning, the assumption of the measurement precision of a construct is said to be equal for every item irrespective of attribute levels (Jabrayilov *et al.*, 2016). This study claims that item response theory (IRT) and its properties as measurement should be used because it remains constant regardless of another examination from a sample of a focused population. It gives measurement precision of each item in the construct (Stover *et al.*, 2019) to provide information on its performance. IRT could also psychometrically establish response expectations of assessment items as performance scale in reasoning. This is given by the idea that a problem task involves a contextualized problem that situates the students to think (or schematize), and reason out (Newell, 1980). Hence, it is necessary to establish the assessment items or goals because it enables the alignment of problem-solving ability to the assessment tasks.

Solving Context-rich Problems

Most introductory physics subjects rely on traditional problem-solving that is scored based on the correctness of the numerical solutions. It gives students the partial credit for the solution characteristics compared to the ideal situations. This is attributed to assessing students' performance in well-structured problem-solving. Usually, when students solve well-structured problems, they encounter difficulties further when challenged with multiple and complex tasks (Antonenko *et al.*, 2011). Regrettably, they are generally accepted and widely used by teachers which in turn does not give further description of the students' performance particularly on their ability to solve problems. It supports the Philippine context that Filipino students perform poorly in educational assessments and international tests in physics (Orleans, 2007). Further, the findings of the Program for International Student Assessment (PISA) 2018 showed that Filipino learners gained 357 scientific literacy score that is lower than 489-point average of the Organization for Economic Cooperation and Development (OECD). Hence, using of contextualized problems should be necessary to attract productive student learning in physics at secondary, and university level. These problems are called context-rich or semi-structured which is a sub-category of open-ended problems (Shekoyan *et al.*, 2007). Figure 1 shows an example of well-, and context-rich problem statements based on the example of O'Brien (2014: pp. 23-24).



Well-structured problem

The nucleus of a tin atom in a vacuum has a charge of +50e. (1) What is the potential energy on a proton 1 x10^{-9} meters away from the tin nucleus? (2) What is the potential energy if the proton is now 0.3048 meters away from the tin nucleus? (3) If the proton was released from rest at an initial distance of 1 x10^{-9} meters from the tin nucleus, what is the velocity of the proton when it is 0.3048 meters away from the tin nucleus? (4) What is the proton's velocity relative to the speed of light?

Context-rich problem

After lunch, Jack noticed that his can of breath mints was made of tin. He believes that this tin can could get a proton up to near light speed by the time it was a foot away from the tin. If Jack held a proton, completely stationary, one nanometer away from a tin nucleus in a vacuum (no electrons are present), and then released the proton, is it possible for the proton to achieve light speed?

Figure 1
O'Brien's (2014) sample problem tasks

Unlike well-structured problems, contextualized problems tell a short story that includes a justification for computing specific measures of real objects or phenomena (Heller & Hollabaugh, 1992). It is found that context-rich problems (CRPs) make physics more interesting because it facilitates understanding of physics concepts (Jonsson et al., 2007) meaning, an important aspect of knowledge construction (Santyasa *et. al.*, 2020). It situates the student as a character to a specific context which is represented as a story, targeting a specific and intermediate physics concepts (e.g., thermodynamics, and electricity). Context-rich problems encourage critical thinking that enables the student to examine the context and apply physics principles and math operations to solve the problem. Hence, it must be structured in a way it encourages the student to decide the target variable and determine physics principles and assumptions to simplify the problem towards meaningful solutions.

Assessing Physics Problem-solving

When students solve problems, several authors recommended the use of rubrics to determine their scores. Docktor *et al.* (2016) argued that giving students aggregated scores based on correct answers gives an inadequate description of the students' problem-solving performance. Further, Henderson *et al.* (2004) mentioned that aggregated scores do not provide information to students that are novice and expert solvers. The use of rubrics provides different dimensions of performance which include standards of attainment facilitates scoring across assignments (Jonsson & Svingby, 2007; Mertler, 2001). Several scholars (Heller *et al.*, 1992; Huffman, 1997; Murthy, 2007; Ogilvie, 2007) have developed problem-solving rubrics that present similar themes however, they may vary on their use as criteria across dimensions that can be difficult to use. This led to the development of Minnesota assessment of problem-solving (MAPS) rubric of Docktor (2009). This rubric had been tested and analyzed via different sets of scoring and categories on varied types of problem solutions. The rubric arrived at 5 dimensions namely, specific application of physics, useful description, mathematical progression, mathematical progression, and physics approach. Given its reliability and validity however, the students' ability to solve problems must be given a closer look to provide diagnostic information to assist teachers in helping the students acquire problem-solving ability.

Problem-solving ability or students' attributes are progress variables that conceptualize the developmental perspective of assessment (Wilson, 2008). When students' responses and conceptions are linked to the progress variables, it then defines what is to be learned. This means that the progress variables are set of goals that could situate the solver in a problem space (Wood, 1983) making the problem, goal-oriented (Newell, 1980). Solver's responses rely on their reasoning as a cognitive structure (Biggs & Collis, 1982), and a description of structural thinking (Eseryel *et al.*, 2013). Treating the attributes as progress variables could serve as intended cognitive learning outcomes.

The Use of Item Response Theory

Validity and reliability are important to educational assessments since the conception of test scores in producing information like cognitive performance, and effectiveness of a test. For over 100 years, the implementation of conventional measurements under classical test theory (CTT) has emerged in the testing field (Zanon *et al.*, 2016) due to the problems of simple mathematical models (Ostini & Nering, 2006). CTT assumes a common measurement precision for all individuals regardless of an attribute or item levels. It limits the possibility of group comparison per item levels given the same test with aggregated scores of a particular concept. Further, its sample dependency produces different levels of item difficulty which implies that CTT can only be done on the same sample. It also assumes equal measurement errors for all persons. This problem lies in the idea of different ability levels that will show different errors that evaluate any other construct (Zanon *et al.*, 2016). To address its issue, item response theory (IRT) has emerged as its extension. This kind of modern measurement approach scales items and individuals, and evaluates item characteristics of unidimensional construct (Revicki *et al.*, 2009). It predicts the ability of a person of a specific attribute to establish a relationship between a person's set of traits and performance under a particular item (Hambleton *et al.*, 1991). Further, IRT uses location parameters as item difficulty for dichotomous data. In multiple response formats such as rating scale, and ordered responses, it indicates the category thresholds between responses. Thresholds represent the cut-off points that correspond to a z-score and the level of a latent variable when a person might choose another response (Schivinski *et al.*, 2018). The use of item precision parameter could discriminate individuals across the latent wherein larger discrimination provides more information about the student's latent ability. Hence, IRT assumes that items under a unidimensional construct are unequally informative. Location thresholds, and discrimination parameters determine the accuracy of instrument and measurement via test information function (TIF). It is visualized as a curve that indicates the precision of the entire scale wherein the peak represents the greatest information, and a flat curve represents high discrimination across latent traits. Overall, IRT models build a hypothetical unidimensional line along which the location of items and persons are aligned to the measures of ability and difficulty. In other words, fit items describe a single construct under the IRT assumptions.

METHOD

The use of exploratory sequential design generated a critical examination of students' problem-solving ability. The first phase involved interviews of 8 science students with physics subjects in the University of Mindanao, Davao City, Philippines. The interview and probing questions were anchored to the MAPS rubric by Docktor (2009) since it provided analysis of the aforementioned scholars' problem-solving criteria. Saturation (Legard *et al*., 2003), and coding in the qualitative phase generated 37 latent attributes that describe the students' problem-solving ability which will then be used as assessment items. In phase 2, these items were surveyed to 370 students to ensure the observation of response sufficiency based on the guidelines of Linacre & Wright (1998).

Quantitative analysis used a rating scale-graded response model (RS-GRM) in the family of polytomous IRT. This study also used Mokken's scale analysis (Mokken, 1971) because IRT is known to have strong assumptions of the unidimensionality of the construct. The measures of MSA such as scalability coefficient generates Mokken's Rho to partition items into Mokken's scale through an automated item selection algorithm (Ark, 2007). Its analysis was used in determining the response categories aside from the step ordering and calibration of the rating scale model (Andrich, 1978). Meaning, a uniform response category is a fundamental assumption of RSM because the attributes or items are tied to the latent trait (Ostini & Nering, 2006). RSM was then used to provide item-fit statistics such as infit and outfit based on the guidelines of Linacre & Wright (1998), and Smith (2000). Representing assessment items as learning expectations, moving the analysis from rating scale model to graded response model (GRM) was necessary. GRM variant of (Muraki, 1990) was used to further give item fit statistics such as RMSEA, and S-X^2 (Pearson X^2 statistic) p-values of each items to examine the model through the guidelines of Browne and Cudeck (1992), and Orlando & Thissen (2000, 2003).

In IRT, the student's problem-solving ability represents a latent trait (θ) and a continuum from for example, -2 to +2 where the average score is 0, and the greater its value, the greater item difficulty. GRM was used to determine the threshold parameters as item difficulty, denoting lower values as easier ability on each item scale. Further, its analysis involved slope parameters in each item to represent discrimination in measuring item differential capability (Watanabe *et al.*, 2017). Lastly, test information function (TIF) was used to determine the reliability of how the latent trait provides information to a specific range of person ability.

FINDINGS AND DISCUSSION

Checking IRT Assumptions

Based on qualitative analysis of 8 recorded interviews, there are 37 items as latent attributes of problem-solving ability. Treating the attributes as a single latent construct allowed the use of IRT as a measurement theory. Hence, these items were surveyed to 370 science students. Based on the guidelines of Linacre & Wright (1998) that each should have 10 or more response observations, the scale was optimized to a 3-point Likert scale. The attributes then, underwent Mokken's scale analysis (MSA) to check the IRT assumptions. This kind of analysis has less restrictive assumptions with less-demanding data to maintain the measurement properties before proceeding to parametric and non-parametric IRT (Ark, 2007). Unidimensionality, local dependence, and latent monotonicity were used in checking the assumptions with hypothesis testing on the item (Hi) and test (H) scalability coefficients. These coefficients convey which total score could rank accurately the persons on latent trait through test scalability coefficient (Ark, 2007). Further, the monotonicity of MSA was utilized to determine the number of violation of assumptions (crit. statistic). Hence, it generated score reliability of Mokken's rho and unbiased reliability estimator in partitioning items into Mokken's scale through aisp. Table 1 shows the results.

Table 1
MSA measures

Item no.	Attributes	Hi	# vi	crit.	aisp
1	Organizing the gathered information	0.45	21	43	1
2	Compare and contrast of solutions to the problem	0.4	21	48	1
3	Justifying the quality of the solution	0.47	26	56	1
4	Remembering what's going on in the problem situation	0.44	18	25	1
5	Determining formulas useful to the solution process	0.42	17	59	1
6	Compare and contrast of shared experiences	0.39	21	53	1
7	Checking the quality of the solution	0.49	19	35	1
8	Gaining awareness of the solution requirements	0.46	17	47	1
9	Agreeing to the problem solution and quality	0.44	28	64	1
10	Explaining the solution validity	0.49	24	60	1
11	Recalling the past experience related to the problem	0.37	26	62	1
12	Reviewing the basic and prior concepts in relation to the problem	0.43	20	43	1
13	Justifying the problem solving process	0.49	23	54	1
14	Justifying the concept using personal experience	0.45	9	39	1
15	Telling personal experience in order	0.44	18	54	1
16	Giving concepts and examples related to the problem	0.46	8	43	1
17	Manipulating equations as mathematical solution process	0.4	17	50	1
18	Evaluating the problem scenario	0.44	16	27	1
19	Compare and contrast of the quality of the solution	0.49	18	51	1

Table 1
Continued

Item no.	Attributes	Hi	# vi	crit.	aisp
20	Checking the validity of the solution	0.46	17	24	1
21	Asking self-regulatory questions	0.26	39	109	0
22	Elaborating the problem scenario	0.47	17	26	1
23	Identifying what the problem is asking	0.47	14	59	1
24	Evaluating the quality of the solution	0.5	24	67	1
25	Extending examples related to the concept	0.43	10	21	1
26	Checking errors in the solution process	0.4	15	49	1
27	Asking further questions	0.32	32	70	1
28	Incorporating symbols in the illustration	0.44	12	24	1
29	Arguing with the gathered concepts	0.4	20	26	1
30	Giving physics-related assumptions	0.42	14	29	1
31	Illustrating free body diagrams	0.41	16	26	1
32	Explaining the concepts of example situations	0.49	10	40	1
33	Illustrating the problem scenario	0.49	19	59	1
34	Finding out misconceptions in the solution process	0.47	15	43	1
35	Determining the variables of the problem	0.44	16	38	1
36	Outlining the concepts and solutions to the problem	0.49	15	19	1
37	Making hand-gestures to imagine the situation	0.33	38	75	1

It shows that the majority of item scalability coefficients are moderately scalable ($0.4 \leq H < 0.5$) which made the whole test also moderately scalable ($H=0.434$). This means that there are some items that may seriously violate the assumptions of monotonicity given with the number of violations such as item 21 (crit.=109). These values are necessary assumptions to decide which items to remove further from the scale (Palmgren et al., 2018). Overall, the MSA assumptions are met based on the Mokken's rho of 0.96. The majority of the items are locally independent, and they belong to a uniform scale via automated item selection (aisp) except item 21 (asking self-regulatory questions).

Rating Scale Model

Demonstration of internal consistency (e.g., point biserial correlations, Cronbach's alpha, and quantity of correct responses out of overall score) of items within a specific category are aggregated measures which is common to classical problem-solving assessment. Meaning, its classical measures vary across samples (Reeve, 2003). In contrast, IRT assumes that the person may have differences in the behavior of selecting a response from the scale that may produce item difficulty measures. Table 2 shows RSM interpretation of the assessment items' performance. It lists the item fit statistics of 37 items such as p-value and mean-square (MNSQ) and z-standardized (ZSTD) of infit and outfit.

Table 2
RSM item fit statistics

Items/Attributes	p-value	Outfit		Infit	
		MNSQ	ZSTD	MNSQ	ZSTD
Organizing the gathered information	0.12	1.08	1.09	0.92	-1.16
Compare and contrast of solutions to the problem[b]	0.02	1.15	1.62	1.02	0.37
Justifying the quality of the solution	0.91	0.90	-1.31	0.92	-1.25
Remembering what's going on in the problem situation	0.79	0.94	-0.80	0.96	-0.64
Determining formulas useful to the solution process	0.20	1.06	0.66	1.10	1.41
Compare and contrast of shared experiences [b]	0.00	1.35	3.60	1.26	3.64
Checking the quality of the solution	0.99	0.82	-2.34	0.86	-2.14
Gaining awareness of the solution requirements	0.53	0.99	-0.10	0.96	-0.51
Agreeing to the problem solution and quality	0.47	1.00	0.05	0.95	-0.69
Explaining the solution validity	0.96	0.87	-1.83	0.90	-1.57
Recalling the past experience related to the problem [b]	0.00	1.22	2.51	1.26	3.60
Reviewing the basic and prior concepts in relation to the problem	0.05	1.13	1.61	1.10	1.48
Justifying the problem solving process	0.98	0.85	-2.01	0.87	-1.90
Justifying the concept using personal experience	0.03	1.15	1.91	1.12	1.73
Telling personal experience in order [c]	0.01	1.17	1.89	1.20	2.68
Giving concepts and examples related to the problem	0.05	1.13	1.64	1.11	1.61
Manipulating equations as mathematical solution process [c]	0.02	1.12	1.85	1.19	2.53
Evaluating the problem scenario	0.20	1.06	0.82	1.05	0.77
Compare and contrast of the quality of the solution	1.00	0.80	-2.77	0.82	-2.85
Checking the validity of the solution	0.87	0.91	-1.08	0.95	-0.72
Asking self-regulatory questions [a]	0.00	2.26	8.50	1.63	7.32
Elaborating the problem scenario	0.78	0.94	-0.81	0.96	-0.63
Identifying what the problem is asking	0.81	0.93	-0.72	1.00	0.05
Evaluating the quality of the solution	1.00	0.76	-3.37	0.78	-3.44
Extending examples related to the concept	0.05	1.13	1.63	1.11	1.50
Checking errors in the solution process [b]	0.01	1.19	2.01	1.21	2.87
Asking further questions [a]	0.00	1.68	6.05	1.38	4.98
Incorporating symbols in the illustration [c]	0.00	1.26	2.97	1.20	2.84
Arguing with the gathered concepts [b]	0.00	1.40	4.40	1.35	4.61
Giving physics-related assumptions [b]	0.00	1.20	2.22	1.21	2.82
Illustrating free body diagrams [a]	0.00	1.27	3.09	1.27	3.69
Explaining the concepts of example situations	0.54	0.99	-0.14	0.93	-1.10
Illustrating the problem scenario	0.19	1.06	0.84	1.07	0.98
Finding out misconceptions in the solution process	0.35	1.02	0.34	1.04	0.58

Table 2
Continued

Items/Attributes	p-value	Outfit		Infit	
		MNSQ	ZSTD	MNSQ	ZSTD
Determining the variables of the problem	0.06	1.12	1.48	1.15	2.14
Outlining the concepts and solutions to the problem	0.97	0.86	-1.88	0.83	-2.68
Making hand-gestures to imagine the situation [a]	0.00	1.54	5.99	1.31	4.27

[a] Items removed in the first iteration. [b] Items removed in the second iteration. [c] Items removed in the third iteration.

Series of item deletion and iteration produced good infit and outfit of the 24 items. Most of the deleted items gained lower p-values (<0.05) because they exceed the "fit" criteria (MNSQ > 1.4, ZSTD > +2.0). The deletion of 13 items resulted in strong test scalability (H=0.50) of 24 items via MSA. The assessment gained a productive measurement to problem-solving with its optimized 3-response category. Further, averaged MNSQ infit (0.98) and outfit (0.98) imply that the 24 items resemble 95% of the latent trait based on marginal reliability as a measure of construct validity. The measurement systemcan distinguish students of high and low problem-solving ability, given with high person reliability of 0.95.

Graded Response Model

The assessment items correspond to the student's problem-solving ability wherein the optimized scale could be uniformly ordered. Meaning, it should be operating equivalently across items, and categories should be the same size across items and not unique to each item (Ostini & Nering, 2006). The context of item location parameters should be based on GRM. In this way, the items are permitted to vary in discrimination to demonstrate the changes in the item location over time while keeping the category boundaries fixed (Ostini & Nering, 2006). Thus, scoring students' problem-solving ability relies on the 3-ordered response expectation. GRM shows the estimation parameters in table 3.

Table 3
GRM item parameters

Attributes	Slope (a)	Thresold		Item-fit		RMSEA
		b1	b2	S-X^2	p.val.	
Organizing the gathered information	1.77	-1.21	0.80	30.77	0.58	0.000
Justifying the quality of the solution	2.01	-1.12	0.58	20.68	0.95	0.000
Remembering what's going on in the problem situation [c]	1.98	-1.13	0.50	45.61	0.22	0.021
Determining formulas useful to the solution process	1.63	-1.47	0.19	39.63	0.27	0.019
Checking the quality of the solution	2.41	-1.01	0.37	26.23	0.66	0.000
Gaining awareness of the solution requirements [c]	2.15	-1.07	0.41	38.14	0.46	0.003

Table 3
Continued

Attributes	Slope (a)	Threshold		Item-fit		RMSEA
		b1	b2	S-X^2	p.val.	
Agreeing to the problem solution and quality	1.81	-1.11	0.78	31.95	0.62	0.000
Explaining the solution validity	2.39	-0.70	0.67	24.11	0.90	0.000
Reviewing the basic and past concepts in relation to the problem [c]	1.64	-0.88	0.77	55.25	0.14	0.025
Justifying the problem solving process	2.45	-0.70	0.66	28.36	0.65	0.000
Justifying the concept using personal experience [b]	1.65	-0.58	1.05	39.75	0.12	0.026
Giving concepts and examples related to the problem	1.63	-0.56	1.03	19.95	1.00	0.000
Evaluating the problem scenario	1.94	-0.72	0.73	43.61	0.25	0.020
Compare and contrast of the quality of the solution	2.64	-0.76	0.67	20.02	0.86	0.000
Checking the validity of the solution	2.28	-0.97	0.39	37.99	0.18	0.025
Elaborating the problem scenario	2.05	-0.81	0.67	40.12	0.38	0.012
Identifying what the problem is asking [c]	2.03	-1.34	0.18	44.99	0.24	0.020
Evaluating the quality of the solution	2.67	-0.93	0.54	31.90	0.37	0.013
Extending examples related to the concept [a]	1.69	-0.74	0.97	60.80	0.12	0.026
Explaining the concepts of example situations	1.96	-0.65	1.03	25.71	0.85	0.000
Illustrating the problem scenario [a]	2.09	-0.57	0.70	62.27	0.04	0.034
Finding out the misconceptions in the solution process	1.91	-0.55	0.91	24.93	0.92	0.000
Determining the variables of the problem	1.68	-0.99	0.58	31.74	0.82	0.000
Outlining the concepts and solutions to the problem	2.21	-0.87	0.73	44.35	0.13	0.027

[a] Items removed in the first iteration. [b] Items removed in the second iteration. [c] Items removed in the third iteration.

Table 3 shows the deletion of 7 items based on GRM iterations because they gained high RMSEA and low S-X^2 p-values, and some items cannot be manifested as problem-solving ability. Although GRM's RMSEA values of 24, and 17 items are just the same given as 0.051, 17 items showed an improved standardized root mean square residual (SRMSR) value of 0.043 than the 24 items (SRMSR=0.048). The reason being, SRMSR is most appropriate in evaluating the goodness of fit (GOF) index with its value of ≤ 0.05 indicating adequate fit that corresponds to the average size of misfit (Maydeu-Olivares, 2013). Retainment of 17 items based on RS-GRM iterations showed an improved RMSEA (0.051) and SRMSR (0.043) which explains 94% of the latent trait.

The discrimination estimates ranged from 1.63 to 2.67 indicating that the items distinguished high and low problem-solving ability because of their positive values. The item difficulty ranges from -1.47 to 1.03 which accounts for the 2 estimated thresholds. In item 24 (evaluating the quality of the solution) for example, its first threshold (b1=-0.93) elicits the probability of affirming fist response category in optimized scale. It means higher its value as in b2=0.53 corresponds to the probability of endorsing the second, and third response category in the scale. Hence, it predicted the person ability and established a relationship between the performance and person trait in every item.

Measurement Information

Based on the measurement precision, figure 2 shows x-axis as estimated problem-solving score or person ability depicted on a z-score. Test information function is inversely proportional to standard errors (SE), making SE=1/(information)1/2. This means more information, the better the measurement precision (Revicki *et al.*, 2009) with smaller measurement errors (SE<<0.5).

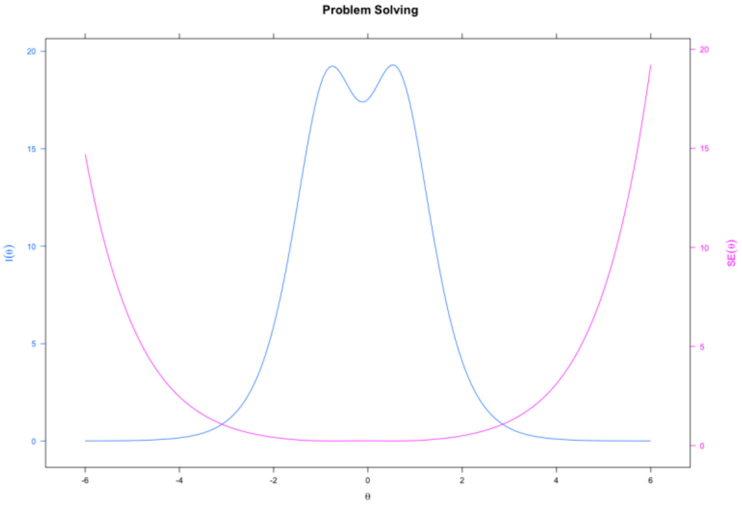
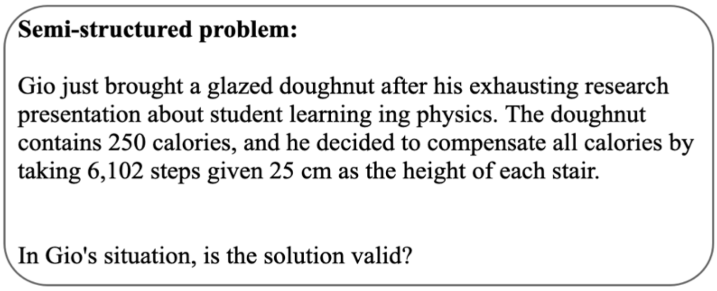


Figure 2
Test information function (IΘ) with standard error (SEΘ)

The figure shows that a test information of 5 provides 0.81 reliability (SE^2). Meaning, the 17 attributes provide most of the information to the person ability from -2.0 to +2.0 given the range of item difficulty. The measurement system of the problem-solving ability is precise, and able to identify between novice, and expert problem solvers.

IMPLICATIONS

These valid set of assessment items belong to the cognitive problem-solving ability that represents the set of outcomes or progress variables (Wilson, 2008). These outcomes could be used in stimulating the solver to respond, and schematize to a contextualized situation hence, making it a goal-oriented problem. Figure 3 shows an example.



Semi-structured problem:

Gio just brought a glazed doughnut after his exhausting research presentation about student learning ing physics. The doughnut contains 250 calories, and he decided to compensate all calories by taking 6,102 steps given 25 cm as the height of each stair.

In Gio's situation, is the solution valid?

Figure 3
Problem-solving task

In the figure, a context-rich problem involves a contextualized situation, and a question that is anchored to the valid set of assessment items as in item 10 (explaining the validity of the solution). It tells a story that attracts the students to reason out and calculate quantities to facilitate conceptual understanding (Heller and Hollabaugh, 1992; Jonsson and Svingby, 2007). These items could be also anchored when designing activities in physics courses such as collaborative problem solving and project-based learning because it could monitor and contribute to learner's progress (Baran *et al.*, 2018) Its implication to monitoring progress in problem-based learning conceptualizes the developmental perspective of formative assessment. Meaning, they are set of items or outcomes that define what is to be learned as a problem solver. It reflects a goal-oriented problem (Newell, 1980), reasoning as a cognitive structure based on Biggs & Collis (1982) taxonomy framework, and a description of structural thinking (Eseryel et al., 2013). Thus, it could provide student diagnosis as a useful information to guide teachers in teaching problem-solving.

CONCLUSION

The use of rating scale-graded response model can be used as a measurement method in developing and validating assessment items as a set of outcomes. Moving the analysis from checking the IRT assumptions to RSM, and to GRM generated a series of iterations that produced a valid set of 17 items with an optimized scale. Measurement information such as discrimination, and difficulty of an item provide a useful information in describing the students' problem-solving ability that is attributed to uniformly ordered response expectations. The 3-ordered response should be further and qualitatively explored in the context of the solver's actual reasoning as cognitive structure.

REFERENCES

Adams, W. K. & Wieman, C. E. (2007). Problem solving skill evaluation instrument-validation studies. *AIP Conference Proceedings, 883,* 18–21. https://doi.org/10.1063/1.2508681

Andrich, D. (1978). Application of a Psychometric Rating Model to Ordered Categories Which Are Scored with Successive Integers. *Applied Psychological Measurement*, *2*(4), 581–594. https://doi.org/10.1177/014662167800200413

Antonenko, P. D., Ogilvie, C. A., Niederhauser, D. S., Jackman, J., Kumsaikaew, P., Marathe, R. R. & Ryan, S. M. (2011). Understanding student pathways in context-rich problems. *Education and Information Technologies*, *16*(4), 323–342. https://doi.org/10.1007/s10639-010-9132-x

Ark, L. A. van der. (2007). Mokken Scale Analysis in R. *Journal of Statistical Software*, 20(11), 1–19. https://doi.org/10.18637/jss.v020.i11

Baran, M., Maskan, A., & Yasar, S. (2018). Learning physics through project-based learning game techniques. *International Journal of Instruction*, *11*(2), 221–234. https://doi.org/10.12973/iji.2018.11215a

Biggs, J. B. & Collis, K. F. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. New York: Academic Press.

Browne, M. W. & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Cappelleri, J. C., Jason Lundy, J. & Hays, R. D. (2014). Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Docktor, J. (2009). Development and Validation of a Physics Problem-Solving Assessment Rubric. *In Dissertation* (Issue September). https://conservancy.umn.edu/handle/11299/56637

Docktor, J. L., Dornfeld, J., Frodermann, E., Heller, K., Hsu, L., Jackson, K. A., Mason, A., Ryan, Q. X. & Yang, J. (2016). Assessing student written problem solutions: A problem-solving rubric with application to introductory physics. *Physical Review Physics Education Research*, *12*(1), 1–18. https://doi.org/10.1103/PhysRevPhysEducRes.12.010130

Eseryel, D., Ifenthaler, D. & Ge, X. (2013). Validation study of a method for assessing complex ill-structured problem solving by using causal representations. *Educational Technology Research and Development*, *61*(3), 443–463. https://doi.org/10.1007/s11423-013-9297-2

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. SAGE Publications, Inc.

Heller, P. & Hollabaugh, M. (1992). Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups. *American Journal of Physics*, *60*(7), 637–644. https://doi.org/10.1119/1.17118

Heller, P., Keith, R. & Anderson, S. (1992). Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving. *American Journal of Physics, 60*(7), 627–636. https://doi.org/10.1119/1.17117

Henderson, C., Yerushalmi, E., Kuo, V. H., Heller, P. & Heller, K. (2004). Grading student problem solutions: The challenge of sending a consistent message. *American Journal of Physics*, *72*(2), 164–169. https://doi.org/10.1119/1.1634963

Hollingworth, R. W. & McLoughlin, C. (2001). Developing science students' metacognitive problem solving skills online. *Australasian Journal of Educational Technology*, *17*(1), 50–63. https://doi.org/10.14742/ajet.1772

Huffman, D. (1997). Effect of explicit problem solving instruction on high school students' problem-solving performance and conceptual understanding of physics. *Journal of Research in Science Teaching*, *34*(6), 551–570. https://doi.org/10.1002/(SICI)1098-2736(199708)34:6<551::AID-TEA2>3.0.CO;2-M

Jabrayilov, R., Emons, W. H. M. & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement,* *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Jonassen, D. H. (1997). Instructional design models for well-structured and III-structured problem-solving learning outcomes. *Educational Technology Research and Development*, *45*(1), 65–94. https://doi.org/10.1007/BF02299613

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Jonsson, G., Gustafsson, P. & Enghag, M. (2007). Context Rich Problems as an Educational Tool in Physics Teaching–A Case Study. *Journal of Baltic Science Education*, *6*(2), 26–34. http://oaji.net/articles/2014/987-1404288026.pdf

Larkin, J., McDermott, J., Simon, D. P. & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science*, *208*(4450), 1335–1342. https://doi.org/10.1126/science.208.4450.1335

Legard, R., Keegan, J., Ward, K.: In-depth interviews. In: Ritchie, J., Lewis, J. (eds.) Qualitative Research Practice: A Guide for Social Science Students and Researchers, pp. 139–169. Sage, London (2003)

Linacre, J. M. & Wright, B. D. (1998). *A User's Guide to BIGSTEPS: Rasch model computer program*. MESA Press.

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement: Interdisciplinary Research & Perspective*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research and Evaluation*, *7*(25), 2000–2001. https://doi.org/https://doi.org/10.7275/gcy8-0w24

Mokken, R. J. (2011). *A Theory and Procedure of Scale Analysis: With Applications of Political Research*. De Gruyter Mouton. https://doi.org/10.1515/9783110813203

Muraki, E. (1990). Fitting a Polytomous Item Response Model to Likert-Type Data. *Applied Psychological Measurement*, *14*(1), 59–71. https://doi.org/10.1177/014662169001400106

Murthy, S. (2007). Peer-assessment of homework using rubrics. *AIP Conference Proceedings, 9*51, 156–159. https://doi.org/10.1063/1.2820920

Newell, A. (1980). Physical Symbol Systems. *Cognitive Science*, *4*(2), 135–183. https://doi.org/10.1016/S0364-0213(80)80015-2

O'Brien, J. R. (2014). Physics education research: *The effects of context-rich problem solving in groups in introductory electricity and magnetism courses* [Worcester Polytechnic Institute]. https://digitalcommons.wpi.edu/mqp-all/3533

Ogilvie, C. A. (2007). Moving Students From Simple to Complex Problem Solving. *In Learning to Solve Complex Scientific Problems* (pp. 159–185). Taylor & Francis Group, LLC. https://doi.org/10.4324/9781315091938

Orlando, M. & Thissen, D. (2000). Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Orlando, M. & Thissen, D. (2003). Further Investigation of the Performance of S - X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *27*(4), 289–298. https://doi.org/10.1177/0146621603027004004

Orleans, A. V. (2007). The condition of secondary school physics education in the Philippines: Recent developments and remaining challenges for substantive improvements. *Australian Educational Researcher*, *34*(1), 33–54. https://doi.org/10.1007/BF03216849

Ostini, R. & Nering, M. (2006). *Polytomous Item Response Theory Models*. SAGE Publications, Inc. https://doi.org/10.4135/9781412985413

Palmgren, P. J., Brodin, U., Nilsson, G. H., Watson, R. & Stenfors, T. (2018). Investigating psychometric properties and dimensional structure of an educational environment measure (DREEM) using Mokken scale analysis – a pragmatic approach. *BMC Medical Education*, *18*(1), 235. https://doi.org/10.1186/s12909-018-1334-8

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*, *93*(3), 223–231. https://doi.org/10.1002/j.2168-9830.2004.tb00809.x

Reeve, B. B. (2003). Item response theory modeling in health outcomes measurement. *Expert Review of Pharmacoeconomics & Outcomes Research*, *3*(2), 131–145. https://doi.org/10.1586/14737167.3.2.131

Revicki, D. A., Chen, W.-H., Harnam, N., Cook, K. F., Amtmann, D., Callahan, L. F., Jensen, M. P. & Keefe, F. J. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain, 146*(1), 158–169. https://doi.org/10.1016/j.pain.2009.07.029

Santyasa, I. W., Rapi, N. K., & Sara, I. W. W. (2020). Project based learning and academic procrastination of students in learning physics. *International Journal of Instruction*, *13*(1), 489–508. https://doi.org/10.29333/iji.2020.13132a

Schivinski, B., Brzozowska-Woś, M., Buchanan, E. M., Griffiths, M. D. & Pontes, H. M. (2018). Psychometric assessment of the Internet Gaming Disorder diagnostic criteria: An Item Response Theory study. *Addictive Behaviors Reports*, 8(May), 176–184. https://doi.org/10.1016/j.abrep.2018.06.004

Shekoyan, V., Etkina, E., Hsu, L., Henderson, C. & McCullough, L. (2007). Introducing Ill-Structured Problems in Introductory Physics Recitations. *AIP Conference Proceedings*, *951*, 192–195. https://doi.org/10.1063/1.2820930

Smith E. V., Jr (2000). Metric development and score reporting in Rasch measurement. *Journal of applied measurement*, *1*(3), 303–326. https://pubmed.ncbi.nlm.nih.gov/12029173/

Stover, A. M., McLeod, L. D., Langer, M. M., Chen, W.-H. & Reeve, B. B. (2019). State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. *Journal of Patient-Reported Outcomes*, *3*(1), 50. https://doi.org/10.1186/s41687-019-0130-5

Watanabe, Y., Madani, A., Ito, Y. M., Bilgic, E., McKendy, K. M., Feldman, L. S., Fried, G. M. & Vassiliou, M. C. (2017). Psychometric properties of the Global Operative Assessment of Laparoscopic Skills (GOALS) using item response theory. *The American Journal of Surgery*, *213*(2), 273–276. https://doi.org/10.1016/j.amjsurg.2016.09.050

Wilson, M. (2008). Cognitive Diagnosis Using Item Response Models. *Zeitschrift Für Psychologie / Journal of Psychology, 216*(2), 74–88. https://doi.org/10.1027/0044-3409.216.2.74

Wood, P. K. (1983). Inquiring Systems and Problem Structure: Implications for Cognitive Development. *Human Development, 26*(5), 249–265. https://doi.org/10.1159/000272887

Zanon, C., Hutz, C. S., Yoo, H. & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica, 29*(1), 18. https://doi.org/10.1186/s41155-016-0040-x