Research Article doi: 10.12973/eu-jer.10.2.825

European Journal of Educational Research
Volume 10, Issue 2, 825 - 840.
ISSN: 2165-8714
http://www.eu-jer.com/

Implementation of Four-Tier Multiple-Choice Instruments Based on the Partial Credit Model in Evaluating Students' Learning Progress

Lukman Abdul Rauf Laliyo
Universitas Negeri Gorontalo, INDONESIA

Syukrul Hamdi*
Universitas Negeri Yogyakarta, INDONESIA

Masrid Pikoli
Universitas Negeri Gorontalo, INDONESIA

Romario Abdullah
Universitas Negeri Gorontalo, INDONESIA

Citra Panigoro
Universitas Negeri Gorontalo, INDONESIA

Received: May 5, 2020 ▪ Revised: November 23, 2020 ▪ Accepted: March 23, 2021

Abstract: One of the issues that hinder the students' learning progress is the inability to construct an epistemological explanation of a scientific phenomenon. Four-tier multiple-choice (hereinafter, 4TMC) instrument and Partial-Credit Model were employed to elaborate on the diagnosis process of the aforementioned problem. This study was to develop and implement the four-tier multiple-choice instrument with Partial-Credit Model to evaluate students' learning progress in explaining the conceptual change of state of matter. This research applied a development research referring to the test development model by Wilson. The data were obtained through development and validation techniques on 20 4TMC items tested to 427 students. On each item, the study applied diagnostic-summative assessment and certainty response index. The students' conceptual understanding level was categorized based on the combination of their answer choices; the measurement generated Partial-Credit Model for 1 parameter logistic (IPL) data. Analysis of differences was based on the student level class using Analysis of Variants (One-way ANOVA). This study resulted in 20 valid and reliable 4TMC instruments. The result revealed that the integration of 4TMC test and Partial-Credit Model was effective to be treated as the instrument to measure students' learning progress. One-way ANOVA test indicated the differences among the students' competence based on the academic level. On top of that, it was discovered that low-ability students showed slow progress due to the lack of knowledge as well as a misconception in explaining the Concept of Change of State of Matter. All in all, the research regarded that the diagnostic information was necessary for teachers in prospective development of learning strategies and evaluation of science learning.

Keywords: *Learning progress, four-tier, change of state of matter, partial-credit model.*

To cite this article: Laliyo, L.A.R., Hamdi, S., Pikoli, R., Abdullah, M., & Panigoro, C. (2021). Implementation of four-tier multiple-choice instruments based on the partial credit model in evaluating students' learning progress. *European Journal of Educational Research, 10*(2), 825-840. https://doi.org/10.12973/eu-jer.10.2.825

Introduction

Central to the notion of science learning is the development of students' scientific understanding of basic concepts of sciences (Hadenfeldt et al., 2013), particularly, change of state of matter (Emden et al., 2018). Aside from the issue, several studies have also highlighted the students' inability to provide an epistemological explanation of basic concepts of sciences (Chi et al., 2018). Efforts to solve the issues, however, have shown little progress, as the students might have more complex perceptions regarding the alternative concept they understand (Morell et al., 2017).

Education practitioners have recommended the utilization of learning progress concept as the instructional method to provide guidance and direction and to adjust the curriculum with the learning process and assessment (Claesgens et al., 2009; Duncan & Hmelo-Silver, 2009; Rogat et al., 2011). Learning progress is defined as a sophisticated and systematic way of thinking. This method is applicable for a learning process, in which the students will undergo gradual progress when learning a topic in a long duration. Its effectiveness is highly dependent on the learning process and the students' learning experience (Duschl et al., 2011). The concept involves certain sets of gradual levels that represent conceptual understanding, ranging from low level up to comprehensive level.

The notion of learning progress is highly distinctive to each student and is dependent to one's learning experience (Rogat et al., 2011); therefore, there is no learning roadmap that is suitable for all kinds of students (Smith et al., 2006).

* **Corresponding author:**
Syukrul Hamdi. Universitas Negeri Yogyakarta, Indonesia. ✉ syukrulhamdi@uny.ac.id

Each student constructs one's understanding in a different way; moreover, the construction process is varied depending on the students' conceptual understanding level (Aktan, 2013). This is to say that each student undergoes a different rate of learning progress, understanding level, and knowledge construction. Simply put, the development of scientific comprehension among students is not linear (Neumann et al., 2013). Therefore, this study regards each level of students' conceptual understanding as a success in progressing for more advanced level of understanding (Hadenfeldt et al., 2013). A student who faces difficulty in a certain level of understanding will see a lack of progress to a more advanced level. This in turn hinders the student's ability to construct an epistemological explanation on the basic concepts of science. Within this context, the learning progress is treated as the method to evaluate students' conceptual understanding. The diagnostic information generated is reliable to be treated as a reference for the teachers in developing accurate and valid instructional components to guide the students to progress to the next level.

Among the diagnostic instruments that are considered applicable is the four-tier multiple-choice (4TMC) instrument. It is the development of two-tier multiple-choice test recommended by Treagust (1988) and Chandrasegaran et al., (2007). The use of two-tier instrument is familiar in identifying students' understanding in select topics such as electrochemistry (Lu & Bi, 2016), covalent bond (Peterson et al., 1989), and chemical equilibrium (Tyson et al., 1999). Despite its reputation in academia, the two-tier test has raised criticism due to its sole focus on the facts and negligence towards students' understanding (Klassen, 2006). Therefore, several experts propose the renewed version of the test by adding distractor answer choices to strengthen the diagnostic value of the items (Herrmann-Abell & DeBoer, 2016). In addition, some have highlighted the test's weakness in cases where students' tended to pick the answer choice and the reasoning randomly. This illustrates that the students were uncertain and possessed several misconceptions in the first tier question. In such cases, teachers faced difficulty in differentiating between guessed answers and misconceptions (Habiddin & Page, 2019; Hasan et al., 1999).

The criticism laid against the model has sparked the innovation of three-tier and four-tiers instruments. Both instruments feature two multi-level questions, also similar with two-tier test. In the three-tier test, however, the measurement of students' certainty level is conducted simultaneously in both first and second-tier questions; in the meantime, the measurement is conducted separately in the first two tiers (Caleon & Subramaniam, 2010). The value of students' certainty rate ranges from one (very uncertain) to five (very certain).

Three-tier test lacks validity in measuring the students' certainty rate regarding both the answer choice and the reasoning, whether or not the value of certainty rate refers only to the answer choice, to the reasoning, or both. Such weakness will in turn obstructs the evaluation and classification process of students' responses (Arslan et al., 2012). In the four-tier instrument, the measurement of certainty rate also involves the answer choice in the first tier and the reasoning in the third tier (Arslan et al., 2012). Regarding this feature, four-tier test is considered more accurate than the three-tier test. Students who pick wrong answer choices with high certainty indicate that they have a very high misconception on the measured item (Hoe & Subramaniam, 2016).

Four-tier instruments are used in studies discussing topics such as physics education (Caleon & Subramaniam, 2010), chemical thermodynamics (Sreenivasulu & Subramaniam, 2013), transition metal (Sreenivasulu & Subramaniam, 2013), acid-base reaction (Hoe & Subramaniam, 2016), and chemical kinetics (Habiddin & Page, 2019). However, it is worth noticing that studies on chemistry topic which employ four-tiers instruments tend to focus on describing alternative conception. To put it another way, the higher the certainty rate is, the stronger the students' alternative conception will be. Despite its potentials, the scholarly discussion has overlooked the implementation of a four-tier diagnostic instrument to measure students' learning progress. Therefore, further analysis is essential on the application of 4TMC test in several domains analyzes by Partial-Credit Model approach.

The use of Partial-Credit Model has been introduced since the 2000s in the science education research; it features the instrument that integrates diagnostic assessment and summative assessment (Liu, 2012). On top of that, the diagnostic assessment approach is introduced to conduct an in-depth analysis of the construction process of students' conceptual understanding (Claesgens et al., 2009; Hadenfeldt et al., 2013; Lu & Bi, 2016). This study employs 4MTC and Partial-Credit Model as a diagnostic tool to evaluate students' learning progress in explaining the change of state of matter, besides focusing on the Concept of Change of State of Matter, this research employs in-depth analysis using Item Response Theory, namely Partial Credit Model.

One of the features of the Partial-Credit Model is that the model facilitates one to identify any correlation between the construct map and the students' competence in ways that the students' competence can be analyzed by referring to the difference in item difficulty level. The 4TMC instrument indicates that there are students with very high ability as well as students with low ability in each group. Such a gap serves as the basis for qualitative interpretation to elaborate on the difference in students' competence. The insight is applicable in the learning process of chemistry subject. The instrument is expected to be beneficial for teachers in developing a formative test to identify the students' progress of conceptual understanding. On top of that, teachers are able to implement the instrument as a diagnostic instrument to evaluate students' conceptual understanding in providing feedback on their learning progress. Further, the teachers will be able to develop instructional strategies that are specifically designed to tackle the students' difficulty in developing an epistemological explanation regarding the concept of change of state of matter. The study focus revolves

around three research questions: 1) What is the quality of the developed 4TMC instrument based on the Partial-Credit Model?. 2). How is the effectiveness of 4TMC instrument to evaluate the students' learning progress in explaining concepts of change of state of matter. 3) How is the learning progress in students ranging from the senior high school level up to the senior (fourth) year of college in explaining the concepts?

Methodology

Development Model

This research used a development research referring to the test development model from Wilson. Wilson (2005, 2008) introduces four steps of measurement instrument development in figure 1.

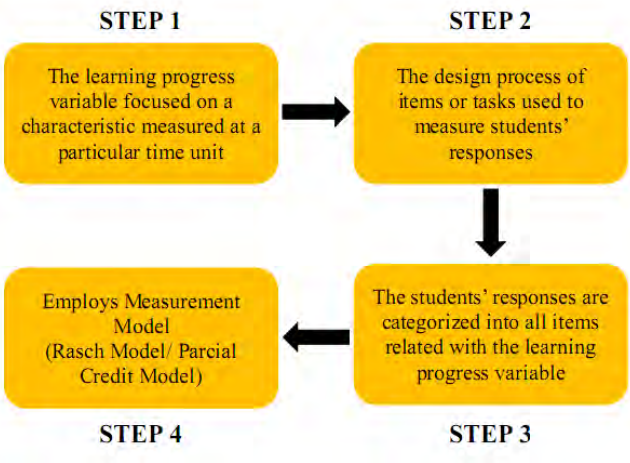


Figure 1. Measurement instrument development

This recommendation is proven valid to be implemented in developing measurement instrument for different construct variables (Chi et al., 2018; Hadenfeldt et al., 2013; Laliyo et al., 2019; Lu & Bi, 2016; Wilson, 2009). The present study also included two questions related to certainty rate (Arslan et al., 2012; Habiddin & Page, 2019: Hasan et al., 1999). The obtained data were analyzed by Partial Credit Model (PCM) approach by WINSTEPS version 4.5.3 software.

Construct Map: Determining Level of Understanding

The first step was to develop the construct of measured variables. The study involved four concepts of change of state of matter: liquid-gas (LG), solid-liquid (SL), solid-gas (SG), and liquid-solid (LS). Gas-liquid (GL) and Gas-Solid (GS) materials were not included in this study as they are included in the basic level of knowledge. The change of a substance from gas to solid (GS) is known as freezing, while from gas to liquid (GL) is called condensing. These two types of changes in the form of substances are very easy to answer by students at a higher level since the materials have always been presented in textbooks, from high school to university students, on the topic of changes in the form of the substance. These concepts were implemented in a gradual manner through five levels of conceptual understanding (Table 1). Such method functions as the pathway of conceptual development that involves learning objectives from the lowest to the highest level of conceptual understanding (Duncan & Hmelo-Silver, 2009; Hadenfeldt et al., 2013; Rogat et al., 2011). In other words, the set of levels, as mentioned previously, was adjusted to the students' needs so as to develop their conceptual understanding. This took into account that each student might progress on different and non-linear development of conceptual understanding; therefore, the levels, as illustrated in Table 1, was considered valid to illustrate the ideal conceptual development pathway (Neumann et al., 2013).

Table 1. Level of conceptual understanding in explaining concept of change of state of matter

Conceptual Understanding Level		Change of State of Matter/Item			
		LG	SL	SG	LS
5	Submicroscopic diagram of change of state of matter	5LG-5	10SL-5	15SG-5	20LS-5
4	Correlation between state of matter and the process of change of state of matter	4LG-4	9SL-4	14SG-4	19LS-4
3	Process of change of state of matter	3LG-3	8SL-3	13SG-3	18LS-3
2	Concept of state of matter	2LG-2	7SL-2	12SG-2	17LS-2
1	Factual phenomenon of state of matter	1LG-1	6SL-1	11SG-1	16LS-1

Description: (LG = liquid-gas, SL = solid-liquid, SG = solid-gas, LS = liquid-gas)

Item Design and Assessment Scheme

The second phase involved an item design. In the 4TMC instrument, all the items consisted of four-tier multiple-choices. To put it another way, each item contains four questions that combine between diagnostic-summative test (Hoe & Subramaniam, 2016; Lu & Bi, 2016; Treagust, 1988) with certainty response index (hereinafter, CRI) test (Arslan et al., 2012; Hasan et al., 1999). The first-tier questions (Q1) aimed to identify whether or not the students understand the content. Moreover, questions in the second tier (Q2) were employed to clarify the students' certainty regarding their answers in the Q1. Third-tier questions (Q3) functioned to diagnose the students' reasoning regarding their answers in the Q1. Further, questions in the second tier (Q4) were employed to clarify the students' certainty regarding their answers in the Q3. Q1 and Q3 questions in each item involved five answer choices; one among them was the correct answer, while three were the distractor, and another answer choice was open-ended answer choice. This open-ended option allows the students to decide the answer by themselves, should they find no correct answer as in accordance with their conceptual understanding. In the meantime, the Q2 and Q4 questions involved two close-ended answer choices; the first choice was for those who are uncertain of their answer, and the second choice was for the students who are very certain of their answer (Arslan et al., 2012). The distractor choices were employed in Q1 and Q3 questions to validate the diagnostic strength of the questions (Herrmann-Abell & DeBoer, 2016). Therefore, in the Q1 and Q3 tiers, the students would have only 0.20 or 20 percent probability of choosing the correct answer. The item Category of Grade of Students' Conceptual Understanding in Table 2.

Table 2 Category of grade of students' conceptual understanding)*

Questions				Conceptual Understanding		Category	Rating
Q1	**Q2**	**Q3**	**Q4**				
Correct	Certain	Correct	Certain	CCCC	Scientific Knowledge	SK	5
Correct	Certain	Incorrect	Certain	CCIC	Misconception False Positive	MFP	4
Incorrect	Certain	Correct	Certain	ICCC	Misconception False Negative	MFN	3
Incorrect	Certain	Incorrect	Certain	ICIC	All-Misconception	AM	2
Correct	Certain	Correct	Uncertain	CCCU	Lack of Knowledge	LOK	1
Correct	Certain	Incorrect	Uncertain	CCIU			
Correct	Uncertain	Correct	Certain	CUCC			
Correct	Uncertain	Correct	Uncertain	CUCU			
Correct	Uncertain	Incorrect	Certain	CUIC			
Correct	Uncertain	Incorrect	Uncertain	CUIU			
Incorrect	Certain	Correct	Uncertain	ICCU			
Incorrect	Certain	Incorrect	Uncertain	ICIU			
Incorrect	Uncertain	Correct	Certain	IUCC			
Incorrect	Uncertain	Correct	Uncertain	IUCU			
Incorrect	Uncertain	Incorrect	Uncertain	IUIU			

(*Hasan, Bagayoko and Kelley, 1999; Arslan, Cigdemoglu and Moseley, 2012; Habiddin and Page, 2019)

As an illustration, in the item 13SG-3, a student picks A in Q1, "very certain" in Q2, A in Q3, and "very certain" in Q4; the combination of the student's answers is ICIC. The result illustrates that the student's answer is incorrect in the Q1 and is very certain of one's error (Q2). Moreover, s/he also provides an incorrect answer in Q3 and is very certain of one's incorrect answer in Q3 (Q4). This indicates that in the item 13SG-3, the student is categorized to have all-misconception understanding (AM). In the Conceptual Understanding Category table, the category is included in fourth grade. Incorporation of the students' answer combinations in each item into the category and grade of students' understanding would result in specific data that are in accordance with the Partial-Credit Model.

Outcome Space and Data Collection

The third step involved the design of the outcome space of the correlation between items and construct maps (Bond & Fox, 2007; Wilson, 2009). The item validation was conducted independently by three expert validators to evaluate the extent of correlation between answer choices in Q1-Q3 in each item and the level of students' conceptual understanding. The validators were asked to clarify that the questions are easy to understand and the students' lack of linguistic competence would not hinder them from providing the right answer. The validators also required to ensure that the questions are in accordance with the syllabus, particularly with the students' conceptual understanding as based on the construct map. The questions in each item were also validated in several aspects, such as: ambiguity, time allocation, directiveness towards a particular answer, and subjective or emotional expression. Fleiss κ measure was employed to acquire information on the validators' approval. From the measure, it was generated that the κ value = 0.97, indicating that the three validators agreed that the 4TMC items were valid in correlating between the answer choices and the students' conceptual understanding.

The next step was to acquire data based on the measurement instrument. The instrument was tested to 427 students in Gorontalo, Indonesia using cluster random sampling technique. The students comprised 171 (40.05%) senior high

school students (or students A), 83 (19.44%) university freshmen majoring chemistry education (or students B), 66 (15.45%) second-year university students majoring chemistry education (or students C), 55 (12.88%) third-year university student majoring chemistry education (or students D), and 52 (12.18%) fourth-year university students majoring chemistry education (or students E). Based on gender, the female participants comprised 369 participants (86.41%), and the male counterparts consisted of 58 participants (13,58%). The participants were given no particular educational treatments and had stated their voluntary consent to participate in the research.

Partial-Credit Model *Measurement and Data Analysis*

The fourth step was to conduct the Partial-Credit Model measurement. This step was implemented to define the correlation between the score generated and the students' conceptual understanding level as elaborated within the construct map. The involvement of Partial-Credit Model measurement lay on the assumption that the item difficulty level is dependent on the students' answer, and that the students' understanding is dependent on the estimation of item difficulty (Linacre, 2012).

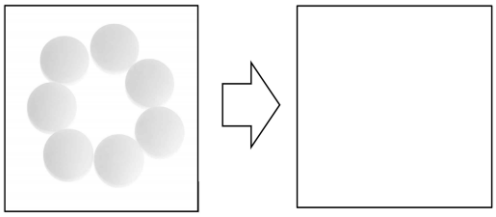
Partial credit model (PCM) was employed to evaluate the learning progress through structured questions; this took into account that the instrument items involved gradual and structured questions (Bond & Fox, 2007; Sumintono & Widhiarso, 2014; Wilson, 2009). The model was stated into the following formula: $\ln[P\_nik/(1 - P\_nik)] - B_n - D\_ik$, in which P_{nik} refers to the probability of student n with B_n ability to pick correct response in the level k of item i; while D_{ik} refers to the difficulty level k of item i, or the threshold point for the test taker who scores k, not k -1. Analysis of differences based on class level of students using One-way ANOVA.

Results

The developed 4TMC instrument adapts the two-level instrument model by Treagust (1988), combined with the CRI theory by Arslan (2012). The function of CRI (certainty response index) is to ensure that students' choice of answers in Q1 and Q3 are the answers that they believe in. This is called diagnostic because it investigates the level of student error in stages, including the ability of students to understand and to use their understanding in explaining the reasons for their choice of answers. Thus, measurement is conducted both at the level of knowledge and reasoning.

The item design referred to the basic criteria to ensure that the students would be able to identify logical reason in Q3 as based on their answer in Q1; moreover, the item design also aimed to clarify the students' certainty of their answers through Q2 and Q4 questions. The 4TMC instrument also allows the students to state their certainty level of Q1 and Q3 answer choices separately. Students with correct understanding regarding the concept of change of state of matter (Q1) and its reasoning (Q3) will pick the "very certain" answer in the Q2 and Q4. If the students are uncertain of their answer regarding the content (Q1) but are certain of the reasoning (Q3), this suggests that the students are able to comprehend the concept/theory but unable to implement such concepts. This study views that it is beneficial to explore potential combinations of Q1/Q3 answer choices and Q2/Q4 certainty rate implementation to provide in-depth elaboration on students' understanding of certain concepts (Habiddin, 2019). The item design is illustrated in Figure 2.

Take a look as this picture!



(a) (b)

After two weeks, the matter in picture (a) turns into picture (b).

Q1. How is the change process of state of matter as described by the previous pictures?

a. Evaporation
b. Naphthalene process
c. Sublimation
d. Melting
e. Depletion

Q2. How certain are you with your answer?

1. Uncertain 2. Very certain

Q3. Why do you choose your answer in the Q1?

a. Because naphthalene occurs in the picture (a), and disappears in the picture (b)
b. Because naphthalene has changed its state to water vapor
c. Because naphthalene is a solid matter that is able to sublimate
d. Because naphthalene is a solid matter that is able to evaporate
e. Other answer...

Q4. How certain are you with your answer in Q3?

1. Uncertain 2. Very certain

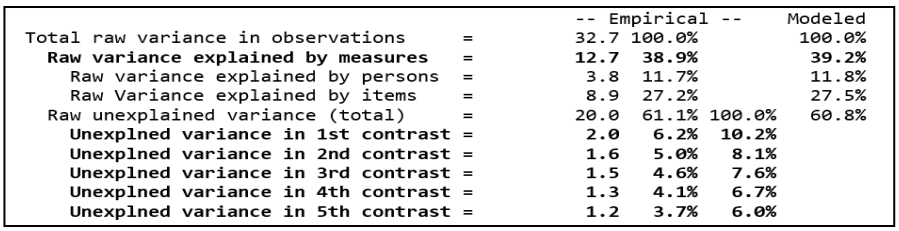
Figure 2. 13SG-3 item design.

Each combination of students' answers in each item was categorized based on the assessment scheme. Every correct response in Q1 and Q3 is labeled with C (correct) code, and wrong answers were labeled with I (incorrect) code. Moreover, in the Q2 and Q4 tier, "very certain" answers are labeled with C (certain) code, and "uncertain" answers were labeled with U (uncertain) code. Therefore, combination of answer choice in each item was generated and written in sequence based on the questions in the Q1, Q2, Q3, and Q4. Such combination was treated as a reference in determining category and grade of students' conceptual understanding in each item, as shown in Table 2. The students' conceptual understanding category was adapted from findings reported by Hasan et al., (1999), Arslan et al., (2012), and Habiddin (2019).

Each combination of students' answer in each item was classified based on five categories of conceptual understanding. The first category is scientific knowledge (SK) with a grade of five; it illustrates that the students possess knowledge that is scientifically correct. The second category is misconception false positive (MFP) (with a grade of four), illustrating that the students have a correct claim of understanding, but they are unable to explain the claim. The third category is misconception false negative (MFN) (with a grade of three), illustrating that the students do not have correct claims of knowledge, but they are able to explain the claim. This category is considered negative because it is possible that the answers provided are guessed answer that is coincidentally correct. The fourth category is all-misconception (AM) with a grade of four; this category signifies that the students are very certain of their incorrect knowledge. Lastly, the fifth category is lack of knowledge (LOK) with a grade of five, signifying that the students lack knowledge in a particular item. Such categories were determined by the students' certainty level in Q2 or Q4. As an instance, one of the possible answer combinations is as follows: CCCU. This combination illustrates that the student is correct in Q1, Q2, and Q3, but is uncertain in Q4; the condition signifies that the student's understanding is ambiguous, hesitant, and is not based on appropriate scientific knowledge.

Effectiveness of Measurement Instruments

Unidimensionality is an essential indicator to evaluate the 4TMC instrument's ability to measure students' capability of explaining the concept of change of state of matter. This indicator is measured by Principal Component Analysis of the residuals to estimate the extent of variance to which the instrument is able to measure what it is supposed to measure (Sumintono & Widhiarso, 2014).

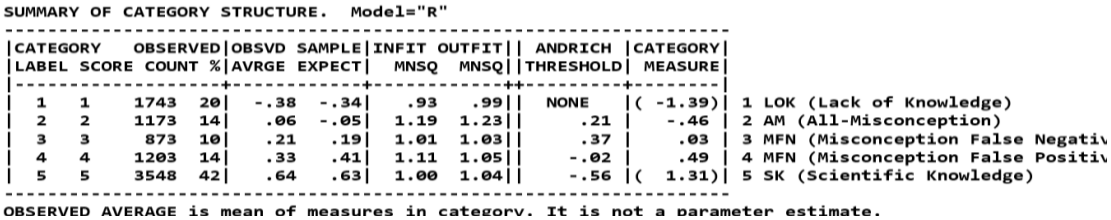


		-- Empirical --			Modeled
Total raw variance in observations	=	32.7	100.0%		100.0%
Raw variance explained by measures	=	12.7	38.9%		39.2%
Raw variance explained by persons	=	3.8	11.7%		11.8%
Raw Variance explained by items	=	8.9	27.2%		27.5%
Raw unexplained variance (total)	=	20.0	61.1%	100.0%	60.8%
Unexplned variance in 1st contrast	=	2.0	6.2%	10.2%	
Unexplned variance in 2nd contrast	=	1.6	5.0%	8.1%	
Unexplned variance in 3rd contrast	=	1.5	4.6%	7.6%	
Unexplned variance in 4th contrast	=	1.3	4.1%	6.7%	
Unexplned variance in 5th contrast	=	1.2	3.7%	6.0%	

Figure 3 Standardized residual variance (in eigenvalue units)

As displayed in the figure 3, the result of raw variance explained by measures of data is 38.9%, the number almost approaches the expectation value of 39.2%. The numbers indicate that the minimum unidimensionality requirements of 20% are achieved, and simultaneously, the limit of PCM unidimension is met (approaching 40%) (Linacre, 2012; Ling Lee et al., 2020). Moreover, the instrument's unexplained variance values are below 7% and considered as ideal (not exceeding 15%), signifying that the item independence rate in instrument falls into "good" category.

Gradual Scale PCM analysis offers a unique verification process on the five grade categories of students' conceptual understanding (see Table 2).



SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
1	1	1743	20	-.38	-.34	.93	.99	NONE	(-1.39)	1 LOK (Lack of Knowledge)
2	2	1173	14	.06	-.05	1.19	1.23	.21	-.46	2 AM (All-Misconception)
3	3	873	10	.21	.19	1.01	1.03	.37	.03	3 MFN (Misconception False Negativ
4	4	1203	14	.33	.41	1.11	1.05	-.02	.49	4 MFN (Misconception False Positiv
5	5	3548	42	.64	.63	1.00	1.04	-.56	(1.31)	5 SK (Scientific Knowledge)

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

Figure 4. Validity of Grade Scale

In Figure 4, it is illustrated that the average observation starts from logit -0.38 in category 1 (lack of knowledge) and increases up to logit +0.64 in category 5 (scientific knowledge). Such finding indicates that the grade category of students' conceptual understanding from 1 to 5 is considered as "very good". Moreover, PCM threshold is also employed to identify the grade's validity; the indicator highlights a transition that occurs in the students' decision making process from one grade to another (Linacre, 2012). The result of PCM-Andrich threshold analysis indicates a consistent increase from the grade 1 to 5, implying that the grade category of conceptual understanding implemented as the evaluation scale to assess the students' competence is categorized as "very good".

Validity

The notion of validity revolves around the question: "does the test measure what it is supposed to measure?". That being said, the developed instrument is considered to have good construct validity if it is able to measure the students' conceptual understanding in explaining the concept of change of state of matter (Linacre, 2012, 2020; Sumintono & Widhiarso, 2014).

Table 3. Item statistics: Misfit order

Item	Measure	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA Corr.
1LG-1	-.86	1.65	4.6	1.62	3.4	.36
8SL-3	-.33	1.18	2.4	1.30	2.7	.44
11SG-1	-.29	1.17	2.3	1.29	2.8	.47
2LG-2	-.35	1.27	3.3	1.29	2.6	.44
5LG-5	.66	1.21	3.3	1.25	2.8	.47
12SG-2	-.21	1.10	1.5	1.14	1.5	.52
6SL-1	.04	1.03	.5	1.12	1.6	.40
19SL-4	-.37	1.08	1.1	.98	-.2	.56
7SL-2	.11	.95	-.9	1.08	1.2	.48
9SL-4	.04	1.01	.2	1.05	.7	.45
16LS-1	-.16	1.00	.1	1.02	.3	.51
4LG-4	.07	.97	-.5	1.01	.1	.48
17LS-2	-.40	1.00	.0	.92	-.7	.51
13SG-3	.30	.99	-.1	.96	-.5	.54
14SG-4	.04	.96	-.8	.88	-1.7	.61

Table 3. Continued

Item	Measure	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PTMEA Corr.
15SG-5	.49	.91	-1.6	.86	-2.0	.55
3LG-3	.15	.85	-3.1	.91	-1.3	.48
20LS-5	.57	.80	-3.8	.90	-1.3	.50
10SL-5	.79	.73	-4.3	.78	-2.5	.48
18LS-3	-.29	.74	-4.0	.72	-3.2	.58

The first step is to ensure that all items match the Partial-Credit Model. Table 3 displays the analysis result of statistic items. The study employs three criteria to measure any misfits or outliers between the students and the items (Linacre, 2012, 2020; Sumintono & Widhiarso, 2014): 1) the accepted Outfit Mean Square (MNSQ) value is between 0.5 < MNSQ < 1.5; 2) the accepted Outfit Z-Standard (ZSTD) value is between -2.0 < ZSTD < +2.0; 3) the accepted Point Measure Correlation (Pt Mean Corr) value is between 0.4 < Pt Measure Corr. < 0.85. As illustrated in figure 4, it is detected that the item 1LG-1 does not meet two criteria (Outfit MNSQ and Outfit ZSTD), while the items 8SL-3, 11SG-1, 2LG-2, and 5LG-5 do not meet the Outfit ZSTD criteria. Moreover, the study does not find any items with negative results in the Point Measure Correlation criteria. This signifies that there is no single item that meets all the three criteria; thus, the measurement instrument possesses good item validity.

The second step is to measure the consistency between the item difficulty level and students' conceptual understanding.

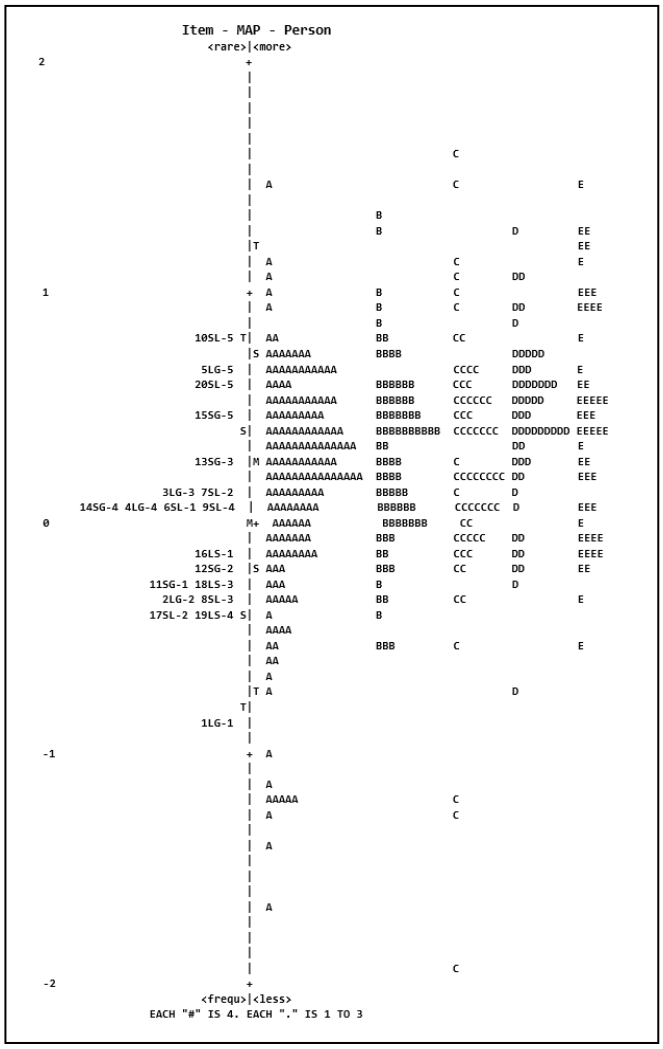


Figure 5 Wright Map: Person-Map-Item
(LG = liquid-gas, SL = solid-liquid, SG = solid-gas, LS = liquid-gas)

Figure 5 displays Wright Map to represent the item difficulty test level and students' conceptual understanding level. The graphical map is a result of empirical analysis on the answer response of the students in each item. According to the Wright Map result, all items in the measurement instrument has majorly encompassed the students' ability. The

result indicates that the most difficult item is 10SL-5 (+0.79 logit), while the easiest item is 1LG-1 (-0.86 logit). However, no equivalent items were found in the understanding level smaller than -0.86 logit (-0.86 to -0.40 logit) as well as the level higher than +0.79 logit; therefore, further investigation is required.

The research discovers several interesting cases regarding the difference between the items and students' conceptual understanding: Firstly, there are four items identified (LG, SL, SG and LS) that measure similar constructs within each level of conceptual understanding. Despite being in the same conceptual understanding level, the items' logit is completely different. For instance, four items were discovered in level 3, each with varying logit (8SL-3 (-0.33) < 18LS-3 (-0.29) < 3LG-3 (+0.15) < 13SG-3 (+0.30)). The numbers indicate that overall, students are more capable of explaining the concept of SL state change compared to LS, LG, and SG. This condition also occurs in the level 4, in which each item has varying logit (19LS-4 (-0.37) < 9SL-4 (+0.04) = 14SG-4(+0.04) < 4LG-4 (+0.07)). Such a finding shows that the students find it easier to explain the correlation between the state of matter and the change process in LS compared to either SL, SG, or LG. Two sample cases above have illustrated that the students' conceptual understanding differs between the change process of LG (evaporation), SG (sublimation), SL (melting), and LS (freezing).

Moreover, it is found that the items in higher conceptual understanding levels tend to have lower logit than those at a lower level. As an instance, the logit of item 19SL-4 in level 4 (-0.37) is smaller than that of item 13SG-3 in level 3 (+0.30). This signifies that students find it harder to explain the item 13SG-3 compared to item 19SL-4. Thirdly, in the same concept of change of state (for example, LS), the logit of item 17LS-2 in level 2 (-0.40) is smaller than that of item 16LS-1 in level 1 (-0.16). As illustrated by the number, students find it easier to explain the SL concept in level 2 rather than to explain the concept's macroscopic fact in level 1. The findings above indicate that the students' conceptual understanding is not consistent with the item sequence. Moreover, the findings also suggest that the item difficulty level (LG, SL, SG, and LS; particularly SL (melting) and LS (freezing)) do not match the level in the construct map.

Measurement reliability

In Partial-Credit Model analysis, the indicator of reliability is observed from the quality of students' response patterns, the instrument, and the interaction between person-item. Within this study, item separation and person separation values are employed as the indicators. The separation index is also converted to Cronbach-equivalent value with an estimation of 0-1. The summary of measurement instrument statistics is displayed in Table 4.

Table 4. Summary of fit statistics

	Student (N=427)	**Item (N=20)**
Mean	0.26	0.00
Standard Error	0.02	0.09
Standard Deviation (SD)	0.48	0.41
Reliability	0.82	0.99
Infit mean-square	1.02	1.03
Outfit mean-square	1.05	1.05
Infit ZSTD	0.00	0.00
Outfit ZSTD	0.10	0.30
Point Raw Score to measure correlation	0.99	-0.99
Separation index (reliability)	2.10	9.54

Cronbach's alpha (KR-20): 0.84
Data Points : 8540
Chi-Square : 21173
df : 8091 (p = 0.0000)

From the Table 4, it is generated that the total data points are 8540 with a Chi-square value of 21173 and the degree of freedom (df) of 8091 (p = 0.0000). These numbers indicate that the measurement is deemed as "very good" and "significant". The column of students and item in the table 4 suggest whether or not the students and the item are considered fit. The average measure value of students is +0.26 logit ($\mu > 0.00$), signifying that the students in overall are competent to explain the concept of change of state of matter. If the separation index value of students (+2.10 logit) is inputted into the person strata (H) formula, or H = [(4*separation) + 1]/3, thus, the generated H value = +3.13 (Linacre, 2012; Sumintono & Widhiarso, 2014). The person strata value (H) of 3 suggests that the students are classifiable into three groups of conceptual understanding (high, moderate, and low). On top of that, if the item's separation index value (+9.54) is processed by the same formula (H), the generated value is 13. Such a number shows that the items in the instrument are classifiable into 14 levels of difficulty. Moreover, the data illustrate that the items are deemed accurate and capable of measuring the students' competence in explaining the focused topic.

From the analysis result of students' answer pattern, the research generates infit and outfit MNSQ values of 1.02 and 1.05, respectively, with expectation value of 1.0. This clarifies that the students' answer pattern towards the instrument

is categorized as "good" (Linacre, 2012; Sumintono & Widhiarso, 2014). In addition, the result generates Infit ZSTD and outfit ZSTD value of 0.0 and 0.10, respectively, with an expectation value of 0.0; the numbers depict that the overall students' answer pattern is in accordance with the model. Moreover, the overall reliability of students section is 0.82, categorized as "good". From the instrument item assessment, it is generated that the infit and outfit MNSQ values are 1.03 and 1.05, respectively, with the expectation value of 1.0, and the infit and outfit ZSTD values are 0.0 and 0.3, with the expectation value of 0.0. The numbers suggest that the overall instrument is deemed as "good", proven by the instrument reliability value of 0.99. The KR-20 (Cronbach's alpha) value results in 0.84, thus signifying a good interaction between the students and the item. As acquired from the findings, the actual data in this study have met the Partial-Credit Model requirements, meaning that further analysis is considered as valid to conduct.

Level of Students' Learning Progress

The second problem of the research is: "How is the learning progress of the participants ranging from senior high school to fourth college year in explaining the focused topic?". To elaborate on that matter, the study employs data generated from the development process of 4TMC instrument to measure the students' conceptual understanding level.

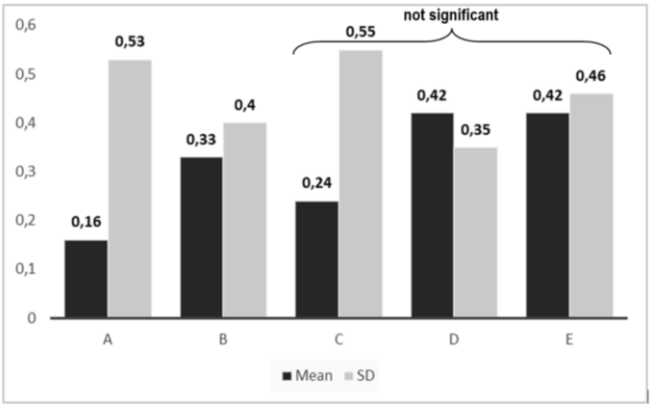


Figure 6. Mean student performance level by grade

(Senior high school students = A, first-year college students = B, second-year college students = C, third-year college students = D, fourth-year college students = E)

Figure 6 displays the average competence calculated in the form of logs based on the students' academic level, ranging from A to E. The figure shows an increasing trend in students' competence development based on their respective academic level (ABCDE). Moreover, it is discovered that the group E shows better learning progress compared to the other groups (D, C, B, and A). Despite that, the One-way ANOVA test indicates a difference among the students' competence based on the academic level, in which F_{count} (6, 0142442) > F_{table} (2,39308); df = 422; p <0.05. The research, therefore, conducted a post hoc Bonferroni test to identify which group that experience significant learning progress. As extracted from the statistical result, group A and B undergo significant learning progress, while group C, D, and E do not experience such significant advancement. This contradicts the common notion that the group CDE are college students with longer formal education experience compared to group A or B. Such finding indicates that the group CDE find it hard to explain the concept of change of state of matter.

Comparison of average competence between groups ABCDE is conducted to map out the difference in the students' learning progress in each conceptual understanding level (displayed in Table 3). The students' competence is calculated based on four items in each level of conceptual understanding. As an example, in the level 1, the students' competence is measured by referring to the mean of item 1LG-1, 6SL-1, 11SG-1, and 16SL-1; the same also applies in the next levels.

Table 5 Measurement of students' average competence in each level of conceptual understanding

Conceptual Understanding Level	**Students' Education Level (Mean, SD)**					
	A (N=171)	**B (N=83)**	**C (N=66)**	**D (N=55)**	**E (N=52)**	**ABDCE (N=427)**
1	0.69 (0.86)	0.80 (0.71)	0.61 (0.91)	1.29 (0.95)	1.05 (0.90)	0.77 (0.86)
2	0.58 (1.04)	0.66 (0.75)	0.68 (0.86)	1.05 (1.00)	0.83 (0.97)	0.69 (0.95)
3	0.19 (0.95)	0.61 (1.00)	0.33 (1.13)	0.84 (0.92)	1.10 (1.24)	0.51 (1.10)
4	0.24 (1.00)	0.53 (0.68)	0.51 (1.12)	0.70 (0.86)	0.51 (0.71)	0.41 (0.57)
5	-1.16 (1.59)	-0.80 (1.46)	-0.86 (1.51)	-0.48 (0.85)	-0.58 (1.51)	-0.84 (1.41)

Based on Table 5, it is found that the students' competence in level 1 (0.77 logit, SD = 0.86) is higher than their competence in level 2 (0.69 logit, SD = 0.95); the same also applies in the next levels. The findings above indicate that the students' conceptual understanding has not developed optimally. On top of that, the item sequence in level 1 is easier to explain compared to that in level 2. The same condition also applies in the next levels. Students find it harder to explain concepts of change of state of matter as the learning progress level increases. Simply put, the students' learning progress level is different in each level of conceptual understanding.

The difference in students' learning progress levels in each conceptual understanding level depicts that longer formal education experience does not necessarily guarantee that the student will have better learning progress in explaining the focused topic. For instance, Table 6 illustrates the comparison of item logit size in level 3 that is calculated based on the students' academic level.

Table 6. Average item logit in level 3

Education Level	N	Item mean (logit) at level 3			
		13SG-3	3LG-3	18LS-3	8SL-3
A	171	0.51	0.33	-0.22	-0.61
B	83	0.55	0.40	-0.43	-0.51
C	66	0.61	0.19	-0.15	-0.66
D	55	0.33	0.20	-0.15	-0.46
E	52	0.57	0.06	-0.30	-0.33

Table 7. Category of item 13SG-3 comprehension

Grade	N	Conceptual Understanding Category - Item 13SG-3 (%)				
		LOK	AM	MFN	MFP	SK
A	171	36	21	3	20	19
B	83	19	36	8	5	31
C	66	36	27	2	8	27
D	55	13	24	4	7	53
E	52	23	12	6	12	48

Category: LOK = Lack of Knowledge, AM = All-Misconception, MFN = Misconception False Negative, MFP = Misconception False Positive, SK = Scientific Knowledge

How is the students' learning progress level in the same item? Table 7 displays the percentage data of students' competence in explaining item 13SG-3 based on five categories of conceptual understanding (LOK, AM, MFN, MFP, and SK). In the SK category, students in group D perform better among all groups (D (53%) > E (48%) > B (31%) > C (27%) >A (19%)). Simply put, more than half students in group D are capable of explaining the item 13SG-3 compared to students in other groups. Meanwhile, in LOK, students in group A and C show higher percentage among all groups (A (36%) = C (36%) > E (23%) > B (19%) > D (13%)). In other words, more than one-third of students in group A or C is incapable of explaining the item 13SG-3 compared to students in other groups due to the limited knowledge on the item. Moreover, in AM, group B shows highest percentage among all groups (B (36%) > C (27%) > D (24%) > A (21%) > E (12%)); it signifies that more than one-fourth of students in group B are incapable of explaining item 13SG-3 compared to other groups due to the misconception on the item. Such findings indicate that the high percentage in LOK and AM category is seen as one of the reasons why the students' competence is different in explaining the same item 13SG-3. To put it another way, the students' learning progress does not develop optimally in explaining item 13SG-3 due to lack of knowledge (LOK) or misconception (AM) on the item.

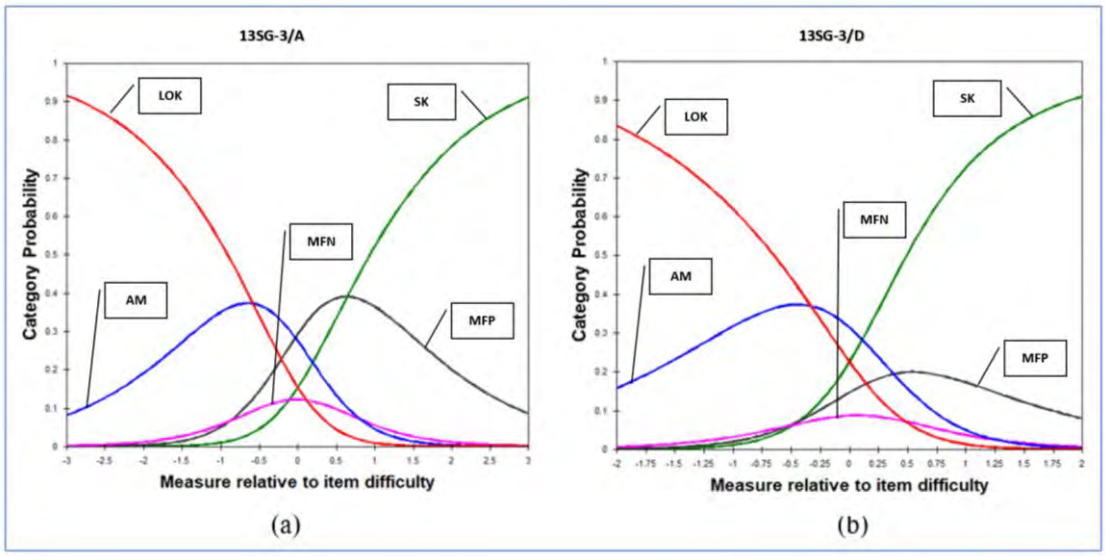


Figure 7. Probability Category Curve of item 13SG-3 of group A, and Probability Category Curve of item 13SG-3 of group D (Category: LOK = Lack of Knowledge, AM = All-Misconception, MFN = Misconception False Negative, MFP = Misconception False Positive, SK = Scientific Knowledge)

Figure 7 illustrates the comparison of the probability category curve (PCC) of students in group A and D in item 13SG-3. The five curve shapes are the visual representation of the distribution of five categories of students' conceptual understanding. From the curves, one can identify which groups that tend to show LOK and AM category traits. It is worth noting that the curve 2(a) and 2(b) tend to be different based on the MFP curve shape, while others are relatively similar. The MFP curve of students A has a higher probability compared to that of students D; simply put, a senior high school student tends to show stronger MFP category compared to a third-year college student. The notion is supported by the finding that senior high school students are relatively incapable of providing correct reason on item 13SG-3 compared to third-year college students. On the other hand, students with low ability in group D tend to show similar curve shape of LOK, AM, and MFN with group A. This implies that both groups' conceptual understanding in the item is relatively similar. In other words, the learning progress of group D, particularly in students with low ability, has not developed optimally despite the fact that that group D consists of third-year college students that progressed through three years of formal education experience in university.

Discussion

The result shows that: firstly, based on the logit size, the items are put in the following order: 13SG-3 > 3LG-3 > 18 SL-3 > 8SL-3. This is to say that it is harder for the students to explain the concept in item 13SG-3 compared to 3LG-3, 18SL-3, and 8SL-3. Secondly, the students' competence in each item is different and not in sequential order based on the education level (ABCDE). The finding leads to an assumption that all students in group E are supposed to perform better in explaining the item sequence in level 3 than those in group D, C, B, and A, since they progressed through longer education experience. However, the calculation result shows a different insight. In the item 13SG-3, students in group C are the most competent among all group (C (0.61) > E (0.57) > B (0.55) > A (0.51) > D (0.33)), while in the item 8SL-3, group E students are the most competent (E (-0.33) > D (-0.46) > B (-0.51) > A (-0.61) > C (-0.66)). Such a finding indicates that the students' competence is varied despite being at the same level. To put it another way, longer formal education tends to have an insignificant effect on the development of students' conceptual understanding.

This echoes previous findings that the learning progress is highly dependent on the students' learning process and experience (Duschl et al., 2011; Park et al., 2017; Wilson, 2009). Learning progress is defined as a sophisticated and systematic way of thinking, in which the students will undergo gradual progress when learning a topic for a long time interval. Students are able to ask questions, form hypotheses, design experiments to test hypotheses, collect data, and draw conclusions (Sutiani et al., 2021). Such a systematic way of thinking is formed by the learning practices and education experience (Emden et al., 2018). Student's way of thinking is affected by learning experience, learning motivation, self regulation and self efficacy to explore understanding of how students go about learning (Haarala-Muhonen et al., 2016; Karagiannopoulou et al., 2020). On top of that, the research findings are in line with previous studies that highlighted that students have distinctive comprehension formed by their own experience (Chi et al., 2018; Emden et al., 2018; Hoe & Subramaniam, 2016; Jin et al., 2019; Rogat et al., 2011; Testa et al., 2019). Such distinctive knowledge has not been explored by evaluation or intervention through learning roadmaps that are in accordance with remedial learning (Smith et al., 2006). In spite of that, it is considered essential to conduct a further analysis that focuses on the modification of conceptual understanding category and analysis variation that is able to define the characteristics of students' alternative conception. The development procedures, as explained in the methodology, has

resulted in 4TMC instrument; however, instrument development is seen as an essential continuous process (Wilson, 2009, 2012).

Based on the research findings, the study identifies several important notes on the development of the 4TMC instrument. Firstly, further analysis of the characteristic of students' response behavior is necessary to conduct regarding the item clarity and the measured concept. The findings have implied that the percentage of LOK and AM understanding category is relatively dominant and tends to increase along with the level of conceptual understanding. Hence, the development of the concept level requires taking into consideration any potential term use that might confuse the students. A further study on the identification of commonly-understood terms or concepts is therefore essential. Secondly, a separate analysis is required to diagnose the factors contributing to the students' lack of knowledge and misconception. Regarding that, further analysis can be conducted by applying the analysis methods developed by previous studies (Caleon & Subramaniam, 2010; Hoe & Subramaniam, 2016). Thirdly, it is discovered that the concepts LG, SG, SL and LS were interpreted differently by the students. Despite being in the same conceptual understanding level, the items' difficulty levels are completely different. Therefore, an evaluation on answer choices requires one to focus on the representation of understanding at the same level.

One of the features of the Partial-Credit Model is that the model facilitates one to identify any correlation between the construct map and the students' competence in ways that the students' competence can be analyzed by referring to the difference in item difficulty level. The 4TMC instrument indicates that there are students with very high ability as well as students with low ability in each group. Such a gap serves as the basis for qualitative interpretation to elaborate on the difference in students' competence. The insight is applicable in the learning process of chemistry subject. The instrument is expected to be beneficial for teachers in developing a formative test to identify the students' progress of conceptual understanding. On top of that, teachers are able to implement the instrument as a diagnostic instrument to evaluate students' conceptual understanding in providing feedback on their learning progress. Providing feedback also improves students' outcome and ability to understand what they learn, increase students ability and creative thinking (Goulas & Megalokonomou, 2021). Through this instrument teacher can give learning feedback to control students learning condition in learning environments both in theory and practice (Dijks et al., 2018; Latifi et al., 2021). Further, the teachers will be able to develop instructional strategies that are specifically designed to tackle the students' difficulty in developing an epistemological explanation regarding the concept of change of state of matter. Through the development of these instructional strategies, teachers will be better able to focus on the goal orientation of learning achievement and motivate students to engage in learning activities (Lee & Keller, 2021; Guo & Leung, 2021; Lin et al., 2021)

Conclusions

The article elaborates on the development and validation procedures of the 4TMC instrument with Partial-Credit Model to evaluate the students' learning progress in explaining the concept of change of state of matter. In addition, the 4TMC instrument was tested on its effectiveness in providing reliable and valid information regarding students' conceptual understanding.

The result revealed that the integration of the 4TMC test and Partial-Credit Model is effective and valid to be treated as the diagnostic instrument to measure students' learning progress. Moreover, it is discovered that students in group A, B, C, D, and E, particularly those with low ability, are hampered in developing an epistemological explanation of the concept. This blames the students' lack of certainty in their answer and reason; thus, assumed as having lack of knowledge or misconception. The low-ability students' curve shape of LOK and AM is consistent in the competence interval of less than 0.1 logit. On the other hand, the students' ability gets lower as the conceptual understanding level increases. Such finding indicates that the learning process and education experience provide a limited contribution for the students in developing a systematic way of thinking regarding the concept of change of state of matter. In spite of that, it is considered essential to conduct a further analysis that focuses on the modification of conceptual understanding category and analysis variation that is able to define the characteristics of students' alternative conception. The development procedures, as explained in the methodology, has resulted in 4TMC instrument; however, instrument development is seen as an essential continuous process.

Recommendations

The Based on the results of the study, there are several recommendations for researchers and teachers. For researchers, the findings of this research can be followed up to examine more in how students build their understanding gradually in explaining the concept of particles in substance form changes. The study can be conducted by developing tests that aim to evaluate and diagnose the process of student knowledge formation and development while being able to identify at the level of education where the confusion of understanding occurs. The evaluation becomes more objective, not only reviewed from the student's point of ability but can be reviewed from the teacher's ability. The model of *PCM* multi-faced item response pattern approach becomes one of the important parts recommended for such objectives. In this way, students' ability to develop epistemological knowledge, and their ability to significantly actualize the knowledge gained can be measured well.

On the other hand, for teachers, the results of this study along with the stages of analysis approach used can be a reference in evaluating the progress of learners' learning, as well as determining alternative thinking frameworks of students in explaining the concept of substance change. The information serves as strategic feedback in formulating instructional strategies and preparing remedial learning, especially for students who have difficulty in developing epistemological explanations of substance changes.

Limitations

The limitations of the research are primarily related to the misrepresentation of student reasoning, which may arise in its efforts to connect phenomena and concepts measured in each item. In this context, the student may not excel to explain, because of his incapableness in using his heuristic reasoning. This instrument is not equipped with items that evaluate the heuristic abilities of the student in question. However, researchers decided to record this incompetence as a misconception or vague knowledge. For further research, it is recommended that the instrument be equipped with items that measure students' emotional and heuristic reasoning according to the conceptual framework to be evaluated.

Acknowledgments

The researchers would like to express their gratitude towards the Directorate of Research and Community Service, Ministry of Research and Technology of Republic of Indonesia, for the financial support through the University Basic Research Excellence Grant Program in the Research and Community Service Office of Universitas Negeri Gorontalo, 2020.

References

Aktan, D. C. (2013). Investigation of students' intermediate conceptual understanding levels: The case of direct current electricity concepts. *European Journal of Physics*, *34*(1), 33–43. https://doi.org/10.1088/0143-0807/34/1/33

Arslan, H. O., Cigdemoglu, C., & Moseley, C. (2012). A three-tier diagnostic test to assess pre-service teachers' misconceptions about global warming, greenhouse effect, ozone layer depletion, and acid rain. *International Journal of Science Education*, *34*(11), 1667–1686. https://doi.org/10.1080/09500693.2012.680618

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). Routledge.

Caleon, I. S., & Subramaniam, R. (2010). Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education*, *40*(3), 313–337. https://doi.org/10.1007/s11165-009-9122-4

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The development of two tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, *8*(3), 293–307

Chi, S., Wang, Z., Luo, M., Yang, Y., & Huang, M. (2018). Student progression on chemical symbol representation abilities at different grade levels (Grades 10–12) across gender. *Chemistry Education Research and Practice*, *19*(4), 1055–1064. https://doi.org/10.1039/c8rp00010g

Claesgens, J., Scalise, K., Wilson, M., & Stacy, A. (2009). Mapping student understanding in chemistry: The perspectives of chemists. *Science Education*, *93*(1), 56–85. https://doi.org/10.1002/sce.20292.

Djiks, M. A., Brummer, L., & Kostons, D. (2018). The anonymous reviewer: the relationship between perceived expertise and the perceptions of peer feedback in higher education. *Assessment & Evaluation in Higher Education*, *43*(8), 1258-1271. https://doi.org/10.1080/02602938.2018.1447645

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching*, *46*(6), 606–609. https://doi.org/10.1002/tea.20316

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, *47*(2), 123–182. https://doi.org/10.1080/03057267.2011.604476

Emden, M., Weber, K., & Sumfleth, E. (2018). Evaluating a learning progression on "Transformation of Matter" on the lower secondary level. *Chemistry Education Research and Practice*, *19*(4), 1096–1116. https://doi.org/10.1039/c8rp00137e

Goulas, S., & Megalokonomou, R. (2021). Knowing who you actually are: The effect of feedback on short-and longer-term outcomes. *Journal of Economic Behavior & Organization*, *183*, 589-615. https://doi.org/10.1016/j.jebo.2021.01.013

Guo, M., & Leung, F. K. S. (2021). Achievement goal orientations, learning strategies, and mathematics achievement: A comparison of Chinese Miao and Han students. *Psychology in the Schools. 58*(1), 107-123. .

https://doi.org10.1002/pits.22424

Haarala-Muhonen, A., Ruohoniemi, M., Parpala, A., Komulainen, E., & Lindblom-Ylänne, S. (2016). How do the different study profiles of first-year students predict their study success, study progress and the completion of degrees? *Higher Education*, *74*(6), 949–962. https://doi.org/ 10.1007/s10734-016-0087-8

Habiddin, & Page, E. M. (2019). Development and validation of a four-tier diagnostic instrument for chemical kinetics (FTDICK). *Indonesian Journal of Chemistry*, *19*(3), 720–736. https://doi.org/10.22146/ijc.39218

Hadenfeldt, J. C., Bernholt, S., Liu, X., Neumann, K., & Parchmann, I. (2013). Using ordered multiple-choice items to assess students' understanding of the structure and composition of matter. *Journal of Chemical Education*, *90*(12), 1602–1608. https://doi.org/10.1021/ed3006192

Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education*, *34*(5), 294–299. https://doi.org/10.1088/0031-9120/34/5/304

Herrmann-Abell, C. F., & Deboer, G. E. (2016). Using rasch modeling and option probability curves to diagnose students' misconceptions. *American Educational Research Association*, *8*(12), 1–12

Hoe, K. Y., & Subramaniam, R. (2016). On the prevalence of alternative conceptions on acid-base chemistry among secondary students: Insights from cognitive and confidence measures. *Chemistry Education Research and Practice*, *17*(2), 263–282. https://doi.org/10.1039/c5rp00146c

Jin, H., Mikeska, J. N., Hokayem, H., & Mavronikolas, E. (2019). Toward coherence in curriculum, instruction, and assessment: A review of learning progression literature. *Science Education*, *103*(5), 1206–1234. https://doi.org/10.1002/sce.21525

Karagiannopoulou, E., Milienos, F. S., & Rentzios, C. (2020). Grouping learning approaches and emotional factors to predict students' academic progress*. International Journal of School & Educational Psychology, 9*(1), 1–18*.* https://doi.org.10.1080/2168363.2020.183241

Klassen, S. (2006). Contextual assessment in science education: Background, issues, and policy. *Science Education*, *90*(5), 820–851. https://doi.org/10.1002/sce.20150

Latifi, S., Noroozi, O., & Talaee, E. (2021). Peer feedback or peer feedforward? Enhancing students' argumentative peer learning processes and outcomes. *British Journal of Educational Technology*, *52*(2), 768-784. https://doi.org/10.1111/bjet.13054

Lee, K., & Keller, J. M. (2021). Use of the ARCS model in education: A literature review. *Computers & Education*, *122*(1), 54-62. https://doi.org/10.1016/j.compedu.2018.03.019

Lin, P. Y., Chai, C. S., Jong, M. S. Y., Dai, Y., Guo, Y., & Qin, J. (2021). Modeling the structural relationship among primary students' motivation to learn artificial intelligence. *Computers and Education: Artificial Intelligence*, *2*(1), 1-7

Laliyo, Botutihe, & Panigoro. (2019). The development of two-tier instrument based on distractor to assess conceptual understanding level and student misconceptions in explaining redox reactions*. International Journal of Learning, Teaching and Educational Research*, *18*(9), 216–237. https://doi.org/10.26803/ijlter.18.9.12

Linacre, J. M. (2012). *A user's guide to W I N S T E P S ® M I N I S T E P Rasch-model computer program: Program manual 3.75.0*. winsteps.com.

Linacre, J. M. (2020). *A User's Guide to W I N S T E P S ® M I N I S T E P Rasch-Model Computer Programs Program Manual 4.5.1*. winsteps.com.

Ling Lee, W., Chinna, K., & Sumintono, B. (2020). Psychometrics assessment of HeartQoL questionnaire: A Rasch analysis. *European Journal of Preventive Cardiology*. Advance online publication. https://doi.org/10.1177/2047487320902322

Liu, X. (2012). Developing measurement instruments for science education research. In B. Fraser, K. G. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 651–665). Springer Netherlands

Lu, S., & Bi, H. (2016). Development of a measurement instrument to assess students' electrolyte conceptual understanding. *Chemistry Education Research and Practice*, *17*(4), 1030–1040. https://doi.org/10.1039/c6rp00137h

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, *54*(8), 1024–1048. https://doi.org/10.1002/tea.21397

Neumann, K., Viering, T., Boone, W. J., & Fischer, H. E. (2013). Towards a learning progression of energy. *Journal of Research in Science Teaching*, *50*(2), 162–188. https://doi.org/10.1002/tea.21061

Park, M., Liu, X., & Waight, N. (2017). Development of the connected chemistry as formative assessment pedagogy for high school chemistry teaching. *Journal of Chemical Education*, *94*(3), 273–281. https://doi.org/10.1021/acs.jchemed.6b00299

Peterson, R. F., Treagust, D. F., & Garnett, P. (1989). Development and application of a diagnostic instrument to evaluate grade-11 and -12 students' concepts of covalent bonding and structure following a course of instruction. *Journal of Research in Science Teaching*, *26*(4), 301–314. https://doi.org/10.1002/tea.3660260404

Rogat, A. (2011). *Developing learning progressions in support of the new science standards: A RAPID workshop series.* CPRE Research Reports. http://repository.upenn.edu/cpre_researchreports/66

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research & Perspective*, *4*(1–2), 1–98. https://doi.org/10.1080/15366367.2006.9678570

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial* [Application of Rasch model in social science research]. Trim Komunikata.

Sutiani, A., Situmorang, M., & Silalahi, A. (2021). Implementation of an Inquiry Learning Model with Science Literacy to Improve Student Critical Thinking Skills. *International Journal of Instruction, 14*(2), 117-138.

Sreenivasulu, B., & Subramaniam, R. (2013). University students' understanding of chemical thermodynamics. *International Journal of Science Education*, *35*(4), 601-635.

Testa, I., Capasso, G., Colantonio, A., Galano, S., Marzoli, I., Scotti di Uccio, U., Trani, F., & Zappia, A. (2019). Development and validation of a university students' progression in learning quantum mechanics through exploratory factor analysis and Rasch analysis. *International Journal of Science Education*, *41*(3), 388–417. https://doi.org/10.1080/09500693.2018.1556414

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education*, *10*(2), 159–169. https://doi.org/10.1080/0950069880100204

Tyson, L., Treagust, D. F., & Bucat, R. B. (1999). The complexity of teaching and learning chemical equilibrium. *Journal of Chemical Education*, *76*(2–4), 554–558. https://doi.org/10.1021/ed077p1560.1

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9781410611697

Wilson, M. (2008). Cognitive diagnosis using item response models. *Journal of Psychology/ Zeitschrift Für Psychologie* , *216*(2), 74–88. https://doi.org/10.1027/0044-3409.216.2.74

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730. https://doi.org/10.1002/tea.20318

Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progression in science* (pp. 317–344). Sense Publishers. https://doi.org/10.1007/978-94-6091-824-7