Journal of Turkish Science Education, 2021, *18*(3), 351-370.
DOI no: 10.36681/tused.2021.78

Journal of Turkish Science Education
http://www.tused.org
© ISSN: 1304-6020

Development of A Diagnostic Assessment Test to Evaluate Science Misconceptions in Terms of School Grades: A Rasch Measurement Approach

Soeharto Soeharto

University of Szeged, Doctoral School of Education, 32–34 Petőfi Sandor Street, H-6722 Szeged, Hungary, soeharto.soeharto@edu.u-szeged.hu, ORDIC ID: 0000-0003-4332-7401

ABSTRACT

This study aims to evaluate the psychometric properties of the developed diagnostic assessment test and to identify student misconceptions in science in terms of school grades. 153 students were gathered by using random sample from 10th to 12th grade in senior high schools. The 32 items of the two-tier multiple-choice diagnostic test were administered to assess student misconceptions in science using the online system (eDia) and paper-based test. The results confirmed the validity and reliability of the developed test based on Rasch measurement. Student misconceptions in science were found statistically significant among school grades, [F(2, 152) = 10.93, p <.01]. The 12th-grade students have higher misconceptions than the students at 10th- and 11th-grade. No statistically significant difference was found between boys and girls for all grade level (p> 0.05). The Stepwise multiple regression confirmed that the grades are the predictor of student misconception in science, [F(152) = 10.208, p <0.01], explaining 25,2% variances of student misconception in science. This study gave preliminary evidence that the developed test well measured student misconceptions and evaluated students' misconception in science concepts.

ARTICLE INFORMATION
Received:
04.12.2020
Accepted:
21.05.2021

KEYWORDS: Rasch measurement, diagnostic test, student misconceptions.

Introduction

Students construct their knowledge on prior learning which usually occurs at the school or other learning environments. Before participating in a learning activity at school, students have their knowledge, skills, and experience that form their initial concepts in science learning. These initial concepts may be contrary to scientific concepts. Even this situation still occurs after the science learning activity is carried out (Eshach et al., 2018; Köse, 2004; Stefanidou et al., 2019). Concepts that are opposite or not following scientific concepts are called misconceptions (Martin, 2005). Allen (2014) stated that misconception is the individual's knowledge based on formal and informal experiences which are unrelated to scientific knowledge. Besides, with the rapid development of science and technology, the increasing amount of knowledge causes changes in the meaning of science concepts (Arslan et al., 2012; Kiray et al., 2015). This condition proves that conceptual learning is essential in science education (Soeharto et al., 2019). Wrong or incomplete knowledge derived from student experience, misinformation in teacher learning, and misunderstandings in examining information in textbooks affects students' concepts (Hakim et al., 2016; Kirbulut & Geban, 2014; Zlatkin-Troitschanskaia et al., 2015).

Indonesia placed to the lowest rank among 41 countries in 2018 PISA report in terms of student science performances (OECD, 2020), which may indicate the most students in Indonesia suffering to comprehending scientific conceptions in the learning process. Many studies (e.g., Butler et al., 2015;

Erman, 2017; Galvin & Mooney, 2015; Kirbulut & Geban, 2014; Köse, 2004; Laliyo et al., 2019; Peşman & Eryılmaz, 2010; Soeharto et al., 2019; Soeharto, 2016) had confirmed that scientific misconceptions was directly related to student academic achievement and affecting student learning activity in science disciplines. Therefore, it can be assumed that if students suffer to master particular science concept which cause students' low science performance in science, students will face problems in understanding related scientific concepts in the learning process or in the future.

In the literature, there have been great effort to identify misconceptions that are specific to certain context. Wandersee et al. (1994) analyzed 103 studies related to misconceptions, Gurel et al. (2015) found 273 articles about misconceptions, and Soeharto et al (Soeharto et al., 2019) also found 111 articles from 2015 to 2019 which were focused on student misconceptions in science. There was three articles (Fajarini et al., 2018; Fariyani et al., 2017; Ratnasari & Suparmi, 2017) about identifying student misconception in Indonesia that address the lack of research concerns in the Indonesian science education research field. However, these recent Indonesian articles merely focused on identification of student misconception in one particular science concepts such as global warming, optics, and heat, and there is no developing instrument for science concepts distributing student misunderstanding in learning science. In this study, sixteen concepts are selected from science subjects. Soeharto et al (Soeharto et al., 2019) found that the multiple-tier test (33.06%) is the most diagnostic tool used to identify science misconceptions in the development trend of using diagnostic tools to identify misconceptions. Therefore, it is decided to develop a two-tier multiple-choice test which is assisted with the Rasch measurement model and to identify and evaluate the development of student misconception with respect to school grade and gender. Rasch measurement model is performed since Rasch measurement can convert research instrument which have interval scale such as Physics measurement tools. Rasch measurement also can tackle with weakness of CTT analysis from previous studies (e.g., Galvin & Mooney, 2015; Laliyo et al., 2019; Taslidere, 2016)

Theoretical Background

Student Misconceptions and The Importance of Research for Science Education

Student misconceptions has been a problem in science education area. Driver and Easley (1978) had pointed out that there are various kind of conceptual understanding among young people related to science concepts and one of the well-known is "student misconceptions". Student misconceptions are grouped into several types: the non-scientific belief, the conceptual misconception, the preconceived notion, the factual misconception, and the vernacular misconception (Keeley, 2012; Leaper et al., 2012; Morais, 2013; Murdoch, 2018). The non-scientific belief is the student knowledge obtained through non-scientific sources such as environment and experience (Leaper et al., 2012). For example, based on their daily experiences, students believe that large and heavy objects will always sink into water if they are put into water. The conceptual misconception is confusing and incorrect student knowledge obtained when students construct their knowledge based on the scientific concept in learning process (Morais, 2013). For example, students have difficulty understanding the concept of collisions because they cannot link it to daily life. The preconceived notion is a popular conception of students obtained from personal life experiences (Murdoch, 2018). For example, students believe that an object can be seen when there is light or because light from the eye leads to the object being seen. Preconceived notions usually occur because students have not yet fully learned the concept of light and how the eye works. The factual misconception is a misconception that is experienced from an early age and is maintained until adulthood. For instance, students believe that they will be struck by lightning if they are outside the house. The vernacular misconception is a misconception that occurs due to the use of words that appear in everyday life but have different scientific meanings. For example, students have difficulty in understanding the concept of mass and weight because they think that mass is equal to weight (Keeley, 2012; Samsudin et al., 2021). This study is focused on investigating conceptual misconceptions in science subjects.

In the literature, there are many studies related to student misconceptions in learning science since the characteristic of misconceptions in science are resistant to change, persistent, and rooted in some science concepts (Boone et al., 2013; Greiff et al., 2018; Morrison et al., 2019; Topalsan & Bayram, 2019). Besides, if students experience misconceptions in learning science, students will find it difficult to learn science at a higher level. Student misconceptions in science can lead students to get low academic performance scores for science education subjects such as physics, biology, and chemistry. This present study identifies science concepts which hold various misconceptions because it can help students and teachers to better understand the subject matter related to science concepts. Identifying misconceptions through using the two-tier multiple-choice test by using Rasch measurement may be essential and become an initial study in the science education area. In this study, for measuring the student misconception correctly, all the Rasch measurement steps will be performed, including checking the item bias based on the background information obtained from participants.

The Development of The Two-Tier Diagnostic Instrument to Assess Misconception in Science

In recent years from 2015 to 2019, multi-tier diagnostic tests are a popular assessment tool which are developed to identify student misconceptions in various research areas (Soeharto et al., 2019). The two-tier test is the first example in the development of a multi-tier test to diagnose student misconceptions. The two-tier multiple-choice test consists of first-tier and second-tier. The first tier assesses student conceptions, and the second tier assesses student reasonings without confident levels (Adadan et al., 2012; Korkmaz et al., 2018). In this study, the first-tier of an item was constructed based on student common misconceptions in science. The first-tier will evaluate student content knowledge. The second-tier was constructed based on possible student reasoning related to scientific conception and possible alterative conceptions. The student answer is scored if the student can answer the content and reason correctly. Two-tier tests were developed as a diagnostic instrument since student conceptions and reasons are linked to understanding scientific misconceptions. Researchers can even find student answers with two-tier tests that have not been thought of before with blank option choice (Tsui & Treagust, 2010). Students are more accessible in responding to the question, and this test is used practically by researchers in various ways, including large-scale use, ease of scoring, and explanations regarding student reasoning (Adadan et al., 2012).

On the other hand, there are criticisms regarding the use of two-tier tests in identifying misconceptions. In his research on misconceptions of geometrical optics in physics subject, Gurel et al. (2015) identified that two-tier tests might produce invalid misconceptions due to a lack of level of uncertainty where the researcher cannot ensure the correctness of student answer to guess, misconception, or concept. Although there are weaknesses in measuring student misconceptions since they cannot confirm students' answers with the confidence tier as in the three-tier or four-tier tests, the weaknesses in the form of guess answers, confident level issues, and missing data on the two-tier can be overcome by running the Rasch measurement model.

Rasch Measurement and Scoring Procedures

Rasch measurement is a measurement model developed by George Rasch, a Danish mathematician. Rasch measurement is based on interactions between item-person interaction and probability estimates. The interaction between items and persons can be described based on mathematical equations. Persons who have high abilities should correctly answer items with easier difficulty levels (Andrich, 2018). The probability in the measurement is governed by the difficulty of the item and person simultaneously. In other words, the probability is closely related to differences between item difficulty and individual abilities ((Boone et al., 2016). Person ability and item difficulty in Rasch measurement is set based on an interval scale, called as logit, and item and person parameters are entirely independent (Bond & Fox, 2007; Sumintono & Widhiarso, 2014). It means that the students' ability in the measurement remains the same regardless of the item's difficulty level, and the item difficulty level remains invariant regardless of the student's ability or test takers. In this study, the Rasch dichotomous model was used to analyze the two-tier multiple-choice diagnostic test, where 1 represents the correct concept, and 0 represents the misconception. The two-tier multiple-choice diagnostic test result was recorded and combined by the following procedure: (a) in which correct responses for both items scored as 1, (b) incorrect response for any tier scored as 0. Unidimensionality and local independence are the two assumptions underlying Rasch measurement and the development. The instrument must meet these two assumptions to achieve a suitable model in terms of data fit criteria. Unidimensionality is the central assumption in the single Rasch model, which shows that the items in used instruments measure the same aspect. Local independence shows the correlation between item responses, which is the latent trait of the measured students. The non-statistically significant correlation between the items used to estimate latent traits should be achieved when latent traits are controlled (Liu, 2007). The presumption of local independence prevents item redundancy and individual reliability inflation (Boone et al., 2016).

Rasch analysis was employed in this study to tackle some limitations of Classical Test Theory (CTT). The CTT has four limitations in describing a measurement model: (a) the measurement is constructed by using the result of ordinal data rather than interval scale (logit); (b) item and person in measurement are dependent; (c) measurement properties in the instrument in terms of reliability and validity are highly dependent on the sample; (d) the data is centered on group-centered statistics but it is not suitable for explaining the measurement of individual respondents ((Barbic & Cano, 2016).

Research Questions

This present study investigates and evaluates the psychometric properties of the developed instrument, which examines student misconceptions in science learning and identifies background factors affecting student misconceptions in the learning context. Thus, the developed instrument in the form of the two-tier diagnostic test was administered to answer six research questions:

(1) Does the developed instrument achieve reliability and validity based on Rasch measurement?
(2) How do items and persons interact in the developed instrument?
(3) How do the student misconceptions develop in science learning?
(4) Is there an instrument bias based on gender according to differential item functioning (DIF)?
(5) How do the student misconceptions develop in terms of school grades?
(6) What are the factors predicting student misconceptions in science?

Methods

Research Design

The quantitative approach was employed, where a two-tier test multiple-choice test was administered to understand student misconceptions in science, especially in physics, biology, and chemistry, and Rasch modelling was used to analyze psychometric properties.

Participants

The participants in this preliminary study were 153 students at public senior high schools and private senior high school schools in Pontianak, part of West Kalimantan province, Indonesia. The samples were recruited by using stratified random sampling according to student grades. In this study, five classes from 5 different schools were randomly selected for the analysis. Data were collected from 123 students by using the paper-based test and 30 students by using the online Electronic Diagnostic Assessment System, the eDia, developed by the Center for Research on Learning and Instruction at the University of Szeged (Csapó & Molnár 2019). The eDia system can support item writing, editing, and scoring by using logfile analysis as well as administering the test, and giving feedback. The eDia was used in the various research areas in teaching and learning, including reading, science, and mathematics, that can be accessed using internet browser applications such as Google Chrome and Firefox (Csapó & Molnár 2019; Greiff et al., 2018). The demographic profile of the participants is presented in Table 1. The data collection was performed from May to June 2019. Students spent 120 minutes completing the test under the surveillance of researchers and teachers.

Table 1

The Demographic Profile of The Participants in This Study

Demographic		Frequency	Percentage (%)
Gender	Girls	68	44.4
	Boys	85	55.6
Grade	10th	57	37.3
	11th	55	35.9
	12th	41	26.8
School category	Public	109	71.2
	Private	44	28.8
Living place	City	77	50.3
	District	76	49.7

Instruments

Background Questionnaire and School Performance

The background questionnaire was adapted from the Indonesian version's PISA 2015 SES questions (OECD, 2016). The questionnaire is embedded in the developed multi-tier diagnostic test body in the online and paper-based format. The background questionnaire in this study consists of information such as gender, parents' level of education, parents' jobs, and student performance in the science subjects of the previous semester. The background questionnaires were functioned to depict demographic profile and to evaluate predictors that affect student misconceptions in science using stepwise regression analysis.

The Two-Tier Multiple-Choice Diagnostic Test

To identify students' misconceptions in science, 32 items were developed and divided into three science subjects as physics, biology, and chemistry. Sixteen selected concepts among the misconceptions in science were shown in Table 2. Concepts and item numbers in the developed two-tier multiple-choice diagnostic test. In identifying common misconceptions in science, Literature review studies and misconceptions in science handbooks were investigated (AAAS, 2019; Allen, 2014; Csapó 1998; Soeharto et al., 2019). Then the selected concepts had been adjusted according to Indonesian education curriculum, the Curriculum 2013, especially on the senior high school level. All items in the test were translated using the back-forward translation from English to Indonesian and then from Indonesian to English by researchers. The sample task in Indonesian and English versions can be seen in Table 3.

Table 2

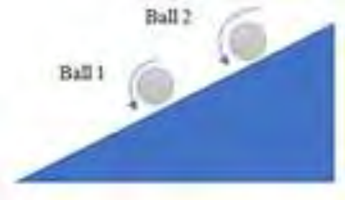
Concepts and Item Number in The Developed Two-Tier Multiple-Choice Diagnostic Test

Subject	Concepts	Item number
Physics	Kinetic energy, thermal energy, atoms and molecules, forces, light	1, 2, 3, 4, 5, 6, 7, 8, 9,10, 11, 12
Biology	Cells, breathing, microbes and disease, human body systems, feeding relationships	13, 14, 15, 16, 17, 18, 19, 20, 21, 22
Chemistry	Substances and chemical reactions, chemical compounds, chemicals equilibrium, hydrocarbons, redox reaction	23, 24, 25, 26, 27, 28, 29, 30, 31, 32

To cope with the diagnostic instruments for assessing student misconceptions in this research, it is decided to construct a diagnostic test in a two-tier multiple-choice test. The first tier will represent question-related student conception in science and the second tier represent student reasonings about their conceptions in science task. A blank option was also provided to give students a chance if their reasoning option is not available. A correct answer was scored to 1 point, and an incorrect answer was scored to 0 points for all the items. Students get 1 point if they answer the task correctly in the first and second tier.

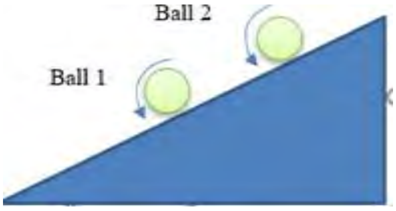
Table 3

Sample Task in Indonesian and English Version

Indonesian Version	1. Dua bola dengan yang identik bergulir di bidang miring. Bola 2 lebih cepat dari Bola 1.  Bola mana yang memiliki energi kinetik lebih besar? a) Bola 2 memiliki energi kinetik yang lebih besar. b) Bola 1 memiliki energi kinetik yang lebih besar. c) Bola 1 dan Bola 2 memiliki jumlah energi kinetik yang sama. d) Kedua bola tidak memiliki energi kinetik. Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya? a) Energi kinetik bola tidak tergantung pada kecepatan b) Energi kinetik bola tergantung pada kecepatan, massa dan posisi c) Energi kinetik bola tergantung pada kecepatan d) Energi kinetik bola tergantung pada kecepatan dan posisi karena kedua bola identik e) …………………………………………………………………………………………

English Version

1. Two balls with identical characteristics are rolling on the sloping board. Ball 2 is faster than Ball 1.



Which ball has more kinetic energy?

a) Ball 2 has higher kinetic energy than Ball 1.
b) Ball 1 has higher kinetic energy than Ball 2.
c) Ball 1 and Ball 2 have the same amount of kinetic energy.
d) Neither ball has any kinetic energy.

Which one of the following is the reason for your answer to the previous question?

a) The kinetic energy of a ball does not depend on the speed.
b) The kinetic energy of a ball depends on speed, mass, and position.
c) The kinetic energy of a ball depends on the speed.
d) The kinetic energy of a ball depends on speed and position because both balls are identic.
e) ....

Procedures, Data Analysis, and Rasch Measurement

Before conducting data collection in schools, researchers asked permission to administer the tests at schools and granted ethical research approval. The paper-based tests were conducted in student classrooms with the guidance and supervision of researchers and teachers. Online tests were conducted in each school computer laboratory using the eDia system. Statistical Package for the Social Sciences (SPSS) version 25 (IBM SPSS, 2017) and the Winsteps version 4.7.0 software (Linacre, 2020) were employed in this study. Winsteps was used to perform data analysis by using Rasch modelling. Winsteps performed Rasch analysis from simple rectangular dataset. Winsteps can be utilized to analyze multiple-choice, dichotomous, and multiple rating-scale and partial credit items. This software can be downloaded in trial and full version in Winsteps website (www.winsteps.com). The SPSS version 25 was used to analyze using statistical methods such as descriptive statistics, regressions, and ANOVA. All samples in the data set were investigated because this preliminary study wanted to explore item and person interaction.

To analyze an instrument's psychometric quality, the most common method is employing software or statistic calculations based on Classical Test Theory (CTT) principles. However, CTT has several limitations, such as the sample-dependent and biased derived scores against central scores (Bradley et al., 2015). In CTT, missing data presents a problem in calculating the data as a whole. Reliability measures are described using Cronbach's alpha, and measurement evidence is based on the correlation between items and other measures, which may not be reliable and valid. It is challenging to assess individual items' characteristics to determine the effectiveness of items in the population and their contribution to measure the overall latent construct. There are many measurement problems with surveys, questionnaires, and rating scales, which concluded that CTT used measurement could produce various responses and analysis biases (Bradley et al., 2015; Zlatkin-Troitschanskaia et al., 2015). Therefore, the Rasch measurement was employed to tackle measurement issues in CTT (Barbic & Cano, 2016). Rasch analysis can explain the difficulty level of an item accurately and precisely, detect the suitability and interaction of items and persons (item-person maps), identify outliers (person misfit), detect item bias (differential item functioning (DIF)), which is useful for describing and identifying students conceptions in science in this study (Boone et al., 2016; Sumintono & Widhiarso, 2014).

Findings

Scalling, Reliability, and Validity of The Developed Instrument

The psychometric properties of the developed instrument based on Rasch measurement model was analyzed in this study. Winsteps run the analysis based on the Joint Maximum Likelihood Estimation (JMLE) equations. In this formulation, the raw data were converted to interval data (logit) (Linacre, 1998, 2020). The logit scale can express person ability and item difficulty ranging from positive infinity to negative infinity. The 32 items of the misconception test and 153 participants were processed with a two-facet item and person model using the Rasch measurement model with the Winsteps software. The mean measure (logit) of the items is 0.00, and the standard deviation (SD) is relatively high (1.84), which means that the variation or dispersion of item measurement in terms of item difficulty was wide across the logit scale. The mean measure was 0.75 logit for students, indicating all respondents tended to be strongly involved in misconception in science, but the person SD was 0.87, almost achieving 1, showing person variation is ideal for data analysis. The mean OUTFIT mean-square and the average outfit z-standardized (ZSTD) was acceptable (ranging from -2 to +2), and outfit mean-square (MNSQ) statistics are 0.96, which is near their expected value of 1 for item and student, and the chi-squared score showing the data achieved the normal distribution criteria and Rasch model fits globally (Boone et al., 2013; Engelhard Jr, 2013; Linacre, 2020). The item separation was 5.81, indicating various levels of item difficulties, and the person separation was 1.91 showing that data consists of 2 levels, high and low performance. The reliability of items and person were excellent (Fisher, 2007; Taber, 2018). The summary statistics of item and person can be seen in Table 4.

Table 4

The Summary of the Statistic Based on Persons and Items

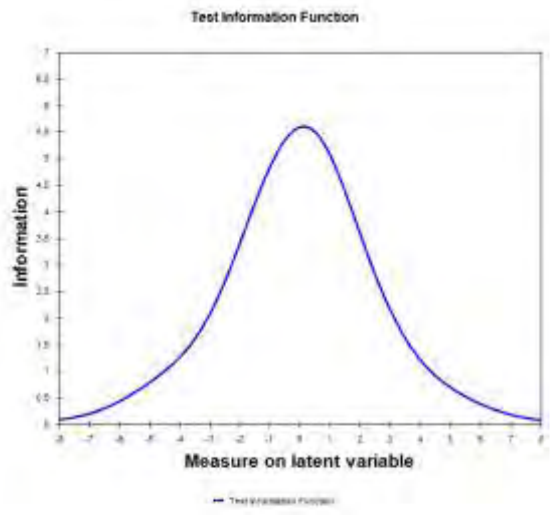
	Persons	Item
N	153	32
Measure	0.75	0
Mean	19.7	94
SD	0.87	1.84
SE	0.08	0.33
Mean Outfit MNSQ	0.96	0.96
Mean Outfit ZSTD	0.12	-0.09
Separation	1.91	5.81
Reliability	0.76	0.97
Cronbach's Alpha	0.8	
Chi-squared (χ2)	4443.85 (df= 4431)	
Probability	0.4429*	

*Note: *Normally distributed*

The reliability is calculated based on item internal consistency by using Cronbach's alpha value for all items and based on the item and person reliability parameter in Rasch measurement. Cronbach alpha for the whole item was 0.8 which indicate high internal consistency reliability (Taber, 2018). The reliability parameter in Rasch measurement was 0.76 and 0.97 for person and item statistics representing good reliability (more than 0.67) (Fisher, 2007). All items in the developed instrument are not deleted and retained in the developed instrument. To achieve validity, the unidimensionality was assessed as well as local independence of the instrument. The unidimensionality shows that the instrument measures the same dimension, which is student misconception in science. The instrument can achieve unidimensionality if the value of the raw variance explained by the measure is more than 30% (Chou & Wang, 2010; Linacre, 1998). The analysis result confirmed that the developed instrument passed the minimum threshold for the variance explained by measure was 37.4% with 12.18 eigenvalue. The local independence is achieved when the raw residual correlation among items is lower than 0.3 (Christensen et al., 2017; Hagell, 2014). The instrument's local independence in this study was below 0.3, which indicated that no items have local dependence. The test information function in Figure 1 have given additional proof of test quality to measure student misconception in science with large range of difficulty level from -8 to +8. It means that the develop test can cover items from the easiest difficulty to the most difficult based on person ability. Therefore, it can be concluded that the developed two-tier multiple-choice test used in this study is valid and reliable.

Figure 1

Test Information Function for The Two-Tier Multiple-Choice Test

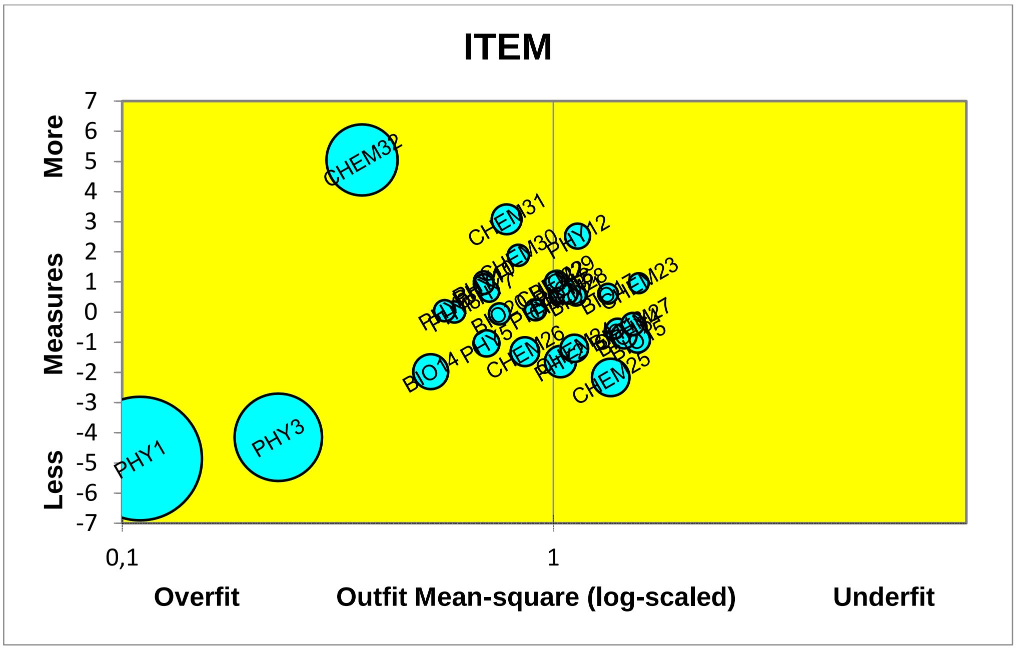


Item Fit

Item fit analysis was carried out to see whether the developed two-tier multiple-choice diagnostic test could measure student misconceptions at the senior high school level. The ideal MNSQ outfit and infit value are 1 based on the Rasch measurement model, but the acceptable values ranging from (0.5-1.5) below 1.6 are still acceptable, and besides that it can also be seen based on the point measure correlation range from 0.4 to 0.85 as an additional indicator (Andrich, 2018; Bond & Fox, 2007). The results of the analysis showed that the mean of infit and outfit MNSQ is 0.99 (SD = 0.18) and 0.9 (SD = 0.39), respectively. However, there are 3 misfit items based on the MNSQ outfit value, namely items PHY1 (0.11), PHY3 (0.23), and CHEM32 (0.36). These three items must be removed or corrected before administering the test in larger sample. The item measure is calculated in logit units ranging from the least difficult (-4.86 logit) to the most difficult (5.05 logit), which means that the instrument is around 4 or 5 categories in the item difficulty level. However, since this study is the preliminary study for developing instrument, those three items are retained to item analysis and improvement for other test version in future study. The distribution of item fit order is shown in Figure 2.

Figure 2

Bubble Chart of Item Fit Order Based on Infit MNSQ



Person Ability

Person ability measure describes the student ability in answering items on the test. Person ability in this study ranging from -2.11 logit to 2.43 (M = 0.75, SD = 1). Person ability was categorized into 4 types on logit value of item (LVI) based on Sumintono & Widhiarso (2014), low misconception 16.33% (2.43 <LVI <1.75), moderate misconception 49.01% (0.75 <LVI <1.75), high misconception 14.37% (0.75 < LVI <- 0.25), and very high misconception 20.26% (-0.25 <LVI <- 2.11). Overall, 37% of students answered incorrectly, which shows that students have misconceptions on the basic concepts in science learning. Misconceptions in each subject in science were also checked based on the percentage of students' incorrect answers to see how the misconceptions were distributed based on the science subjects, physics (33.4%), biology (35.22%), and chemistry (47.97%).

Figure 3

The Wright Item-Person Map of Student Misconception In Science Subjects

```
MEASURE                                PERSON - MAP - ITEM
                                            <more>|<rare>
     5                                            +  CHEM32
                                                  |
                                                  |
                                                  |
                                                  |
                                                  |
     4                                            +
                                                  |
                                                  |T
                                                  |
                                                  |
     3                                            +  CHEM31
                                                 T|
                                                  |
                                   P P P P P P P  |  PHY12
                                                  |
     2                                 P P P P P  +
                       P P P P P P P P P P P P P  |S CHEM30
                                                 S|
                             P P P P P P P P P P  |
                         P P P P P P P P P P P P  |
                 P P P P P P P P P P P P P P P P  |
     1 P P P P P P P P P P P P P P P P P P P P P  +  BIO22  CHEM23 CHEM29 PHY10
                                                  |  BIO18  PHY11
                 P P P P P P P P P P P P P P P P M|  BIO17  BIO19  CHEM28 PHY7
                             P P P P P P P P P P  |  BIO21
                                               P  |
     0                                     P P P  +  PHY6
                                                  |M BIO20  PHY8   PHY9
                                 P P P P P P P P S|
                                       P P P P P  |  CHEM27
                               P P P P P P P P P  |  BIO13  PHY4
                                   P P P P P P P  |  BIO16
    -1                                       P P  +  BIO15
                                           P P P  |  PHY5
                                               P T|  CHEM24 CHEM26
                                                  |
                                                  |  PHY2
    -2                                         P  +S
                                                  |  BIO14
                                           P P P  |  CHEM25
                                                  |
                                                  |
                                                  |
    -3                                            +
                                                  |
                                                  |
                                                  |
                                                  |
                                                  |T
    -4                                            +
                                                  |  PHY3
                                                  |
                                                  |
                                                  |
    -5                                            +  PHY1
                                            <less>|<freq>
```

To comprehend the interaction between item and person, the item-person analysis was run by using the Wright map which illustrates the student ability on the left side and item difficulty on the right side. The Wright map is item-person maps that can compare items and people simultaneously in the context of a measurement on the one interval scale (logit), and assess the interactions between items and person, as well as check students' individual abilities. If the item is in line with the person, it means that the student has a 50% chance ($p = 0.5$) of answering correctly because the difficulty level of the item is the same as student ability. If the person is located above the item, it means that the student has the correct chance to answer the question more than 50% ($p> 0.5$). If the difficulty level of the item is higher than the student's ability, the chance of the student to answer correctly is lower than 50% ($p <0.5$)

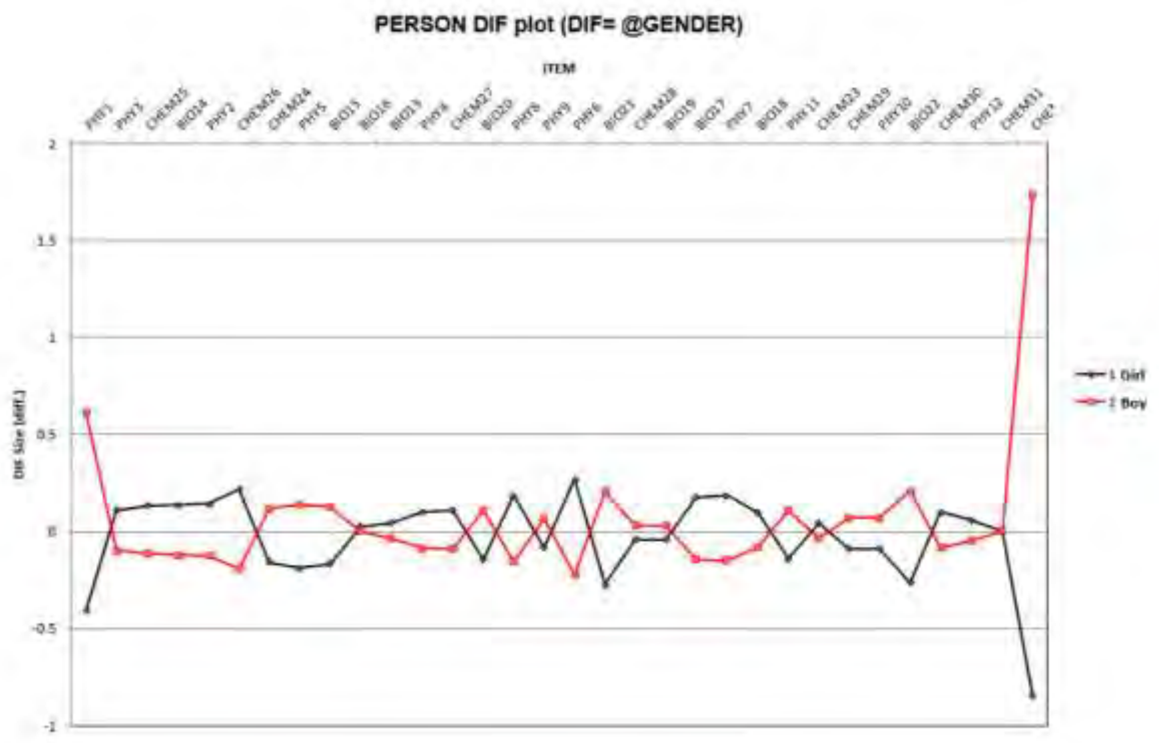
(Griffin, 2010; Linacre, 2020). In this study, the easiest item is shown at the bottom on the right of y axis (CHEM25, PHY1, and PHY3) while the most difficult item is shown at the bottom on the right of y axis (CHEM31 and CHEM32). The good items in the instrument have to cover all student abilities in the item-person map (Griffin, 2010). However, there are three misfit items, which are CHEM32 (too difficult), and PHY1 and PHY3 (too easy) having logit more than two standard deviation. In general, if the misfit items are omitted, the test still shows good performance and acceptable because the developed test can cover all scales of person abilities. Therefore, it can be concluded that the developed test is matching with the target group of in testing student misconception in science subjects. The Wright item-person map of student misconception in science subjects can be seen on Figure 3.

Item Bias Based on Differential Item Functioning (DIF)

DIF analysis was conducted to check whether there were items bias based on gender. DIF analysis suggested on participant responses based on subgroups for each item in the test of measuring student misconceptions on science learning (Adams et al., 2020; Boone et al., 2014; Rouquette et al., 2019). DIF analysis is divided into three types namely negligible, slight to moderate (| DIF | ≥ 0.43 logits), moderate to large (| DIF | ≥ 0.64 logits) (Zwick et al., 1999). DIF analysis (Figure 4) showed that the items PHY1 and CHEM32 have DIF bias in the moderate to large category. These two items were also misfit item. Items PHY1 and CHEM32 explained that these two items were more difficult for boys than girls to answer correctly.

Figure 4

DIF Based on Gender



The Differences of Student Misconceptions In Science-Based School Grade

ANOVA was conducted to determine the comparison of student misconceptions in science with respect to school grades on the test and subtest. The analysis showed that there were significant differences between school grades which confirmed student misconception test and subtest score across four cohorts with the physics subtest [F (2, 152) = 6.35, p <.01], biology subtest [F (2, 152) = 7.84, p <.01], chemistry subtest [F ((2, 152) = 5.06, p <.01], and the entire test [F (2, 152) = 10.93, p <.01]. Because the equal variances are not assumed, Dunnett T3-test was run to identify specific differences between the school grades in Table 5. Dunnett T3-test was utilized when comparing one group to other groups. Dunnett T3-test is the most powerful ANOVA post-hoc tests than others. Overall, the entire test's significant differences were found for all school grade pairs, except for the differences in all subtests (Physics, Biology, and Chemistry), which showed that the 10th-grade students had a higher mean score of misconceptions than the 11th-grade students on the subtest and the entire test. This trend showed that 10th-grade students more misunderstanding science concepts than 11th-grade students. However, the 12th-grade students suffered from the highest conceptual misconceptions than students in the 10th- and 11th-grades. This phenomenon showed misconceptions that are resistant, persistent to change, and rooted deeply in science concepts which made students more difficult to understand science learning with the increase of grade level.

Table 5

The Dunnett-T3 Multiple Comparisons of Student Misconception on School Grades

Grade	Physics		Biology		Chemistry		Test	
	Mean differences	p	Mean differences	p	Mean differences	p	Mean differences	p
10th & 11th	0.58	0.54	0.06	0.99	0.30	0.72	0.93	0.56
10th & 12th	-1.35	0.06	-1.31*	0.01	-0.82	0.07	-3.61*	0.00
11th & 12th	-1.94*	0.00	-1.38*	0.00	1.13*	0.01	-4.55*	0.00

Gender Differences Among School Grades

In general, the boxplot showed that boys and girls were identified as having equivalent mean scores of student misconception in science for each cohort shown in Figure 5. Mean scores of student misconceptions in science range from 0.28 to 0.47, where the mean score for boys in 12th-grade (0.47) explained that boys were suffering misconceptions higher than girls (0.44). However, for the whole grades, the average score among boys and girls is relative at the same level. The length of the boxplot in Figure 5 showed that the standard deviation for the 12th grade is higher than the 10th and 11th grade, showing that girls experience more varied misconceptions than boys. In table 6, a comparison was made between boys and girls-based on t-test with the maximum likelihood estimate of the students' conceptual misconception in science. No statistically significant differences were found in the test and whole grade school level (p> 0.05). This also indicates that each cohort is not different between girls and boys. Therefore, it can be concluded that the estimate of girls and boys had an equivalent value. However, unexpectedly, in the 12th-grade boys, the mean score of student misconceptions in science was slightly higher than girls, and the opposite was the variation in misconceptions where the misconceptions of female students were more varied than boys.

Table 6

The T-Test Comparing Student Misconceptions Between Girls and Boys

Grade	Girl		Boy			
	N	Mean(SD)	N	Mean(SD)	t	p
10	13	0.35(0.12)	44	0.34(0.13)	0.16	.60
11	37	0.33(0.14)	18	0.28(0.09)	1.51	.11
12	18	0.44(0.21)	23	0.47(0.18)	-0.48	.41

Figure 5

Comparison of Student Misconception among School Grades

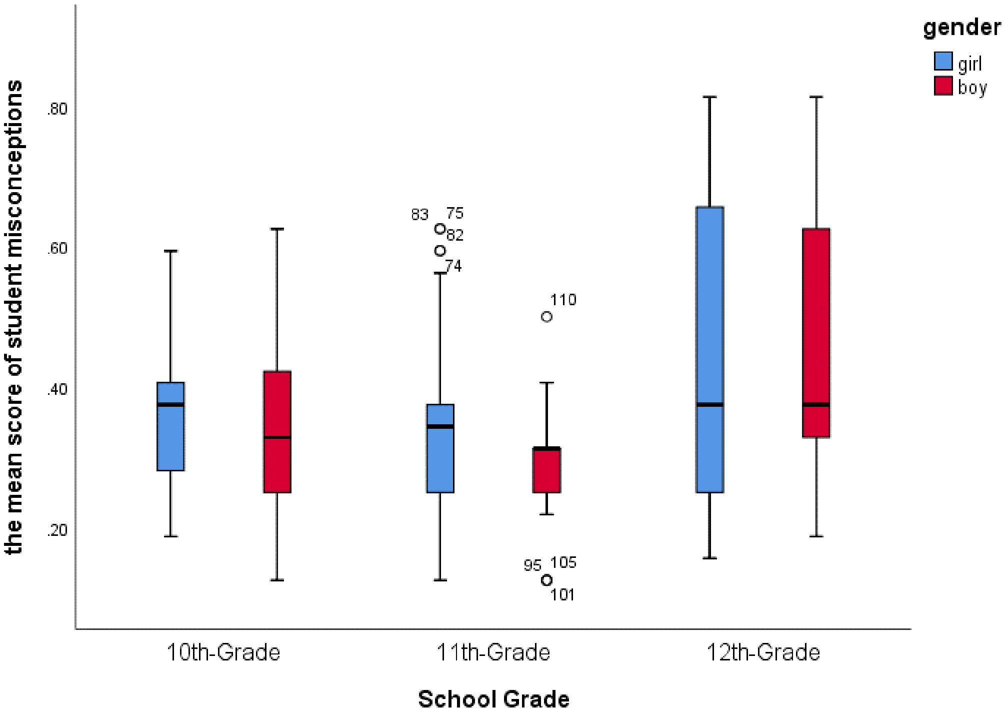


Table 7 showed the student misconceptions for all science subjects. Boys (48%) and girls (47%) suffered from high misconceptions in chemistry subject. However, overall, boys and girls had the same or equivalent percentage of misconceptions, and no significant differences were found based on the t-test conducted on all science subjects. These results were in line with the study about student misconceptions in science on gender subgroups (Taslidere, 2016; Treagust, 1988; Tsui & Treagust, 2010).

Table 7

The T-Test Comparing Misconceptions in Science Subjects Between Girls and Boys

Subject	Girl	Boy		
	Mean (SD)	Mean (SD)	t	p
Physics	0.34 (0.23)	0.32 (0.22)	0.29	0.38
Biology	0.34(0.21)	0.35(0.18)	-0.46	0.15
Chemistry	0.48(0.18)	0.47(0.17)	0.07	0.40
All subject (Science)	0.37 (0.16)	0.36 (0.16)	0.05	0.70

Predicting student misconceptions in science

To explore how other factors predict student misconceptions in science, the stepwise multiple regression was run with school category, school grade, father education, mother education, school performance as predictors. The analysis result showed that only school grade predictor could significantly explain 25.2% of the variance on student misconception mean scores [F(152) = 10.208, p <0.01]. These results indicated that school grade is an essential factor in the development of student misconceptions in learning science at senior high school.

Discussion

The preliminary result indicated that the developed two-tier multiple-choice diagnostic test is valid and reliable for identifying student misconceptions in science for 10th, 11th, and 12th grades in school contexts. With the all items, the test can identify conceptual misconceptions and cover all student ability areas even though three misfit items must be revised or deleted in further research. The item-person analysis indicated that all item can cover student ability from low to high ability although three misfit items have to be revised for further research in large sample. Nonetheless, the stabilized value for misfit item depend on the number of samples (Boone et al., 2013; Khine, 2020; Planinic et al., 2019). In the development student misconceptions, it was found that there are significant differences in student misconception between science disciplines based on ANOVA test. This findings are in line with previous studies in student and item evaluation related to energy (Park & Liu, 2019), which is one of the science concepts chosen in this study. Student misconception mean scores in science may range across school grades but remain persistent and resistant to the same concept indicating that students still suffer from misconception in science even if they have been in upper grade level (Taslidere, 2016; Tsui & Treagust, 2010; Wandersee et al., 1994). Moreover, the finding in Figure 5 showed that the students at 12th grade had higher misconceptions than the students in 10th and 11th grades. But, the 10th and 11th-grade pairs did not have substantial significant misconception score. This condition might occur based on characteristic of misconceptions that are persistent, resistant, and root deeply in science concepts ((Arslan et al., 2012; OECD, 2016, 2020; Treagust, 1988; Wandersee et al., 1994) whereby students in grade 12th actually already had misconceptions related to particular science concepts when they were in grade 10th and 11th. So that, student misunderstandings were getting worse by time in the upper grade level. The DIF analysis showed that two items are biased based on gender, PHY1, and CHEM32. However, these items are still retained to analyze the psychometric properties of the developed test.

The online and paper-based tests in this study offers several solutions to the initial stages of instrument development process. This study might be the first study that assesses student misconceptions in science based on the Rasch measurement model. Rasch measurement can solve several problems in assessing misconceptions that cannot be resolved based on CTT, for example, detecting the difficulty level of an item accurately and precisely, determining the misfit of items and persons, and identifying DIF items (Adams et al., 2020; Boone et al., 2013). Technology-based testing offers several solutions to cover an even broader competency range in development tests on different difficulty levels. This present study identifies student misconceptions in science subjects, physics (33.4%), biology (35.22%), and chemistry (47.97). In comparing school grades, regression analysis was able to explain 25.2% variance of student misconceptions in science. Stepwise regression showed that only school grade predictor could significantly explain 25.2% of the variance on student misconception mean scores [F(152) = 10.208, p <0.01].

Conclusion and Implications

To sum up, it was concluded that this study can provide comprehensive knowledge related to evaluation and development of student misconceptions in science. All the items in the developed instrument are valid and reliable and covering student ability based on item-person Wright maps even there are three misfit item and DIF issue based on gender. ANOVA test have verified that there are significant differences between science concepts across science disciplines and school grades whereby grade school predicted student misconception in science based on stepwise multiple regression. Independent sample t-test verified that no significant difference was found between boys and girls.

There are several limitations in the measurement in this study as well. Items were not developed based on all scientific concepts studied in Indonesia. The items were selected based on concepts which hold misconceptions according to the previous research (AAAS, 2019; Allen, 2014; Csapó 1998; Gurel et al., 2015; Soeharto et al., 2019). Therefore, further research is needed to find new science concepts, where students find it challenging to understand and have conceptual misconceptions. The participants also were drawn from a small population in West Kalimantan province may be a limitation in this study. It was realized that some of the results in the educational context could not be generalized.

This research is early-stage research, so it is necessary to research a larger sample to identify misconceptions in science at school contexts. However, this research is probably the first in using Rasch model analysis in developing a two-tier test by combining online and paper assessments. This study's exposure might encourage the emergence of other studies related to scientific misconceptions with Rasch measurement analysis. It is hoped that the successfully developed of the instrument will inspire other researchers to create a diagnostic assessment based on Rasch measurement. For educators and instructors, it is hoped that this report related to evaluation and development of student misconceptions in science can be an initial signal or alert to overcome student problem in understanding science concepts. If an educator realizes what is the specific concepts which are difficult to understand in learning activity, they can cope with the problem and be more concerning to design proper and correct lesson plan to make the student to understand and to improve in science performance.

Acknowledgment

The author would like to appreciate the forum discussion of Doctoral of Education at University of Szeged and the Tempus Public Foundation as a scholarship institution that supports the corresponding author's degree through Stipendium Hungaricum scholarship program.

Declaration of Competing Interest

Author declare no potential conflict of interest in this study concerning the authorship, research, and publication of this article

References

AAAS. (2019). *Project 2061 | American Association for the Advancement of Science*. https://www.aaas.org/programs/project-2061

Adadan, E., Savasci, F., & Martin, R. E. (2012). An analysis of 16–17-year-old students' understanding of solution chemistry concepts using a two-tier diagnostic instrument. *International Journal of Science Education*, *34*(4), 513–544.

Adams, D., Joo, M. T. H., Sumintono, B., & Oh, S. P. (2020). Blended Learning Engagement in Higher Education Institutions: A Differential Item Functioning Analysis of Students' Backgrounds. *Malaysian Journal of Learning and Instruction*, *17*(1), 133–158.

Allen, M. (2014). *Misconceptions in primary science*. McGraw-hill education.

Andrich, D. (2018). Advances in social measurement: A Rasch measurement theory. *Perceived Health and Adaptation in Chronic Disease*, 66–91.

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

Barbic, S. P., & Cano, S. J. (2016). The application of Rasch measurement theory to psychiatric clinical outcomes research: Commentary on… Screening for depression in primary care. *BJPsych Bulletin*, *40*(5), 243–244.

Bond, T. G., & Fox, C. M. (2007). Rasch modeling applied: rating scale design. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences (2nd Ed., Pp. 219–233). Mahwah, NJ: Lawrence Erlbaum Associates Publishers*.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer.

Boone, W. J., Townsend, J. S., & Staver, J. R. (2016). Utilizing multifaceted Rasch measurement through FACETS to evaluate science education data sets composed of judges, respondents, and rating scale items: An exemplar utilizing the elementary science teaching analysis matrix instrument. *Science Education*, *100*(2), 221–238.

Bradley, K. D., Peabody, M. R., Akers, K. S., & Knutson, N. (2015). Rating Scales in Survey Research: Using the Rasch model to illustrate the middle category measurement flaw. *Survey Practice*, *8*(2).

Butler, J., Mooney Simmie, G., & O'Grady, A. (2015). An investigation into the prevalence of ecological misconceptions in upper secondary students and implications for pre-service teacher education. *European Journal of Teacher Education*, *38*(3), 300–319. https://doi.org/10.1080/02619768.2014.943394

Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement*, *70*(5), 717–731.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194.

Csapó, B. (1998). *Iskolai tudas*. Osiris Kiadó.

Csapó, Beno, & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in Psychology*, *10*(JULY). https://doi.org/10.3389/fpsyg.2019.01522

Driver, R., & Easley, J. (1978). Pupils and Paradigms: a Review of Literature Related to Concept Development in Adolescent Science Students. *Studies in Science Education*, *5*(1), 61–84. https://doi.org/10.1080/03057267808559857

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Erman, E. (2017). Factors contributing to students' misconceptions in learning covalent bonds. *Journal of Research in Science Teaching*, *54*(4), 520–537. https://doi.org/10.1002/tea.21375

Eshach, H., Lin, T., & Tsai, C. (2018). Misconception of sound and conceptual change: A cross-sectional study on students' materialistic thinking of sound. *Journal of Research in Science Teaching*, *55*(5), 664–684.

Fajarini, F., Utari, S., & Prima, E. C. (2018). Identification of students' misconception against global warming concept. *International Conference on Mathematics and Science Education of Universitas Pendidikan Indonesia*, *3*, 199–204.

Fariyani, Q., Rusilowati, A., & Sugianto, S. (2017). Four-tier diagnostic test to identify misconceptions in geometrical optics. *Unnes Science Education Journal*, *6*(3).

Fisher, W. P. J. (2007). Rating Scale Instrument Quality Criteria. *Rasch Measurement Transactions*, *21*(1), 1095. http://www.rasch.org/rmt/rmt211m.htm

Galvin, E., & Mooney, S. G. (2015). Identification of Misconceptions in the Teaching of Biology: A Pedagogical Cycle of Recognition, Reduction and Removal. *Higher Education of Social Science*, *8*(2), 1–8. https://doi.org/10.3968/6519

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education*, *126*, 248–263.

Griffin, P. (2010). *Item response modelling: An introduction to the Rasch model.* Assessment Research Centre Faculty of Education, The University of Melbourne.

Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a

Hagell, P. (2014). Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the Rasch model: the primacy of theory over statistics. *Open Journal of Statistics*, *4*(6), 456–465.

HAKİM, A., KADAROHMAN, A., & SYAH, Y. M. (2016). Effects of the natural product mini project laboratory on the students conceptual understanding. *Journal of Turkish Science Education*, *13*(2), 27–36.

IBM SPSS. (2017). IBM SPSS Statistics for Windows, version 25. In *Armonk, NY: IBM SPSS Corp.[Google Scholar]*.

Keeley, P. (2012). Misunderstanding misconceptions. *Science Scope*, *35*(8), 12. https://www.questia.com/library/journal/1G1-294903479/misunderstanding-misconceptions

Khine, M. S. (2020). Rasch Measurement. In *Rasch Measurement*. https://doi.org/10.1007/978-981-15-1800-3

Kiray, S. A., Aktan, F., Kaynar, H., Kilinc, S., & Gorkemli, T. (2015). A descriptive study of pre-service science teachers' misconceptions about sinking-floating. *Asia-Pacific Forum on Science Learning & Teaching*, *16*(2).

Kirbulut, Z. D., & Geban, O. (2014). Using three-tier diagnostic test to assess students' misconceptions of states of matter. *Eurasia Journal of Mathematics, Science and Technology Education*, *10*(5), 509–521. https://doi.org/10.12973/eurasia.2014.1128a

Korkmaz, S. D., Bahadir, A., Aybek, E. C., & Suat, P. A. T. (2018). Evaluating the gifted students' understanding related to plasma state using plasma experimental system and two-tier diagnostic test. *Journal of Education in Science Environment and Health*, *4*(1), 46–53.

Köse, S. (2004). Effectiveness of conceptual change texts accompanied with concept mapping instructions on overcoming prospective science teachers' misconceptions of photosynthesis and respiration in plants. *Published Ph. D., Karadeniz Technical University, Institute of Natural and Applied Sciences, Trabzon*.

Laliyo, L. A. R., Botutihe, D. N., & Panigoro, C. (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.

Leaper, C., Farkas, T., & Brown, C. S. (2012). Adolescent girls' experiences and gender-related beliefs in relation to their motivation in math/science and English. *Journal of Youth and Adolescence*, *41*(3), 268–282.

Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best? *Journal of Outcome Measurement*, *2*, 266–283.

Linacre, J. M. (2020). *Winsteps® (Version 4.7.0) [Computer Software].* (4.7.0). Winsteps.com. https://www.winsteps.com/

Liu, X. (2007). Elementary to high school students' growth over an academic year in understanding concepts of matter. *Journal of Chemical Education*, *84*(11), 1853.

Martin, R. E. (2005). *Teaching science for all children: Inquiry methods for constructing understanding*. Allyn & Bacon.

Morais, M. de F. (2013). *Creativity: Challenges to a key concept for the XXI century*.

Morrison, G. R., Ross, S. J., Morrison, J. R., & Kalman, H. K. (2019). *Designing effective instruction*. John Wiley & Sons.

Murdoch, J. (2018). Our preconceived notions of play need to challenging. *Early Years Educator*, *19*(9), 22–24. https://repositorium.sdum.uminho.pt/handle/1822/26465

OECD. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. https://doi.org/10.1787/9789264266490-en

OECD. (2020). *Science performance (PISA) (indicator)*. https://doi.org/doi: 10.1787/91952204-en

Park, M., & Liu, X. (2019). An Investigation of Item Difficulties in Energy Aspects Across Biology, Chemistry, Environmental Science, and Physics. *Research in Science Education*. https://doi.org/10.1007/s11165-019-9819-y

Peşman, H., & Eryılmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *The Journal of Educational Research*, *103*(3), 208–222.

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2), 1–14. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Ratnasari, D., & Suparmi, S. (2017). Effect of problem type toward students' conceptual understanding level on heat and temperature. *Journal of Physics: Conference Series*, *909*(1), 12054.

Samsudin, A., Afif, N. F., Nugraha, M. G., Suhandi, A., Fratiwi, N. J., Aminudin, A. H., Adimayuda, R., Linuwih, S., & Costu, B. (2021). Reconstructing Students' Misconceptions on Work and Energy through the PDEODE* E Tasks with Think-Pair-Share. *Journal of Turkish Science Education*, *18*(1), 118–144.

Soeharto, Csapó, B., Sarimanah, E., Dewi, F. I., & Sabri, T. (2019). A review of students' common misconceptions in science and their diagnostic assessment tools. *Jurnal Pendidikan IPA Indonesia*, *8*(2), 247–266. https://doi.org/10.15294/jpii.v8i2.18649

Soeharto, S. (2016). Implementation of Text Transformation in Physics Education to Reduce Students' Misconception. *JETL (Journal Of Education, Teaching and Learning)*, *1*(2), 56. https://doi.org/10.26737/jetl.v1i2.38

Stefanidou, C. G., Tsalapati, K. D., Ferentinou, A. M., & Skordoulis, C. D. (2019). Conceptual Difficulties Pre-Service Primary Teachers Have with Static Electricity. *Journal of Baltic Science Education*, *18*(2), 300.

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial (edisi revisi)*. Trim Komunikata Publishing House.

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Taslidere, E. (2016). Development and use of a three-tier diagnostic test to assess high school students' misconceptions about the photoelectric effect. *Research in Science & Technological Education*, *34*(2), 164–186.

TOPALSAN, A. K., & BAYRAM, H. (2019). Identifying Prospective Primary School Teachers' Ontologically Categorized Misconceptions on the Topic of" Force and Motion". *Journal of Turkish Science Education*, *16*(1), 85–109.

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.

Tsui, C., & Treagust, D. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education*, *32*(8), 1073–1098.

Wandersee, J. H., Mintzes, J. J., & Novak, J. D. (1994). Research on alternative conceptions in science. *Handbook of Research on Science Teaching and Learning*, *177*, 210.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education*, *40*(3), 393–411.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*(1), 1–28.