



Research article

Exploring Indonesian student misconceptions in science concepts

Soeharto Soeharto^{a,*}, Benó Csapó^{b,c}^a Doctoral School of Education, University of Szeged, 30-34, Petöfi S. sgt., Szeged, H-6722, Hungary^b Institute of Education, University of Szeged, Hungary^c MTA-SZTE Research Group on the Development of Competencies, Hungary

ARTICLE INFO

Keywords:

Rasch analysis
 Science concepts
 Misconceptions
 Confirmatory factor analysis

ABSTRACT

This study aims to investigate the development and the differences in student misconceptions in science based on gender and grade level, and to evaluate the developed two-tier multiple-choice diagnostic test in confirming the test's validity and reliability. A sample of 856 participants from 10th–12th graders and prospective science teachers were collected. The two-tier multiple-choice diagnostic test with 32 items covering biology, chemistry, and physics was administered to evaluate students' science misconceptions at the senior high school and university levels. The results met validity and reliability criteria using confirmatory factor analysis and Rasch parameters. The single-factor model has CFI = .973, RMSEA = .006, CI (.001, .014) and SRMR = .017 and the three-factor model has CFI = 0.939, RMSEA = .010, CI (.01, .017) and SRMR = .017. Based on the Rasch parameter, the infit and outfit MNSQ values achieve the acceptable fit (0.96 to 1) with good item reliability (.99) and person reliability (.80). All items have positive PTMA. Student misconceptions had significant differences in terms of grade and gender. We confirmed that prospective science teachers have higher misconceptions than 11th and 12th graders and slightly higher ones than 10th graders. Boys have a better conceptual understanding than girls based on the mean of correct answers. The multiple linear regression with the stepwise method confirmed that gender significantly predicted student misconceptions of science concepts, with 9% of variance explained. This study provided evidence that students and prospective teachers experience various misconceptions about science concepts.

1. Introduction

Students continuously develop attributes like knowledge, attitudes and experiences to learn new scientific concepts based on their interactions with the environment and construct their understanding of science by incorporating such attributes into their learning activities. In some cases, the construction of science-related concepts may lead to an incorrect grasp of these ideas, which persists even after learning in science class (Eshach et al., 2018; Prodjosantoso et al., 2019; Stefanidou et al., 2019). Allen (2014) also stated that students experience misconceptions in formal and informal settings unrelated to scientific knowledge. Moreover, students' misconception of science concepts is also triggered by the continuous development of science and technology; consequently, the meanings of science concepts change (Kaltakci-Gurel et al., 2017; Kiray et al., 2015). This condition makes conceptual learning an essential topic in science education to improve student achievement in science subjects. Students' misunderstanding or misconception refers to incorrect generalisations associated with their life experiences, teachers'

misinformation, student and teacher misconceptions and reflections of misconceptions in science textbooks (Chazbeck and Ayoubi, 2018). Students' understanding of science concepts may be different based on the scientific context and scientific facts; therefore, this study uses the term 'misconception' to represent students' misunderstanding or alternative conceptions.

Many studies (e.g. Laliyo et al., 2020; Park and Liu, 2019; Prodjosantoso et al., 2019; Slater et al., 2018; Wernecke et al., 2018) have found that student misconceptions are closely related to science learning and affect students' academic achievement in science subjects. Misconceptions in science learning are persistent and resistant to change (Wandersee et al., 1994). If students experience difficulties comprehending particular science concepts, they would have the same difficulties understanding related concepts in the future, resulting in low performance in learning science. Indonesia's science performance was the lowest of 41 countries in the 2018 PISA report (OECD, 2020). This is an initial indicator that demonstrates Indonesian students facing misconceptions in solving science problems. Therefore, this study

* Corresponding author.

E-mail address: soeharto.soeharto@edu.u-szeged.hu (S. Soeharto).

investigates students' misconceptions in science and explores the development of such misconceptions across grade levels, students at school and prospective science teachers (PSTs). PSTs are also chosen because many studies (e.g. [Butler et al., 2015](#); [Gurbuz, 2015](#); [Laliyo et al., 2020](#); [Liampa et al., 2019](#); [Stefanidou et al., 2019](#); [Tiruneh et al., 2017](#)) have found that they also experience difficulties in and misconceptions of various science concepts.

In Indonesia, Core competencies and learning indicators embedded in the national curriculum. the Ministry of Education and Culture (MOEC) composed the Indonesian national curricula (Curriculum, 2013). Teacher has cumulative task to teach student based on the Core competencies and learning indicators in each discipline. Students study science in general at the junior high school level (7th grade to 9th grade). The specific subject in science like Biology, Physics and Chemistry is only taught at the senior high school level (10th grade to 12th grade) ([Faisal and Martin, 2019](#)).

However, Student misconception in science rarely seems to be assessed on learning activity and tests in school level, or even on Indonesian national examination. Teacher focused on helping students achieving learning indicators based on national curricula without realizing if students may suffer misconception in particular science concepts. Assessment to identifying misconception or alternative conception in science is a pivotal aspect to improve student understanding related to science concepts. The well-constructed student conceptions in science will lead to students' development and achievement in the science education area. This study attempts to investigate students' misconceptions in science at the senior high school level and university level. Our instrument also attempts to map item difficulty level and compare the development of student misconceptions based on gender and grade school.

2. Literature review

2.1. Student misconceptions

Terms associated with students' understanding of science concepts include preconceptions, conceptual misconceptions, mental models, alternative conceptions and conceptual changes ([Allen, 2014](#); [Galvin and Mooney, 2015](#); [Gurbuz, 2015](#); [Gurel et al., 2015](#); [Mintzes et al., 2005](#); [Morais, 2013](#); [Taslidere, 2016](#); [Wandersee et al., 1994](#)). In this study, we focused on investigating misconceptions of science concepts. Conceptual misconception refers to incorrect student knowledge or misunderstanding that results in confusing knowledge as students build science-related insights ([Morais, 2013](#)). For example, students may find it difficult to understand the concept of light because they are unable to link the implementation in actual activity or practice. Many factors are connected with the process leading to knowledge construction based on initial beliefs, including knowledge from sensory experiences, cultural backgrounds, peers, teachers, textbooks and classroom learning ([Kiray and Simsek, 2021](#); [Liu and Mckeough, 2005](#); [Soeharto et al., 2019](#); [Vosniadou, 2012](#)).

Educators and scholars have described different conceptual changes experienced by an individual derived from their intuitive beliefs, life experiences, cultural influences and learning and teaching processes ([Arslan et al., 2012](#); [Galvin and Mooney, 2015](#); [Keeley, 2012](#)). Different terminologies and meanings regarding the nature of students' conceptual understanding reflect the application of misconceptions in various research areas ([Auhtors, 2019](#)). Misconceptions in science learning have been constantly studied because such misconceptions are persistent, resistant to change in students' minds and rooted in science concepts ([Taslidere, 2016](#); [Treagust, 1988](#); [Treagust and Duit, 2008](#); [Wandersee et al., 1994](#)). In addition, if students experience misconceptions or fail to correctly understand science concepts, they would find it difficult to understand and solve science problems, which would lead to low academic attainment in science disciplines ([Mintzes et al., 2005](#)). Students' misconceptions connected to science concepts need to be identified early so that teachers can construct knowledge that meets competency

requirements in science learning. Hence, misconceptions in science concepts are pivotal and essential to investigate in the science education field.

2.2. Prospective science teacher misconceptions

Studies involving preservice science teachers or PSTs have shown that misconceptions in science occur throughout the different education levels, even among senior teachers or professional teachers ([Becker and Cooper, 2014](#); [Duit, 2014](#); [Kiray and Simsek, 2021](#); [Laliyo et al., 2019](#); [Liampa et al., 2019](#); [Stefanidou et al., 2019](#)). [Kaltakci-Gurel et al. \(2017\)](#) found that PSTs sometimes share misconceptions that students hold in their knowledge. These misconceptions exist in learning design and learning activities, which directly reinforce students' misconceptions instead of correcting them. In Indonesia, science teachers have a special agenda called 'remediation' to correct students' misconceptions. Remediation activities are usually held after students' examinations in science disciplines, where science teachers reconstruct students' knowledge regarding particular science concepts. [Galvin and Mooney \(2015\)](#) highlighted the importance of identifying misconceptions of PSTs, undergraduate students in teacher training and education majors to improve the quality of science teachers and reduce student misconceptions. If science teacher misconceptions are not corrected, science teachers may fail to properly teach science concepts to students in their learning activities ([Arslan et al., 2012](#); [Gurbuz, 2015](#)).

2.3. Instruments for identifying science misconceptions

Researchers have developed several widely used instruments to identify student misconceptions in science. These include open-ended questions ([Tanahoung et al., 2010](#)), interviews ([Hamza and Wickman, 2008](#)), multiple-choice tests ([Lau et al., 2011](#); [D. Treagust, 1986](#)), concept maps ([Van Zele et al., 2004](#)) and multitier multiple-choice tests ([Arslan et al., 2012](#); [Galvin and Mooney, 2015](#); [Kirbulut and Geban, 2014](#); [Peşman and Eryilmaz, 2010](#)). The multiple-tier multiple-choice test has been the most popular assessment instrument, having been used by 33.06% of science education researchers from 2015 to 2019 ([Soeharto et al., 2019](#)). A two-tier diagnostic test, a type of multitier instrument used to determine student misconceptions, consists of two levels that assess scale content and student reasoning ([Korkmaz et al., 2018](#)). Students are considered to understand a science concept if they correctly answer the content and reasoning questions. The link between student conceptions and reasoning is the basis for developing a two-tier multiple-choice test ([Tsui and Treagust, 2010](#)). Using a two-tier multiple-choice test, researchers can define student knowledge in different ways and address some problems such as large sample size, scoring and understanding student reasoning ([Adadan and Savasci, 2012](#)).

However, the two-tier multiple-choice diagnostic test has certain deficiencies in identifying student misconceptions. [Gurel et al. \(2015\)](#) stated that the two-tier multiple-choice diagnostic test could not completely assess student misconception because of the lack of certainty in answering items brought about by the researcher's inability to confirm whether a student's answer was a guess or a correct conception. This weakness can be overcome through the Rasch measurement model, which can resolve guessing problems and detect outliers who answer via guessing. Therefore, this study attempts to apply Rasch modelling to identify student misconceptions using the two-tier multiple-choice diagnostic test.

2.4. Research questions

This study aims to investigate student misconceptions in science concepts across school grades, examine student-item interaction regarding science concepts, detect outliers in student misconceptions and predict background factors that influence students' misconception in

sciences. A two-tier multiple-choice diagnostic test was employed to answer the following 5 research questions:

- (1) Did the students provide guesses or inconsistent answers (i.e. misfitting persons) as their science misconceptions were assessed?
- (2) How did students and items interact based on the person–item map and grade levels?
- (3) To what extent does the collected data fit the Rasch and confirmatory factor analysis (CFA) models?
- (4) How do students' science misconceptions differ in terms of gender and grade level?
- (5) Which factors predict student conceptions in science?

3. Methods

3.1. Participants

The participants were recruited via a stratified random sample of 856 students (52.3% females and 47.7% males) from the 10th to the 12th grades at a senior high school and PSTs from three different universities in West Kalimantan Province, Indonesia. The paper-based test was administered, and participants spent 120 min completing the test under the supervision of researchers and teachers. However, not all participants were included in the data analysis, as the study applied Rasch modelling for data scaling to filter outliers.

3.2. Instruments

The research instruments in this study comprised a background questionnaire and a two-tier multiple-choice test, the former being embedded into the latter in a paper-based format. The background questionnaire contained information such as gender, grade level, school category and science score in the previous semester, whereas the two-tier multiple-choice test contained 32 questions from three science subjects: physics, chemistry and biology based on Indonesian national curricula. A total of 16 concepts pertaining to science misconceptions were selected and constructed into the two-tier multiple-choice test form. We chose common science concepts to determine student misconceptions based on a literature review and misconceptions in science learning handbooks (AAAS, 2012; Csapó, 1998; Soeharto et al., 2019). The instrument was checked by two experts in science education and one expert in English-Indonesian lecturer to confirm content validity whereby the pilot study had been done on 153 students at the senior high school level (Soeharto and Csapó, 2021). The physics dimension included kinetic energy, thermodynamics (thermal energy), atoms and molecules, impulse and momentum, light and force. For biology, we chose cells, breathing, microbes and disease, human body systems and feeding relationships. Finally, the chemistry aspect involved substances and chemical reactions, chemical compounds, chemical equilibrium, hydrocarbons and redox reactions. The newly design item and adapted item were also presented in Table 2. All science concepts were adjusted based on the K–12 curriculum (Curriculum 2013) implemented in the Indonesian educational system. Figure 1 represents a sample in the two-tier multiple-choice diagnostic test in this study.

3.3. Procedures and data analysis

To collect data, we asked permission to administer the test in the schools and universities, and the Institutional Review Board at the University of Szeged granted ethical research approval. With the guidance and supervision of researchers and teachers, the test was successfully administered. The statistical tools for data analysis included the Statistical Package for the Social Sciences (SPSS) version 25 (IBM Corp, 2017), MPLUS 8.4 (L. Muthén and Muthén, 2017) and Winsteps version 4.7.0 for Rasch measurement (Linacre, 2022). Students' total scores were converted into the log odd unit scale (logits) assumed as interval data

ranging from negative to positive infinity. Further, this study performed item–person maps, outlier analysis, model fit analysis, reliability and validity analyses, descriptive statistics, stepwise regressions, t-test and analysis of variance (ANOVA). All Rasch analysis procedure follows the guideline for Rasch analysis from Linacre (2021) and Boone et al. (2013).

4. Result

The findings were derived from the following research analyses: (1) scaling outliers based on misfitting person identification and person diagnostic maps (PKMAPs), (2) finding model fit based on confirmatory factor analysis using unweighted least squares (ULS) estimator and the Rasch measurement for item validity and reliability, (3) Wright maps to present item–person interactions, (4) t-test and ANOVA to measure differences based on gender and grade level and (5) multiple linear regression using the stepwise method to find factors that predict students' science conceptions.

4.1. Scaling outliers or misfitting persons

Before performing further analysis, we screened the data for outliers, also known as 'misfitting persons', which refer to student responses that show inconsistency or indicate guesswork. Rasch analysis allows researchers to screen the data for misfitting persons so that the data ascertain the true ability of students' scores to represent their ability to understand scientific concepts. From the dataset, we excluded 102 misfitting students out of 856 which involves 594 students at the senior high school level and 160 students at the university level. data were analysed using Rasch modelling and Winsteps version 4.7.0 based on the joint maximum likelihood estimation formula, wherein the raw data were converted into logits as interval data (Linacre, 2021). Table 1 shows the summary statistics of students and items in this study after excluding misfitting persons.

Misfitting students were identified based on person infit and outfit mean of the squared (MNSQ) criteria. If infit and outfit MNSQ values are outside the acceptable range of 0.5–1.5 (around 1.6 still acceptable), the student is included in the misfitting or outlier category (Andrich, 2018; Bond et al., 2020). Another indicator of misfitting students, person infit and outfit z-standardized (ZSTD), has acceptable values ranging from –2 to +2 in sequence (Bond et al., 2020). However, infit and outfit ZSTD can be ignored if the sample size is more than 500 and if the infit and outfit MNSQ criteria have been met (Azizan et al., 2020; Linacre, 2021).

We adopted PKMAPs to obtain more detailed information on the need for data scaling to detect outliers before further analysis. Stud121, a sample case from the misfitting student category (infit MNSQ: 1.67, outfit MNSQ: 2.19), had inconsistent response patterns in PKMAPs as shown in Figure 2. PKMAPs describe students' ability to respond according to the difficulty level of an item. In Figure 2, the most difficult items are at the top of the diagram, and the easiest ones are at the bottom. Correct student responses are on the left, whereas incorrect ones are on the right. While stud121 correctly answered the two most difficult items, numbers 31 and 32, they were incorrect in the easier items, such as numbers 12, 30 and 22, and such inconsistency in responses might have been due to the student's carelessness. Because the student's correct answers to more difficult items were higher than their logit ability, these responses are considered lucky guesses.

4.2. Wright map based on grade levels

The Wright map in Figure 3 illustrates the interaction between student ability and item difficulty based on grade. Item difficulty level is on the right side of the map, whereas student abilities based on four categories (10th grade, 11th grade, 12th grade and PST) are on the left side. The logit value determines the item's difficulty level (Boone et al., 2013): the higher the item logit, the more difficult the correctly answered item, and the lower the item logit, the easier the correctly answered item.

5. The Blue football player holds a Red football player who moves at speed (v), while lying on the ground. The players of blue football use body friction (F) on the soil and body mass (m) so that in the interval (Δt), the American red football players stop moving (watch the picture carefully). From this event, what is relation between impulse and momentums?



- a) Impulse is equal to the momentum
 b) Impulse is equal to changes in momentum
 c) impulse decreases momentum
 d) impulse increases momentum

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

- a) From the picture, it can be seen that the football player's speed has decreased so the momentum will also decrease.
 b) From the picture, it can be seen that the football player's speed has decreased so the momentum will also increase.
 c) From the picture, the momentum changes at a brief interval due to the friction of the football player's body with the ground.
 d) From the picture, the momentum is the same as impulse because in the same case on the energy transfer system.
 e)

Figure 1. A sample item in the two-tier multiple-choice diagnostic test on impulse and momentums in the physics task.

Figure 3 shows that the most difficult item to correctly answer is item 32 (redox reaction concepts) in the chemistry task, whereas the easiest item to correctly answer is item 1 (kinetic energy concepts) in the physics task.

Table 1. Summary statistics of students and items.

	Senior high school students		University students	
	Persons	Items	Persons	Items
N	754	32	754	32
Mean measure	0.70	0.00	0.70	0.00
Mean	18.7	454.8	18.7	454.8
SD	0.98	2.32	0.98	2.32
SE	0.49	0.11	0.49	0.11
Mean outfit MNSQ	1	1	1	1
Mean infit MNSQ	0.96	0.96	0.96	0.96
Separation	2	12.34	2	12.34
Reliability	0.80	0.99	0.80	0.99
Cronbach's alpha	0.82		0.82	
Raw variance explained by measures	36.1%		36.1%	
Chi-squared (χ^2)	21716.79 (df = 21746)		21716.79 (df = 21746)	
Probability	0.5544*		0.5544*	

* Normally distributed

Simultaneously, the Wright map evaluates student ability and item difficulty level using the same linear interval scale of item measure (logit). In addition, we found that student ability did not show significant differences for each grade level, indicating that the students experienced persistent misconceptions in science. Through the Wright map, we were able to evaluate how items and persons corresponded to the theoretical prediction.

4.3. Item reliability

Table 1 demonstrates that the internal consistency was assessed using Cronbach's alpha for all items and the item-person reliability parameter based on Rasch analysis. The Cronbach's alpha value for all items was 0.82, indicating high internal consistency and reliability (Taber, 2018); hence, all items were retained. Meanwhile, the Rasch model showed good person and item reliability values, which were 0.80 and 0.99, respectively (values higher than 0.67 indicate good reliability) (Fisher, 2007). Generally, in terms of reliability indicators, the two-tier multiple-choice diagnostic test met the acceptable threshold.

4.4. Validity of the two-tier multiple-choice diagnostic test

Confirmatory factor analysis for model fit.

One of the best measures for the construct validity of a research instrument is CFA. To perform CFA, we employed MPLUS 8.4 (Muthén and

Table 2. Item fit analysis.

Item	Science concept	Correct answer (%)	Measure (logit)	Infit MNSQ	Outfit MNSQ	PTMA	
PHY1	Kinetic energy	99.47	-5.34	0.96	0.12	0.20	(AAAS, 2012)
PHY2	Kinetic energy	83.69	-1.29	1.07	1.16	0.35	Authors
PHY3	Thermodynamics–Thermal energy	98.41	-4.18	1.03	0.36	0.23	Authors
PHY4	Thermodynamics–Thermal energy	70.69	-0.34	1.21	1.35	0.28	Authors
PHY5	Impulse and momentum	79.18	-0.91	0.76	0.60	0.64	Authors
PHY6	Impulse and momentum	59.81	0.27	0.93	0.97	0.49	Authors
PHY7	Atoms and molecules	43.24	1.1	0.81	0.75	0.56	(AAAS, 2012)
PHY8	Atoms and molecules	61.27	0.2	0.66	0.59	0.72	Authors
PHY9	Forces	61.94	0.16	0.59	0.53	0.77	(AAAS, 2012)
PHY10	Forces	37.53	1.38	0.80	0.68	0.56	Authors
PHY11	Light	43.10	1.1	0.76	0.67	0.61	Authors
PHY12	Light	20.56	2.35	1.04	0.92	0.27	Authors
BIO13	Cells	72.02	-0.42	1.23	1.45	0.25	(AAAS, 2012)
BIO14	Cells	87.93	-1.73	1.19	0.70	0.36	Authors
BIO15	Breathing	78.51	-0.86	1.05	1.15	0.41	Authors
BIO16	Breathing	73.34	-0.5	0.97	1.29	0.43	Authors
BIO17	Microbes and disease	51.86	0.67	1.36	1.32	0.15	Authors
BIO18	Microbes and disease	39.12	1.3	1.17	1.15	0.25	Authors
BIO19	Human body systems	50.80	0.73	0.98	0.97	0.44	(AAAS, 2012)
BIO20	Human body systems	61.41	0.19	0.82	0.76	0.60	Authors
BIO21	Feeding relationships	50.93	0.72	1.38	2.00	0.06	Authors
BIO22	Feeding relationships	43.10	1.1	1.03	0.98	0.38	Authors
CHEM23	Substances and chemical reactions	57.96	0.37	1.40	1.60	0.10	(AAAS, 2012)
CHEM24	Substances and chemical reactions	80.90	-1.05	1.03	0.95	0.43	Authors
CHEM25	Chemical compound	88.73	-1.82	0.93	1.35	0.37	Authors
CHEM26	Chemical compound	82.10	-1.15	0.96	0.96	0.46	Authors
CHEM27	Chemical equilibrium	70.56	-0.33	1.16	1.25	0.32	Authors
CHEM28	Chemical equilibrium	55.04	0.52	0.86	0.92	0.54	Authors
CHEM29	Hydrocarbons	38.86	1.31	0.95	0.89	0.43	(AAAS, 2012)
CHEM30	Hydrocarbons	24.01	2.12	0.92	0.76	0.39	Authors
CHEM31	Redox reaction	3.85	4.33	0.90	0.57	0.25	Authors
CHEM32	Redox reaction	0.13	8.97	1.00	1.00	0.00	Authors

Muthén, 2017) with two CFA models with the ULS estimator, as it provides more accurate results regarding standard errors, estimates and fit indices than weight least square (WLS) or maximum likelihood (ML) (Muthén 1993). CFA evaluated the model based on standardised root mean square residual (SRMR), comparative fit index (CFI) and the root mean square error of approximation (RMSEA). Goodness-of-fit indices measured how well the rotated matrix matched the original matrix. CFI required a large number of values and compared the real correlation matrix with the reproduced correlation matrix. RMSEA and SRMR pertain to the value of residual statistics, which are expected to be small in the residual matrix. Hence, we observed the following cut-off values to assess model fit: SRMR <.08, CFI >.90 and RMSEA <.06 (Caleon and Subramaniam, 2010; Hu and Bentler, 1999).

The first model proposed a single-factor CFA model with 32 items in a single group; this model showed acceptable goodness-of-fit indices. The results showed that all cut-off criteria values were met, all of which had a significant positive factor loading [CFI = .973, RMSEA = .006, CI (.001, .014) and SRMR = .017]. The second model proposed a three-factor CFA model based on biology, chemistry and physics tasks. The results showed that all cut-off criteria values were met and were less than the single-factor model [CFI = 0.939, RMSEA = .010, CI (.01, .017) and SRMR = .017]. Overall, the single-factor model showed the best fit, indicating acceptability in terms of construct validity and achieving unidimensionality in a single factor.

The measurement invariance was conducted to compare based on senior high school and university level through CFA in measurement models to confirm the measurement model in this study measures the

same underlying latent construct across the different groups. In other words, the instrument is not different if we measure two group levels, students from senior high school and university level. We found that there were no significant differences between senior high school and university levels in terms of group invariance. The invariance testing showed there are no significant invariances when comparing from Metric against Configural (Chi-square = 1.971, $p = 0.9223$), Scalar against Configural (Chi-square = 10.273, $p = 0.5920$), and Scalar against Metrics (Chi-square = 8.302, $p = 0.2168$).

4.5. Rasch analysis for item fit

The criteria applied to validate item-level appropriateness include the infit and outfit MNSQ, infit and outfit ZSTD and point-biserial correlations (PTMA). However, we excluded infit and outfit ZSTD because the sample size was more than 500 (Linacre, 2021). For infit and outfit MNSQ, the acceptable range is 0.5–1.5, with about 1.6 still acceptable (Andrich, 2018; Bond et al., 2020; Boone et al., 2013). All test items had positive PTMA, which evaluates whether items function according to the intended model in measuring a construct. PTMA was used as an additional threshold to confirm item fit. A positive PTMA value indicates that all items are acceptable, but a negative PTMA value shows that an item does not function well when compared with other items (Bond et al., 2020; Boone et al., 2013). Table 2 shows the results of the Rasch analysis using difficulty level (logit), infit and outfit MNSQ and PTMA. The item fit analysis results in Table 2 indicated that all items met the model fit criteria. Moreover, item separation (see Table 1) had a value of 12.34,

Name: stud121 18.00 1.00 1.00 12.00 3.00 1.00 3.00 4.00 2.00
 Ref. Number: 121 Measure: 2.10 S.E. .54 Score: 26
 Test: TDTS misconception 2nd 1 point fix (final data) rev 1.sav

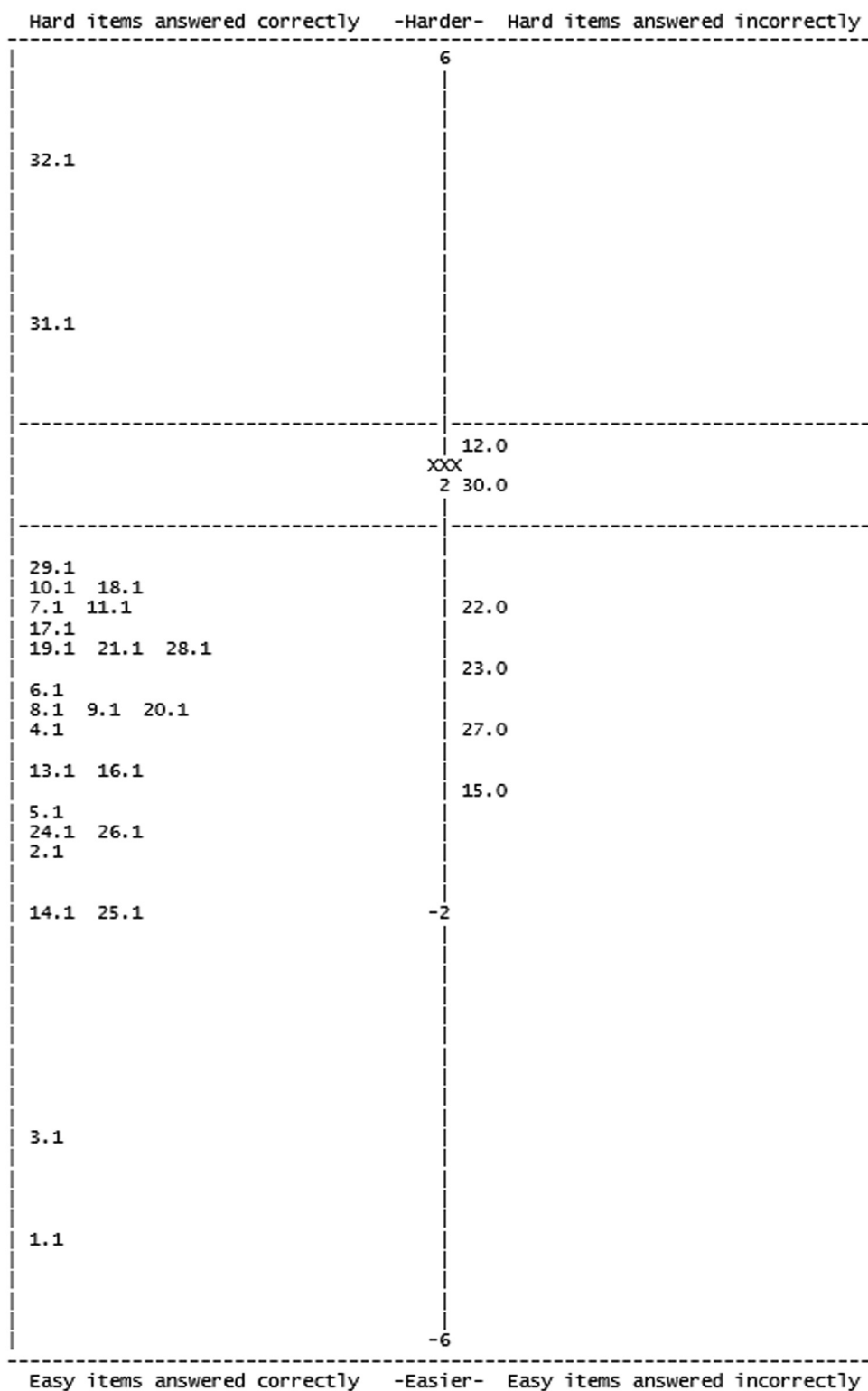


Figure 2. Responses by stud121 based on PKMAPs.

indicating various levels of item difficulty, and the person separation value was 2, showing that the test could distinguish at least two groups of students: high and low performance. Therefore, we included all items in the analysis because the infit and outfit MNSQ and PTMA criteria were fulfilled.

Figure 3 and Table 2 show that item 32 (CHEM32) is the most difficult item, but its value is still within the acceptable range based on infit and outfit MNSQ. Notably, however, this item seemed too difficult and needed to be revised to match sample targets; meanwhile, this result also indicated that students at every level have severe misconceptions (0.13%

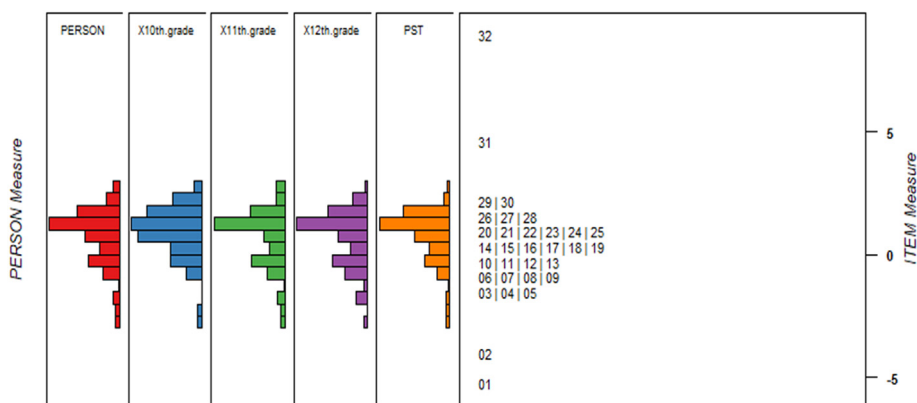


Figure 3. Wright item–person map based on grade levels.

correct answers) regarding redox reactions in chemistry. An item would be considered a misfit only if the three abovementioned criteria (infit MNSQ, outfit MNSQ and PTMA) are not achieved. Generally, we can assume that the collected data used all items in the two-tier multiple-choice diagnostic test from 10th, 11th and 12th graders and PSTs to assess scientific misconceptions matching the Rasch model.

Based on The principal component analysis of Rasch (PCAR), the test has achieved the unidimensionality assumption with the variance explained by measures was 38.5%. The unidimensional test can be achieved if the minimum variance explained by the measure is >30 % (Linacre, 1998). Items in the test have a residual correlation of around 0.1 and 0.28 confirming item dependency achieved whereby the raw residual correlation between pairs of the items <0.3 (Boone et al., 2013). The unidimensionality assumption is used to confirm the items in the instrument measure the same construct namely student misconception in science. This procedure follows the Rasch analysis for the unidimensional model using WINSTEPS Software (Boone et al., 2014; Linacre, 2021).

DIF analysis can be used in several background variables using categorical data in comparing items in a test (Boone et al., 2013). Differential item functioning analysis is categorized into three types: moderate to large ($|DIF| \geq 0.64$ logits), slight to moderate ($|DIF| \geq 0.43$ logits), and negligible (Zwick et al., 1999). To confirm item bias, the differential item functioning (DIF) analysis was utilized based on gender. The results confirm that all items do not have DIF based on gender. We found one item in chemistry (CHEM 32) with significant probability ($p < 0.01$), but the DIF size can be categorized as negligible, DIF contrast <0.43).

4.6. Differences in students' science misconceptions according to grade level

We performed ANOVA to compare students' conception scores across school grades and PSTs on the test and subtest. No significant differences were observed between students' understanding of science concepts in physics [$F(3,750) = 1.83, p > .05$] and chemistry [$F(3,750) = 1.51, p > .05$]. However, we found mean significant differences in the biology subtest [$F(3,750) = 3.34, p < .05$]. For the whole test, the results showed that student conception mean scores differed between grades [$F(3,750) = 2.653, p < .05$]. Because equal variances are not assumed based on Levene statistics ($p < .05$), we performed a Dunnett T3 test for post-hoc analysis to identify differences between cohorts, presented in Table 3.

Table 3 shows that students' conception scores are different between grade levels. Although ANOVA results for the entire test showed significant differences between cohorts, post-hoc analysis showed no significant differences with less than a 5% probability except for the biology subtest for 10th and 11th graders ($p = 0.25$) and for 10th and 12th graders, which showed substantial differences. This might indicate that student misconceptions are resistant to change, persistent and rooted deeply in science concepts, making it more difficult for higher-level students to understand science. Figure 4 shows that students at higher

levels (PSTs) develop higher misconceptions than other cohorts; for instance, Student 272 from the PST cohort correctly answered five of 32 items (around 15%), proving that higher-level students experience higher misconceptions than others.

4.7. Differences in students' science misconceptions based on gender

We conducted an independent-sample t-test to compare students' conceptions in the tests and subtests according to gender. The results showed significant differences in tests and subtests between boys and girls, with mean scores ranging from 4.87 to 19.21 as shown in Table 4. Boys' mean scores for the whole test and subtests were higher than those of girls, showing that boys comprehend science concepts and solve science problems better than girls. In addition, the mean score comparisons showed that the chemistry subtest was more difficult than the other subjects, as the mean scores of boys and girls in that subtest were lower than in the other subtests, confirming the item difficulty (logit) results in Table 2.

4.8. Predicting students' science misconceptions

To evaluate how some factors affect students' misconceptions in science, we performed multiple regression using the stepwise method, in which the predictors are school category, grade level, gender, father's education, mother's education and school performance. The results showed that only the gender predictor could significantly explain 9% of the variance on student misconception mean scores [$F(753) = 6.6, p < .05$]. This indicated that gender was a pivotal factor in predicting the science misconceptions of 10th, 11th and 12th graders and PSTs in Indonesia.

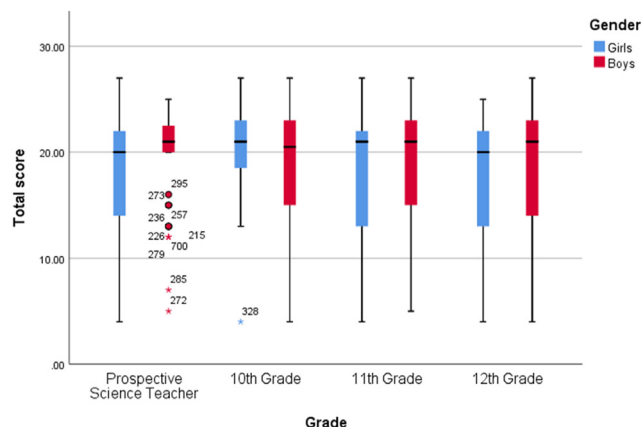


Figure 4. Comparison of student misconceptions between school grades.

Table 3. Dunnett T3 multiple comparisons of student conceptions between senior high school students and prospective science teachers.

Grade	Physics		Biology		Chemistry		Test	
	Mean differences	p	Mean differences	p	Mean differences	p	Mean differences	p
10th & 11th	.52	.24	.51	.02	.19	.83	1.23	.07
10th & 12th	.62	.22	.58	.04	.19	.91	1.40	.09
10th & PST	.26	.93	.35	.37	.10	.99	.72	.69
11th & 12th	.09	.98	.07	.99	-.01	.99	.17	.99
11th & PST	-.25	.94	-.16	.96	-.09	.96	-.51	.92
12th & PST	-.35	.86	-.23	.90	-.09	.98	-.68	.56

5. Discussion

The results showed that the two-tier multiple-choice diagnostic test could reliably assess students' misconceptions at the senior high school (10th, 11th and 12th grades) and PST levels. The test met the criteria for Cronbach's alpha (0.82) (Taber, 2018) and person and item reliability (0.80 and 0.99, respectively) (Fisher, 2007), which meant the test can be used in the same cohort range. The combination of reliability analysis based on internal consistency and item reliability based on Rasch parameters can provide more convincing results for researchers. The two-tier multiple-choice diagnostic test also showed good validity based on unidimensionality criteria, with 36.1% of variance explained by its measures, indicating that the test can evaluate a single dimension of science misconception. Meanwhile, the CFA analysis revealed that the single-factor or one-dimension model had higher fit indices than the three-factor model [CFI = .973, RMSEA = .006, CI (.001, .014) and SRMR = .017]. Infit and outfit MNSQ and PTMA values for all items indicated good item fit. However, the item CHEM32 (redox reaction) seemed too difficult to correctly answer (0.13%) and had a high difficulty level (8.97 logits). When we assessed item fit, we realised that each science concept had different difficulty levels. These findings were consistent with those of Park and Liu (2019), who examined item difficulty in several energy concepts. Therefore, we can assume that the two-tier multiple-choice diagnostic test in this study is valid and reliable in evaluating science concepts.

This study employed data screening analysis to identify outliers or misfitting persons. It excluded 102 of 854 students from the dataset because their person misfit parameters were outside the acceptable range (infit and outfit MNSQ, 0.5–1.6). This finding was similar to demonstrations of outlier detection by Chan et al. (2021) in assessing students' thinking ability. This means we can find misfitting persons in each test evaluation including the science context or beyond students' thinking ability. However, researchers have rarely applied outlier detection, especially in science education (e.g. Kaltakci-Gurel et al. (2017), Arslan et al. (2012), Caleon and Subramaniam (2010), Kiray and Simsek (2021) (Peşman and Eryılmaz (2010)),

Meanwhile, the Wright map depicted the construction and interaction between student conceptions in science and all items in terms of difficulty level (logit). The constructions of the two-tier multiple-choice diagnostic test covered all student ability levels. However, some items needed revising because they were either too difficult or too easy (e.g. CHEM32, PHY01 and PHY02). Item-person maps showed that students

Table 4. Independent-sample t-test comparing student conceptions according to gender.

Subject	Girl	Boy	t	p
	Mean (SD)	Mean (SD)		
Physics	7.38 (2.87)	7.80 (2.73)	-2.03	.042
Biology	5.94 (2.07)	6.24 (1.90)	-2.07	.039
Chemistry	4.87 (1.89)	5.16 (1.81)	-2.14	.032
All subjects (science)	18.20 (5.59)	19.21 (5.09)	-2.58	.010

with better ability are more likely to correctly answer more difficult items, whereas those with lower ability are more likely to incorrectly answer such items. When assessing students' misconceptions, science education research usually involves revising items (e.g. Laliyo et al., 2019, 2020; Park and Liu, 2019).

Table 2 presents a wide range of students' misconceptions in terms of science concepts and item difficulty level. The concepts in chemistry subjects appeared more difficult than those in other subjects, with correct answers ranging from 0.13% to 88.7%, especially in redox reaction. These results were consistent with those of Laliyo et al. (2019), Treagust et al. (2014) and Becker and Cooper (2014). Also, ANOVA showed significant differences in student mean scores between cohorts. However, in the post-hoc analysis, we found that significant differences were present only in biology among 10th, 11th and 12th graders; other subjects in various grade level combinations had no differences in mean scores. Interestingly, PSTs, which are considered higher-level students, experienced higher misconceptions than the other cohorts (see Table 3 and Figure 3). These findings supported the resistant, persistent and deeply rooted nature of science misconceptions (Arslan et al., 2012; Wandersee et al., 1994). Therefore, this study confirmed that higher-level students are more prone to science misconceptions than lower-level ones because of the nature of such misconceptions.

For gender, the independent-sample t-test results confirmed significant differences between girls' and boys' mean scores for subtests and the entire test, which ranged from 4.87 to 19.21, indicating that boys have a higher ability than girls in answering science problems in the test. This is supported by reports that boys are more affected by science motivation and parental role in achievement tests (Taskinen et al., 2015). A study by Shaheen and Kayani et al. (2015) also found that boys have higher ability than girls in understanding science concepts. While the stepwise multiple regression results showed gender as the pivotal factor in predicting students' misconceptions in this study, this does not dismiss the possible effect of other factors on students' science misconceptions, such as textbooks, teacher knowledge and students' mathematical abilities as described in a review of common science misconceptions by Soeharto et al. (2019).

The findings of this study have some implications for science teachers in the class context. Students have carried science misconceptions in grade and gender in particular science concepts. Teachers can use the finding to prepare the lesson plans for specific science concepts that will distribute misconceptions often to tackle student learning difficulties. Researchers can explore further how certain science concepts distribute misconceptions to students specifically with various research content. We also hope this study can lead other researchers to utilize Rasch measurement to identify student misconceptions in science.

6. Conclusions

To conclude, all items in the two-tier multiple-choice diagnostic test met reliability and validity criteria based on CFA and Rasch parameters. Rasch analysis helped to detect misfitting persons or outliers, that is, students with inconsistent answers and lucky guesses. We expect this new method to be used by other researchers before performing further data

analysis. Meanwhile, the Wright map showed the interaction between persons and items. However, because the item CHEM32 was considered too difficult and unsuitable for these cohorts, it must be revised if further tests are to be conducted. Further, we confirmed significant differences in student conception mean scores between all cohorts; however, post-hoc analysis results evinced that differences were present only among 10th and 11th graders, and 10th and 12th graders in the biology subtest. In addition, the independent-sample t-test results confirmed that boys' and girls' mean scores were significantly different in that the former had higher mean scores than the latter, which demonstrated that boys tend to demonstrate better comprehension of science concepts and can solve science problems better than girls. Multiple linear regression results also identified gender as an essential factor in predicting students' science misconceptions.

7. Limitations and future study

This current study fills the gap in assessing and investigating the development of misconceptions by high school students and PSTs. It is the first to employ the outlier detection method and Rasch parameters to measure students' conceptual understanding. However, this study has some limitations. First, because all respondents were from West Kalimantan, one of the provinces in Indonesia, one must exercise caution in generalising the results to all Indonesian students, although the Rasch analysis demonstrated that the sample had local independence, that is, the results are not dependent on the respondents. Second, this study performed quantitative analysis only; a mix of quantitative and qualitative methods may provide more meaningful insights. Finally, this study used only cross-sectional data; hence, we recommend that other researchers measure changes in student misconceptions via time series and longitudinal datasets.

For future study, we intend to explore the relations between students' science misconceptions and thinking skills such as inductive reasoning and student reasoning. In addition, we are interested in investigating science teachers' misconceptions more deeply, as doing so can help us understand the relation between student and teacher misconceptions in teaching and learning activities. Such research will be useful for educators when forming lesson plans and preparing science knowledge before conducting teaching activities.

Declarations

Author contribution statement

Soeharto Soeharto: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Analysis tools or data; Wrote the paper.

Benő Csapó: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Funding statement

This work was supported by University of Szeged Open Access Fund [5649].

Data availability statement

The data that has been used is confidential.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

- Adadan, E., Savasci, F., 2012. An analysis of 16–17-year-old students' understanding of solution chemistry concepts using a two-tier diagnostic instrument. *Int. J. Sci. Educ.* 34 (4), 513–544.
- Allen, M., 2014. *Misconceptions in Primary Science*. McGraw-hill education.
- American Association for the Advancement of Science (AAAS), 2012. AAAS Science Assessment - Project2061. <https://www.aaas.org/programs/project-2061>.
- Andrich, D., 2018. Advances in social measurement: a Rasch measurement theory. In: *Perceived Health and Adaptation in Chronic Disease*. Routledge, pp. 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. *Int. J. Sci. Educ.* 34 (11), 1667–1686.
- Azizan, N.H., Mahmud, Z., Rambli, A., 2020. Rasch rating scale item estimates using maximum likelihood approach: effects of sample size on the accuracy and bias of the estimates. *Int. J. Adv. Sci. Technol.* 29 (4), 2526–2531.
- Becker, N.M., Cooper, M.M., 2014. College chemistry students' understanding of potential energy in the context of atomic-molecular interactions. *J. Res. Sci. Teach.* 51 (6), 789–808.
- Bond, T.G., Fox, C.M., Lacey, H., 2020. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Routledge.
- Boone, W.J., Staver, J.R., Yale, M.S., 2013. *Rasch Analysis in the Human Sciences*. Springer.
- Boone, W.J., Staver, J.R., Yale, M.S., 2014. *Rasch Analysis in the Human Sciences*. Springer.
- 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. *Eur. J. Teach. Educ.* 38 (3), 300–319.
- Caleon, I., Subramaniam, R., 2010. Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *Int. J. Sci. Educ.* 32 (7), 939–961.
- Chan, S.-W., Looi, C.-K., Sumintono, B., 2021. Assessing computational thinking abilities among Singapore secondary students: a Rasch model measurement analysis. *J. Comp. Educ.* 8 (2), 213–236.
- Chazbeck, B., Ayoubi, Z., 2018. Resources used by Lebanese secondary physics teachers' for teaching electricity: types, objectives and factors affecting their selection. *J. Educ. Sci. Environ. Health* 4 (2), 118–128.
- Csapó, B., 1998. *Iskolai Tudás*. Osiris Kiadó.
- Duit, R., 2014. Teaching and learning the physics energy concept. In: *Teaching and Learning of Energy in K–12 Education*. Springer International Publishing, pp. 67–85.
- Eshach, H., Lin, T., Tsai, C., 2018. Misconception of sound and conceptual change: a cross-sectional study on students' materialistic thinking of sound. *J. Res. Sci. Teach.* 55 (5), 664–684.
- Faisal, Martin, S.N., 2019. Science education in Indonesia: past, present, and future. *Asia-Pac. Sci. Educ.* 5 (1).
- Fisher, W.P.J., 2007. Rating scale instrument quality criteria. *Rasch Meas. Trans.* 21 (1), 1095.
- Galvin, E., Mooney, S.G., 2015. Identification of misconceptions in the teaching of biology: a pedagogical cycle of recognition, reduction and removal. *High Educ. Soc. Sci.* 8 (2), 1–8.
- Gurbuz, F., 2015. Physics education: effect of micro-teaching method supported by educational technologies on pre-service science teachers' misconceptions on basic astronomy subjects. *J. Educ. Train. Stud.* 4 (2), 2010–2011.
- Gurel, D.K., Eryilmaz, A., McDermott, L.C., 2015. A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia J. Math. Sci. Technol. Educ.* 11 (5), 989–1008.
- Hamza, K.M., Wickman, P., 2008. Describing and analyzing learning in action: an empirical study of the importance of misconceptions in learning science. *Sci. Educ.* 92 (1), 141–164.
- Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.: A Multidiscip. J.* 6 (1), 1–55.
- IBM Corp., 2017. *IBM SPSS Statistics for Windows*. IBM SPSS Corp, Armonk, NY (Version 25) [Computer software].
- Kaltaki-Gurel, D., Eryilmaz, A., McDermott, L.C., 2017. Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Res. Sci. Technol. Educ.* 35 (2), 238–260.
- Keeley, P., 2012. Misunderstanding misconceptions. *Sci. Scope* 35 (8), 12–13.
- 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 Pac. Forum Sci. Learn. Teach.* 16 (2), 1–28.
- Kiray, S.A., Simsek, S., 2021. Determination and evaluation of the science teacher candidates' misconceptions about density by using four-tier diagnostic test. *Int. J. Sci. Math. Educ.* 19 (5), 935–955.
- Kirbulut, Z.D., Geban, O., 2014. Using three-tier diagnostic test to assess students' misconceptions of states of matter. *Eurasia J. Math. Sci. Technol. Educ.* 10 (5), 509–521.
- 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. *J. Educ. Sci. Environ. Health* 4 (1), 46–53.
- 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. *Int. J. Learn. Teach. Educ. Res.* 18 (9), 216–237.
- Laliyo, L.A.R., Puluhalawa, F.U., Eraku, S., Salimi, Y.K., 2020. The prevalence of students and teachers' ideas about global warming and the use of renewable energy technology. *J. Environ. Account. Manag.* 8 (3), 243–256.

- Lau, P.N.K., Lau, S.H., Hong, K.S., Usop, H., 2011. Guessing, partial knowledge, and misconceptions in multiple-choice tests. *J. Educ. Technol. Soc.* 14 (4), 99–110.
- Liampa, V., Malandrakis, G.N., Papadopoulou, P., Pnevmatikos, D., 2019. Development and evaluation of a three-tier diagnostic test to assess undergraduate primary teachers' understanding of ecological footprint. *Res. Sci. Educ.* 49 (3), 711–736.
- Linacre, J.M., 1998. Detecting multidimensionality: which residual data-type works best? *J. Outcome Meas.* 2, 266–283.
- Linacre, John M., 2021. Winsteps® Rasch measurement computer program User's Guide. Winsteps.com.
- Linacre, J.M., 2022. Winsteps® (Version 5.2.2) [Computer Software]. (5.2.2). Winsteps.com. <https://www.winsteps.com/>.
- Liu, X., Mckeough, A., 2005. Developmental Growth in Students' Concept of Energy : Analysis of Selected Items from the TIMSS Database, 42, pp. 493–517 (5).
- Mintzes, J.J., Wandersee, J.H., Novak, J.D., 2005. Assessing science understanding. In: *Assessing Science Understanding*. Elsevier.
- Morais, M. de F., 2013. Psychology of creativity. In: *Creativity: Challenges to a Key Concept for the XXI century*. Nova Science Publishers.
- Muthén, B.O., 1993. Goodness of fit with categorical and other nonnormal variables. In: *SAGE Focus Editions*, 154. Sage Publications, p. 205.
- Muthén, L., Muthén, B., 2017. Mplus User's Guide. Muthén & Muthén, Los Angeles, CA (eight edition)[Computer software manual].
- OECD, 2020. Science Performance (PISA) (Indicator).
- Park, M., Liu, X., 2019. An investigation of item difficulties in energy aspects across biology, chemistry, environmental science, and physics. *Res. Sci. Educ.*
- Peşman, H., Eryılmaz, A., 2010. Development of a three-tier test to assess misconceptions about simple electric circuits. *J. Educ. Res.* 103 (3), 208–222.
- Prodjosantoso, A.K., Hertina, A.M., Irwanto, 2019. The misconception diagnosis on ionic and covalent bonds concepts with three tier diagnostic test. *Int. J. InStruct.* 12 (1), 1477–1488.
- Slater, E.V., Morris, J.E., McKinnon, D., 2018. Astronomy alternative conceptions in pre-adolescent students in Western Australia. *Int. J. Sci. Educ.* 40 (17), 2158–2180.
- Soeharto, Csapó, B., Sarimanah, E., Dewi, F.I., Sabri, T., 2019. A review of students' common misconceptions in science and their diagnostic assessment tools. *J. Pendidikan IPA Indonesia* 8 (2), 247–266.
- Soeharto, S., Csapó, B., 2021. Evaluation and development of students' misconception using diagnostic assessment in science across school grades: a Rasch measurement approach. *J. Turkish Sci. Educ.* 18 (3), 351–370.
- Stefanidou, C.G., Tsalapati, K.D., Ferentinou, A.M., Skordoulis, C.D., 2019. Conceptual difficulties pre-service primary teachers have with static electricity. *J. Baltic Sci. Educ.* 18 (2), 300–313.
- Taber, K.S., 2018. The use of Cronbach's alpha when developing and reporting research instruments in science education. *Res. Sci. Educ.* 48 (6), 1273–1296.
- Tanahoung, C., Chitree, R., Soankwan, C., 2010. Probing Thai freshmen science students' conceptions of heat and temperature using open-ended questions: a case study. *Int. J. Phys. Chem. Educ.* 2 (2), 82–94.
- Taskinen, P.H., Dietrich, J., Kracke, B., 2015. The role of parental values and child-specific expectations in the science motivation and achievement of adolescent girls and boys. *Int. J. Gender Sci. Technol.* 8 (1), 103–123.
- Taslidere, E., 2016. Development and use of a three-tier diagnostic test to assess high school students' misconceptions about the photoelectric effect. *Res. Sci. Technol. Educ.* 34 (2), 164–186.
- Tiruneh, D.T., De Cock, M., Weldeclassie, A.G., Elen, J., Janssen, R., 2017. Measuring critical thinking in physics: development and validation of a critical thinking test in electricity and magnetism. *Int. J. Sci. Math. Educ.* 15 (4), 663–682.
- Treagust, D., 1986. Evaluating students' misconceptions by means of diagnostic multiple choice items. *Res. Sci. Educ.* 16 (1), 199–207.
- Treagust, D.F., 1988. Development and use of diagnostic tests to evaluate students' misconceptions in science. *Int. J. Sci. Educ.* 10 (2), 159–169.
- Treagust, D.F., Duit, R., 2008. Conceptual change: a discussion of theoretical, methodological and practical challenges for science education. *Cult. Stud. Sci. Educ.* 3 (2), 297–328.
- Treagust, D.F., Mthembu, Z., Chandrasegaran, A.L., 2014. Evaluation of the predict-observe-explain instructional strategy to enhance students' understanding of redox reactions. In: *Learning with Understanding in the Chemistry Classroom*. Springer Netherlands, pp. 265–286.
- Tsui, C., Treagust, D., 2010. Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *Int. J. Sci. Educ.* 32 (8), 1073–1098.
- Ul Khalil Shaheen, M.N., Juman, N.B., Kayani, M.M., 2015. Improving students' achievement in biology using 7E instructional model: an experimental study. *Mediterr. J. Soc. Sci.* 6 (4), 471.
- Van Zele, E., Lenaerts, J., Wieme, W., 2004. Improving the usefulness of concept maps as a research tool for science education. *Int. J. Sci. Educ.* 26 (9), 1043–1064.
- Vosniadou, S., 2012. Reframing the classical approach to conceptual change: preconceptions, misconceptions and synthetic models. In: *Second International Handbook of Science Education*. Springer Netherlands, pp. 119–130.
- Wandersee, J.H., Mintzes, J.J., Novak, J.D., 1994. Research on alternative conceptions in science. In: *Handbook of Research on Science Teaching and Learning*, 177, p. 210.
- Wernecke, U., Schütte, K., Schwanewedel, J., Harms, U., 2018. Enhancing conceptual knowledge of energy in biology with incorrect representations. *CBE-Life Sci. Educ.* 17 (1) ar5.