**UNIVERSITY OF SZEGED**

**DOCTORAL SCHOOL OF EDUCATION**

**LEARNING AND INSTRUCTION**

SOEHARTO SOEHARTO

# ASSESSING STUDENTS' SCIENCE MISCONCEPTIONS AND INDUCTIVE REASONING: CROSS-SECTIONAL STUDIES IN INDONESIAN CONTEXT

THE DISSERTATION ABSTRACT

Supervisor:
Prof. Dr. Benő Csapó
Professor of Education

**LEARNING AND INSTRUCTION EDUCATIONAL PROGRAMME**

**SZEGED, HUNGARY,**
**2023**

## Context of study and statement of the problem

In the PISA report 2018 (OECD, 2020) for student performances in science, Indonesian students performed the worst out of 79 nations, which may indicate that most Indonesian students struggle to understand scientific notion during the learning process. Numerous studies (e.g., Arslan et al., 2012a; Keeley, 2012; Mubarokah et al., 2018; Samsudin et al., 2021; Soeharto, Csapó, et al., 2019; Soeharto & Csapó, 2021, 2022a) have demonstrated the connection between scientific misconceptions and student academic achievement and how they affect student learning activity in science Hence, it stands to reason that if students struggle to understand a particular scientific subject, they will likely struggle in the future or during the learning process, which will lead to poor science achievement.

Through classroom instruction and outside learning, students build their knowledge. Students have prior knowledge, abilities, and experience that shape their initial notions in scientific learning before engaging in a learning activity at school. Although this condition still exists after the science learning activity is completed, these initial conceptions may be in conflict with scientific concepts, called misconceptions in science (Eshach et al., 2018; Köse, 2004; Stefanidou et al., 2019). Various studies had been conducted to determine the different science-learning ideas that cause student misconceptions in science. Soeharto et al. (2019) also discovered 111 articles from 2015 to 2019 that focused on student misconceptions in science. Wandersee et al. (1994) analyzed 103 studies related to misconceptions, Gurel et al. (2015) discovered 273 articles about misconceptions, and Wandersee et al. (Wandersee et al., 1994) examined 103 studies related to misconceptions. There are three publications (Fajarini et al., 2018; Fariyani et al., 2017; Ratnasari & Suparmi, 2017) that talk about detecting student misconceptions in Indonesia and how this relates to the dearth of research issues in the country's field of scientific education. However, these recent Indonesian publications mainly focused on identifying student misconceptions in a single science concept, such as global warming, optics, or heat, and there is no instrument now being developed from science ideas dispersing student misunderstanding in learning science. There is also lack of evidence how is the pattern of misconceptions in sciences and how are students' ability in solving science concepts. Therefore, there is a need to investigate Indonesian students' misconceptions in science with various background factors such as gender and grade levels.

In Indonesia, the 2013 Indonesian core curriculum included thinking skills (Hasan, 2013; Prastowo & Fitriyaningsih, 2020). The learning material was created to link to the fundamental competencies in several disciplines in the three primary domains of attitude, skills, and knowledge supported by this curriculum (Hasan, 2013). This curriculum has a significant issue with evaluation practices, particularly when assessing attitude. It was challenging to adjust the attitude assessment to the setting of the classroom because it was brand-new. According to Badaruddin & Hawi (2022), the majority of teachers expressed frustration over how challenging it was to gauge student attitudes and that their knowledge of the best methods and evaluation tools was still lacking. However, it was simple to evaluate knowledge and abilities (Natsir et al., 2018). The teacher may use a variety of learning models on various resources and subjects to improve students' thinking skills (Prastowo & Fitriyaningsih, 2020). The inductive reasoning test has been used in the general basic skills knowledge test when applying for positions at the government and corporate levels, despite the fact that it is not taught and trained explicitly in schools. Limited data and studies were related to inductive reasoning in classrooms and even in institutes of higher learning. Therefore, there is a need to perform an evaluation of Indonesian student inductive reasoning to be a pioneer for assessment of inductive reasoning in Indonesia.

Consequently, evaluation of student misconceptions in science and inductive reasoning skills in Indonesian context are topics that can be a foundation for further researches in educational area. In this research, The Rasch measurement mainly use to perform objective measurement because it can be useed to validate instrument, investigate items and examine person ability and interaction. The Rasch measurement approach is a widely used statistical method in the educational field, particularly in the area of educational assessment (Masters, 1982; Soeharto, 2021; Soeharto & Csapó, 2022b; Sukarelawan et al., 2021). However, despite its widespread use in the educational field, the Rasch measurement approach is relatively rarely applied in developing measurement instruments in Indonesian context. Rasch measurement approach is based on the idea that measurement should be based on the concept of equal intervals, meaning that the difference between any two scores should have the same meaning, regardless of the specific values of those scores (Bond & Fox, 2015; Boone et al., 2014).

Indonesia implemented the 2013 curriculum for more than 10 years. This curriculum focus on three domains namely attitude, skills, and knowledge. However, there is no specific assessment to identify students' misconceptions in science and inductive reasoning skills. Whereas these both construct is important in guiding students' achievement in academic and work field. To start the investigation of students' ability to understand science concepts and inductive reasoning skills before investigating the structural model or causal relationship stage. There is a need of assessment in comprehensive work to pioneer this research topic.

In addition, the literature review conducted by Soeharto et al. (2019) have confirmed that topics of physics, chemistry, and biology subject in science were distributing misconception for the student in Indonesia from 111 published studied reviewed. However, only four studies that measuring misconception. The studies of inductive reasoning in Indonesian context are also limited in schools and higher education context (Istikomah et al., 2017; Siswono et al., 2020). Furthermore, the inductive reasoning test has been used in the general basic skills knowledge test when applying for jobs at the government and company levels, even though inductive reasoning is not explicitly taught and studied in schools. Therefore, there is a need to do assessment to identify student misconceptions in science and inductive reasoning skills in Indonesian context.

## The structure of a dissertation

This dissertation is composed based on two cross sectional studies from pilot and main study with five different published studies in the assessment topic of student misconceptions in science and inductive reasoning skills (Soeharto, 2021; Soeharto et al., 2019; Soeharto & Csapó, 2021, 2022b, 2022a). The dissertation consists of five different chapters. Chapter one is the introduction which consist of the study context, statement of the problem and organization of dissertation. Chapter two is a review literature on studies related to the research topic in this dissertation. The main focus was on assessment of student misconception in sciences and inductive reasoning skills in Indonesian context. Chapter three focus on study aims, research question, structure of empirical studies, and the methodology section was used in the empirical studies which focus on design, sampling procedure, data collection, data analysis, instrument and validation. Chapter four presents four empirical studies in this dissertation. The systematic review of students' common misconceptions in science and its' diagnostic assessment tools was included in chapter two. This systematic literature review focuses on initial investigation of topics in science causing student misconception and what kind of instruments used in previous studies. The first empirical study is the evaluation and development of students' misconception using diagnostic assessment in science across school grades. This study actually a pilot study as an initial stage in developing two-tier multiple choice test in measuring student misconceptions in science. Study two is the evaluating item difficulty patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts. This study focuses in investigating item difficulty patterns across the science subject using Rasch measurement approach. Study three is an investigation of Indonesian student misconceptions in science concepts in specific using Rasch measurement approach. The last study, study four is a comprehensive assessment of Indonesian inductive reasoning skills and validation of inductive reasoning test using Rasch measurement approach.

## Methods, objectives and empirical studies

Cross-sectional studies are a type of research design that is employed in academic studies to collect data at one moment in time and examine the relationships between various variables (Creswell & Creswell, 2017; Leedy & Ormrod, 2005). The design is employed to spot patterns or trends in data or to try theories regarding the frequency of particular traits in a population (Merriam & Tisdell, 2015). Examining various aspects of misconceptions in science, such as such as exploring students' understanding, measuring item difficulty level and assessing inductive reasoning skills, can be done using cross-sectional studies (Soeharto & Csapó, 2021, 2022a, 2022c).

In this dissertation, researchers gather information from a representative sample of participants during a cross-sectional study at a particular moment. The sample is chosen to guarantee that it accurately reflects the traits of the target community (Yin, 2018). To make sure the sample is representative, different sampling methods can be used, such as stratified sampling or random sampling (Creswell & Creswell, 2017; Leedy & Ormrod, 2005). The sample number should be sufficient to guarantee statistical power and the validity of the findings (Merriam & Tisdell, 2015). Table 1 illustrates the cross-sectional studies that had been conducted in this dissertation.

Table 1. Cross-sectional studies from pilot and main study in this dissertation.

| Timeline | Main objective | Instrument | Sample |
|---|---|---|---|
| May to June 2019 (pilot study) | 1. Conducting pilot study<br>2. Checking the psychometric properties of the developed instrument<br>3. Examining student misconceptions in science learning<br>4. identifying background factors affecting student misconceptions in the learning context. | 1. Background questionnaire<br>2. The two-tier multiple-choice test | 10th, 11th, and 12th<br>N =152 |
| September – June 2021 (main study) | 1. Investigating item difficulty patterns<br>2. Evaluating item–person map interaction<br>3. Checking the DIF based on gender and grade across science disciplines | 1. Background questionnaire<br>2. The two-tier multiple-choice test<br>3. IR Test | 10th, 11th, 12th<br>and PST<br>N =856 |
| September – June 2021 (main study) | 1. Investigating student misconceptions in science concepts across school grades<br>2. examining student–item interaction regarding science concepts<br>3. detecting outliers in student misconceptions<br>4. predicting background factors that influence students' misconception in sciences | 1. Background questionnaire<br>2. The two-tier multiple-choice test<br>3. IR Test | 10th, 11th, 12th<br>and PST<br>N =856 |
| September – June 2021 (main study) | 1. Assessing the adapted Indonesian version of the inductive reasoning test<br>2. Classifying their inductive reasoning levels in accordance with grade and gender. | 1. Background questionnaire<br>2. The two-tier multiple-choice test<br>3. IR Test | 10th, 11th, 12th<br>and PST<br>N =856 |

**Study 1: Evaluation and development of students' misconceptions using diagnostic assessment in science across school grades: A Rasch measurement approach.**

We analyzed the psychometric properties of the developed instrument based on Rasch measurement model. 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 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 achieve the normal distribution criteria and Rasch model fits globally (Boone et al., 2013; Engelhard Jr, 2013; Linacre, 2020). The reliability is calculated based on item internal consistency 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 that indicated 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).. Therefore, we can conclude that the developed two-tier multiple-choice used in this study is valid and reliable.

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). We categorized person ability 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%).

DIF analysis was conducted to check whether there were items bias based on gender. DIF analysis (Figure 2) shows that the items PHY1 and CHEM32 have DIF bias in the moderate to large category. These two items was also misfit item. Items PHY1 and CHEM32 explained that these two items were more difficult for boys than girls to answer correctly.

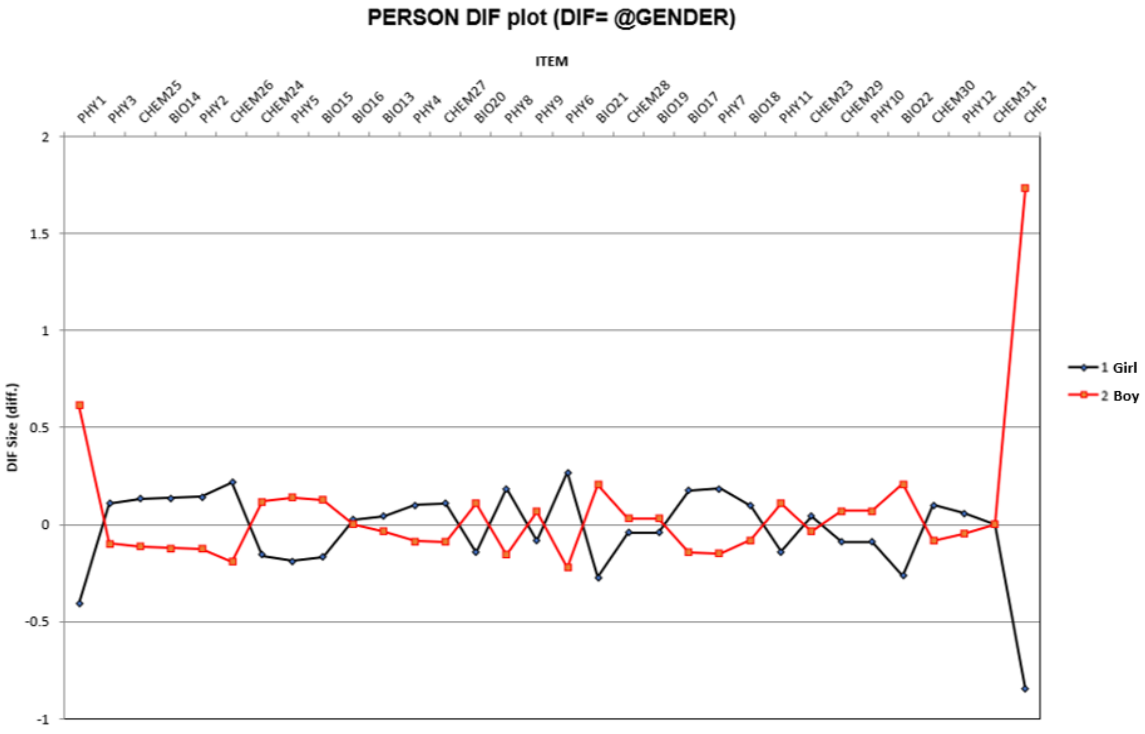

Figure 2. DIF based on gender.

ANOVA was conducted to determine the comparison of student misconceptions across school grades of student misconceptions in science 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], The entire test [F (2, 152) = 10.93, p <.01]. Because the equal variances are not assumed, we ran Dunnett T3-test 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.

Table 1. 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 | .54 | 0.06 | .99 | 0.30 | .72 | 0.93 | .56 |
| 10th & 12th | -1.35 | .06 | -1.31* | .01 | -0.82 | .07 | -3.61* | .00 |
| 11th & 12th | -1.94* | .00 | -1.38* | .00 | 1.13* | .01 | -4.55* | .00 |

No significant differences were found in the test and whole grade school level (p> .05). This also indicates that each cohort is not different between girls and boys. 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).

To explore how other factors predict student misconceptions in science, we ran the stepwise multiple regression 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 <.01. These results indicated that grade school is an essential factor in developing student misconceptions in learning science at senior high school.

**Study 2: Evaluating item difficulty patterns for assessing student misconceptions in science across Physics, Chemistry, and Biology concepts**

In this study we focus on item investigation using larger sample size. We calculated the standard deviation (SD) and the mean of average item difficulty measure for each of the three science disciplines, that is, physics, biology, and chemistry, using item difficulty estimates or logits of items. The mean of items in biology was placed as the easiest on the basis of the mean of item difficulties. Additionally, we also calculated the item difficulty estimates (measure) on the basis of the 16 science concepts as shown in Table 5 in this study. When comparing item difficulty for each concept, the redox reaction (CHEM 32) with 5.06 logits was the most challenging item to solve among all of the items in chemistry, and kinetic energy (PHY1) with −5.13 logits was the easiest item among all of the items in physics.

A two-way Analysis of Variance (ANOVA) was used to analyze the effect of science concepts and science discipline on item difficulty estimates based on logits. The 2 × 2 ANOVA group in this study achieved the assumption of homogeneity variances based on Levene's test ($p > 0.05$). To validate the normality data assumption, the Kolmogorov–Smirnov test was run before conducting the two-way ANOVA. The results showed that the item difficulty estimates did not differ significantly from normality ($p > 0.05$) with kurtosis (2.21) and skewness (−0.14).

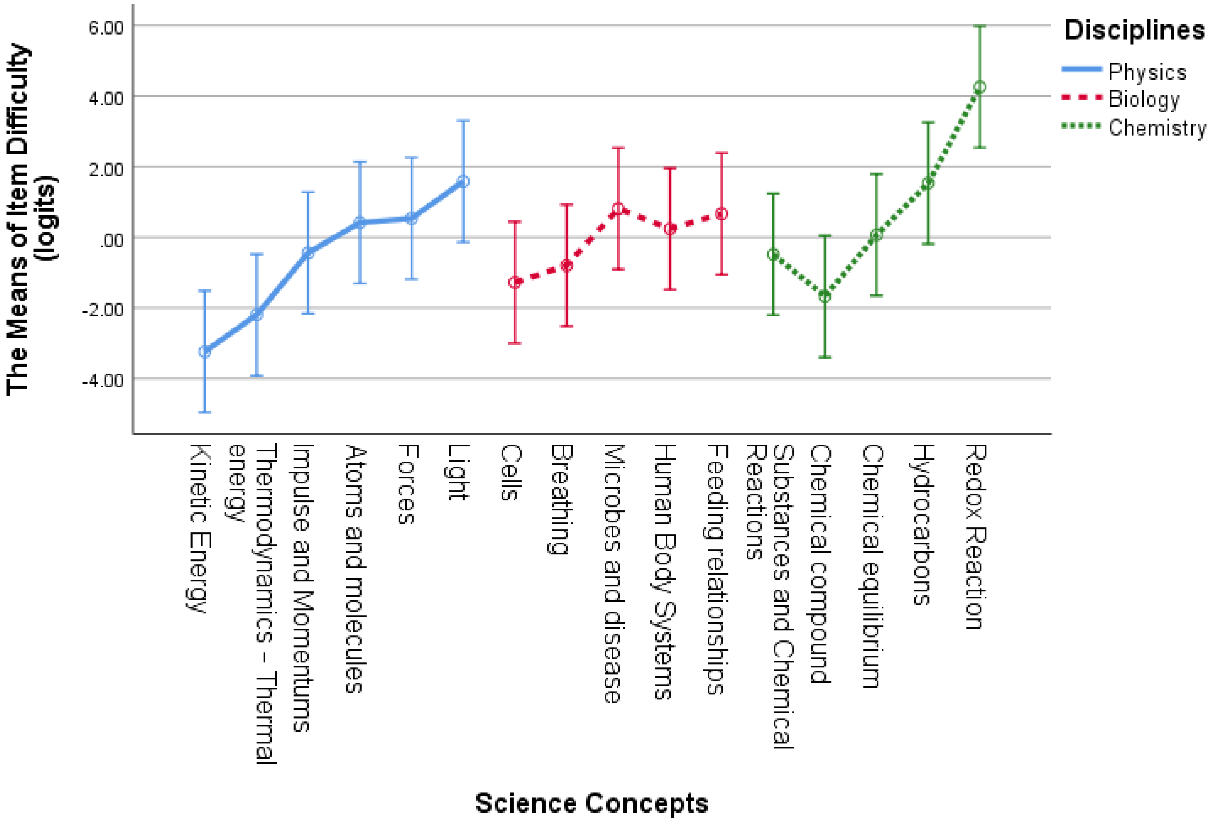

Figure 3. Item difficulty patterns between science concepts and across science disciplines.

The results showed a significant effect of science concepts on item difficulty estimates with a large effect size, F(13) = 4.76, $p < 0.0$. Also, the interaction effect of science disciplines and science concepts showed a significant effect on item difficulty estimates F(15) = 4.59, $p < 0.0$. However, the difference of item difficulties estimates among science disciplines was found to be insignificant, F(2) = 1.30, $p > 0.05$. To visualize the item difficulty pattern from each concept among disciplines, we calculated the mean of item difficulty pattern for each concept in Figure 4 .Both the science concepts and science disciplines can explain 81% of the variance on item difficulty estimates. To sum up, these findings indicated that the item difficulties pattern varies across science concepts, although there are no significant mean differences of item difficulties among disciplines.

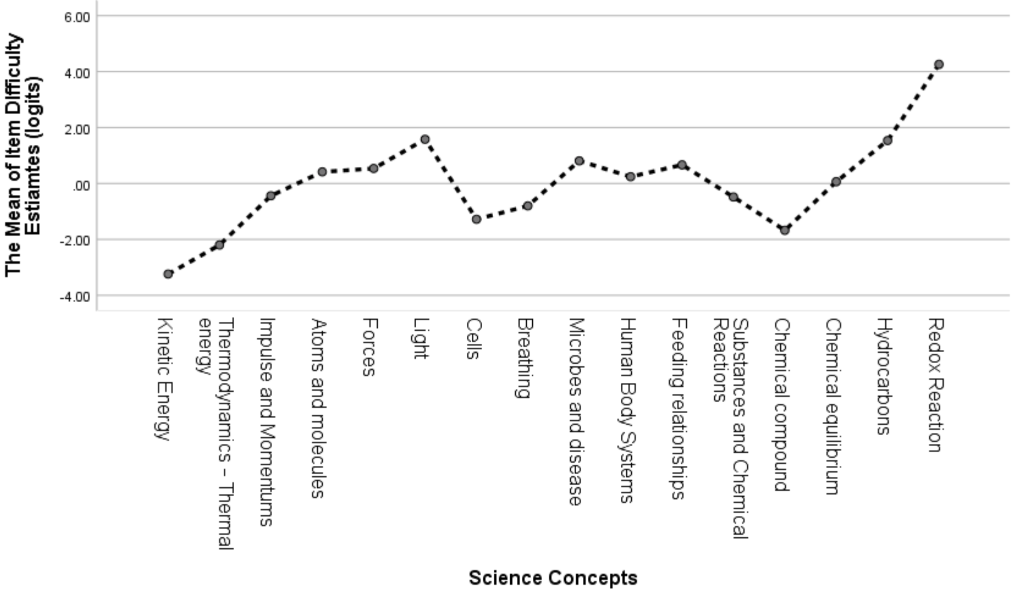

Figure 4. The mean of item difficulty estimates based on science concepts.

DIF analysis was performed to assess differences in item function on the basis of gender and grade on all items in test. Overall, items do not have DIF based on gender, except one item in chemistry (CHEM 32). For DIF based on grade, we compared four different cohorts: 10th grade, 11th grade, 12th grade, and the PST. Four items are categorized to differ based on grade: PHY1, PHY5, CHEM23, and CHEM32 (see Figure 5).

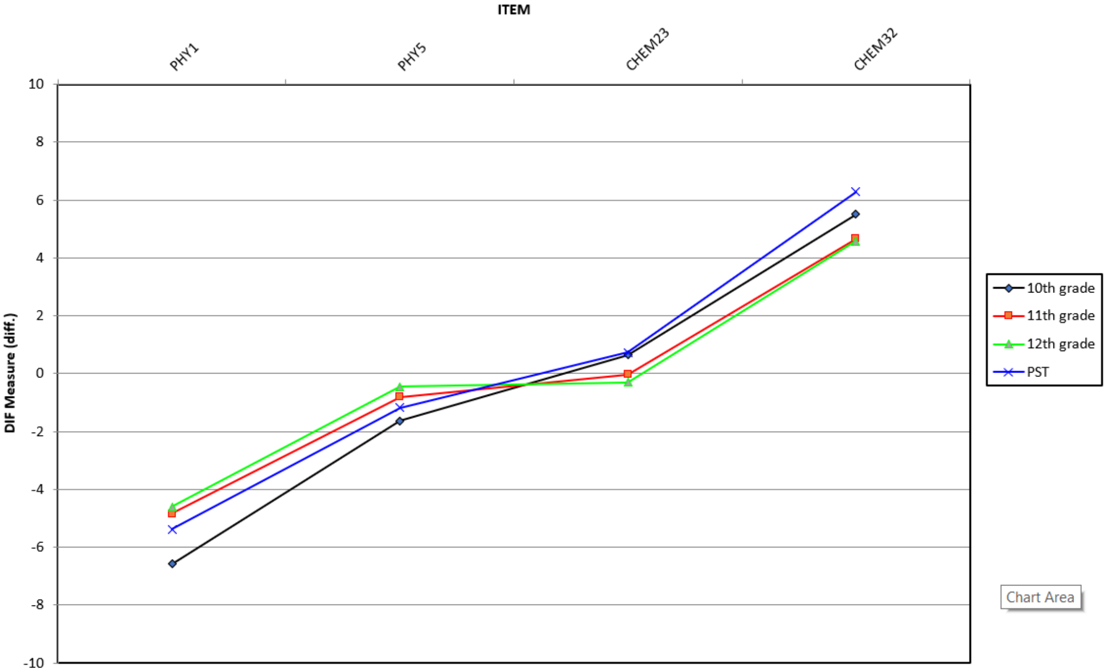

Figure 5. DIF measure based on grade

**Study 3: Exploring Indonesian student misconceptions in science concepts.**

In this study, we focus on exploring student misconception in science concepts. 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. We adopted PKMAPs to obtain more detailed information on the need for data scaling to detect outliers before further analysis.

The Wright map in Figure 6 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.

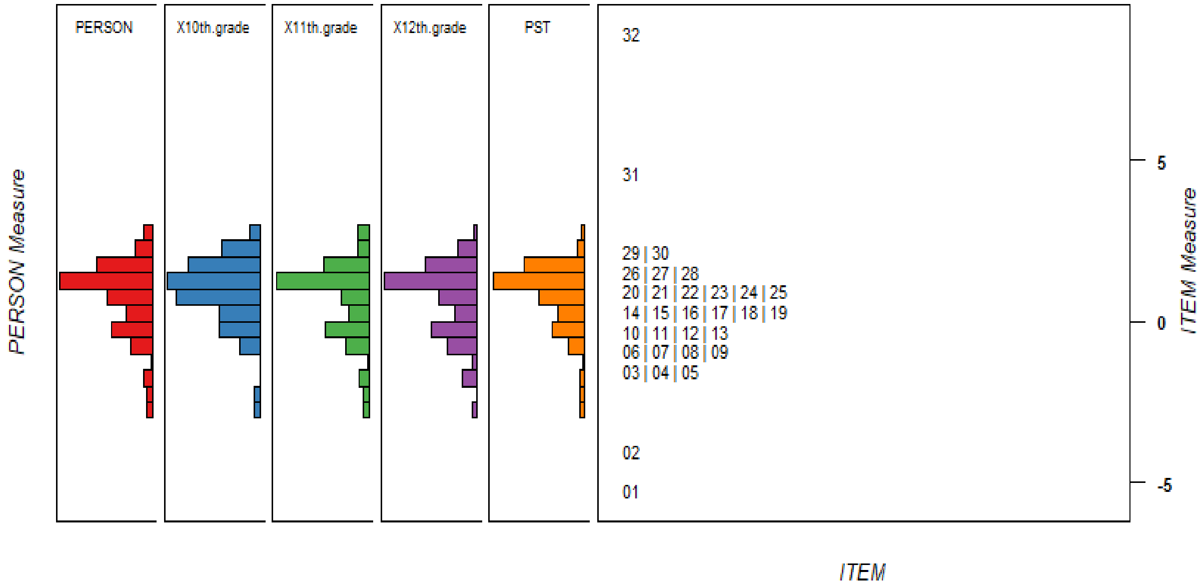

Figure 6. Wright item–person map based on grade levels.

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 2. 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 7 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.

Table 2. 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 |

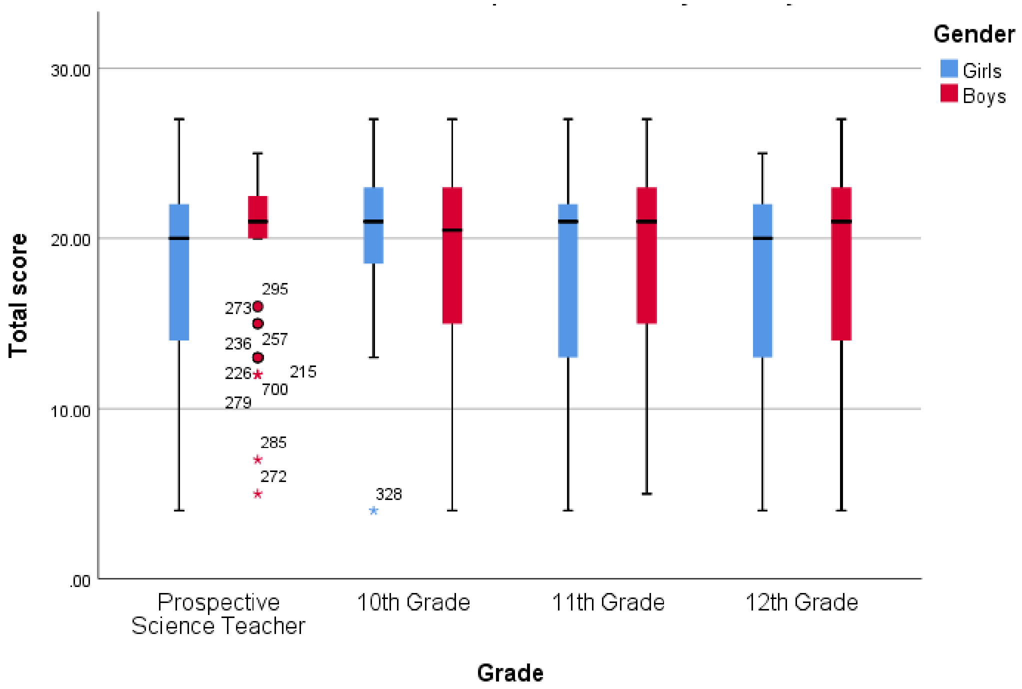

Figure 7. Comparison of student misconceptions between school grades.

**Study 4: Assessing Indonesian student inductive reasoning: Rasch analysis.**

The results confirmed that inductive reasoning for the reasoning test adapted for Indonesia achieved validity in accordance with the Rasch parameter for each task and entire test. It was considered that the FA task met the person separation threshold, with person separation close to 2 logits.

Table 2. The summary of Rasch parameters for inductive reasoning test and task

| Psychometrics Attribute | Task | | | | IR test |
|---|---|---|---|---|---|
| | FA | FS | NA | NS | |
| Number of Items | 10 | 10 | 10 | 10 | 40 |
| Mean | | | | | |
| item outfit MNSQ | 0.95 | 0.98 | 1.16 | 1.54 | 1.01 |
| item Infit MNSQ | 1.00 | 0.98 | 0.98 | .99 | 1.00 |
| person outfit MNSQ | 0.95 | .98 | 1.16 | 1.13 | 1.01 |
| person Infit MNSQ | 1.00 | .99 | 0.98 | 0.96 | 1.00 |
| Item separation | 10.27 | 12.07 | 13.62 | 14.79 | 16.46 |
| Person separation | 1.98 | 2.18 | 2.25 | 2.82 | 2.92 |
| Unidimensionality | | | | | |
| Raw variance by measure | 30.2% | 36.6% | 36.1% | 53.7% | |
| Unexplained variance 1st contrast | 1.72 | 1.97 | 1.70 | 2.03 | |

The reliability criteria were evaluated following several indicators, including Rasch parameters using person and item reliability (Fisher, 2007; Linacre, 2021), Cronbach's Alpha (α) (Taber, 2018) and McDonald’s omega (ω) (Dunn et al., 2014). WINSTEPS software will generate person reliability, item reliability and Cronbach's Alpha (α), and SPSS was utilized to compute McDonald’s omega (ω). Cronbach's Alpha (α) values ranged from 0.61 to 0.77 for all the tasks as well as the entire test, thus indicating sufficient reliability (Taber, 2018), and McDonald’s omega (ω) ranges from 0.54 to 0.75, thus confirming acceptable reliability was achieved for only in the test level with 0.75 (Dunn et al., 2014). However, for person reliability and item reliability, the values range from 0.68 to 1.00. Fisher (2007) noted that values more than 0.67 demonstrated acceptable reliability. Overall, the adapted inductive reasoning test and all its tasks exhibited acceptable criteria for the Rasch reliability parameter.

DIF analysis used in this study was the uniform DIF analysis that compares all ability levels of the two or more groups. However, NS6 had moderate to large DIF. Furthermore, the online-based test was more difficult for students than the paper-based test with regard to NS6 item, with 0.94 logits of DIF size, $p < 0.05$. FS2 and FA7 were classified as having negligible DIF. The DIF analysis based on the test method is illustrated in Figure 8. Based on gender and grade, We can assume the IR test can hold invariance confirming no DIF issue across grade and gender.

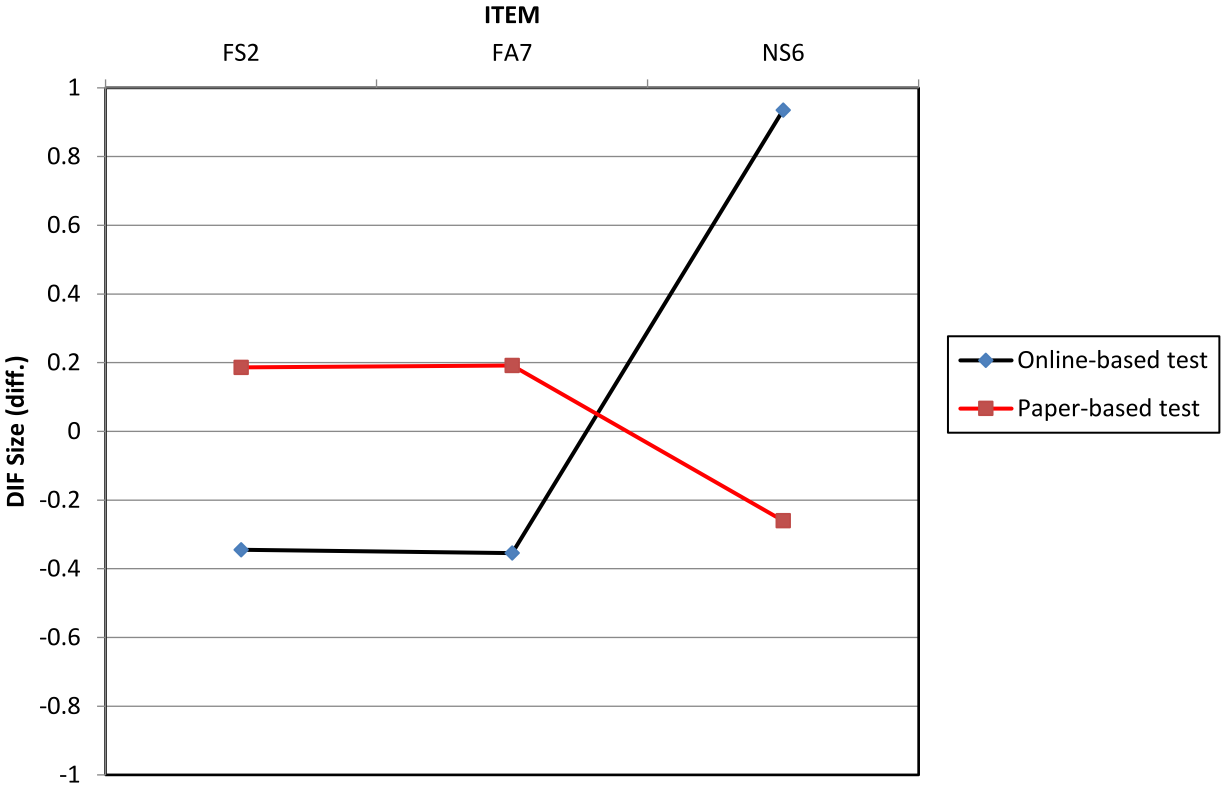

Figure 8. DIF analysis based on the test method

The correlation matrix for all the tasks and the whole test were also evaluated. All correlation values were significant and ranged from 0.16 to 0.76. While the highest correlation was found between the FA task and inductive reasoning test (r = 0.76), the lowest correlation was revealed between the FS and NS tasks, even though the latter relationship was positively significant. This finding implied that students with a higher score on a task would achieve a higher score on the inductive reasoning test. The students' abilities and correlations between the inductive reasoning test and tasks are summarized in Table 4.

Table 4. Result of student abilities and correlation based on inductive reasoning test and tasks

| Test-subscale | M (logits) | SD | Logit range (Min, Max) | Pearson correlation | | | |
|---|---|---|---|---|---|---|---|
| | | | | FA | FS | NA | NS |
| FA | 1.16 | 0.8 | (-2.58, 3.97) | | | | |
| FS | 0.98 | 1.01 | (-2.72, 4.31) | .45** | | | |
| NA | -0.04 | 0.78 | (-2.76, 4.05) | .36** | .24** | | |
| NS | -1.41 | 0.98 | (-4.25, 4.30) | .17** | .16** | .45** | |
| IR test | 0.24 | 0.79 | (-5.41, 1.69) | .76** | .68** | .74** | .56* |

*Note. N = 856 *p < .05, **p < .001, M = Mean, SD = Standard deviation, IR = Inductive reasoning, FA = Figural analogies, FS = Figural series, NA = Number analogies, NS = Number series*

The students' inductive reasoning abilities were also evaluated in accordance with gender and grade. An examination reveals that undergraduate students outperformed students in other grades; M = 0.59; SD = 0.63. The 12th grade students had higher logit values (M = 0.31; SD = 0.66) than the 10th and 11th grade students. Surprisingly, the 10th and 11th graders had the same logit values. Furthermore, the female students had superior performances (M = 0.28; SD = 0.88) in solving inductive reasoning problems in comparison to the male students.

To depict the primary trend between gender and grade related to the development of student inductive reasoning, graphical packages such as the yarrr package (Phillips, 2017) and the ggplot2 package (Wickham, 2016) were employed by using R software to create a pirate plot that combined the boxplot and student logit value distribution in Figure 9.

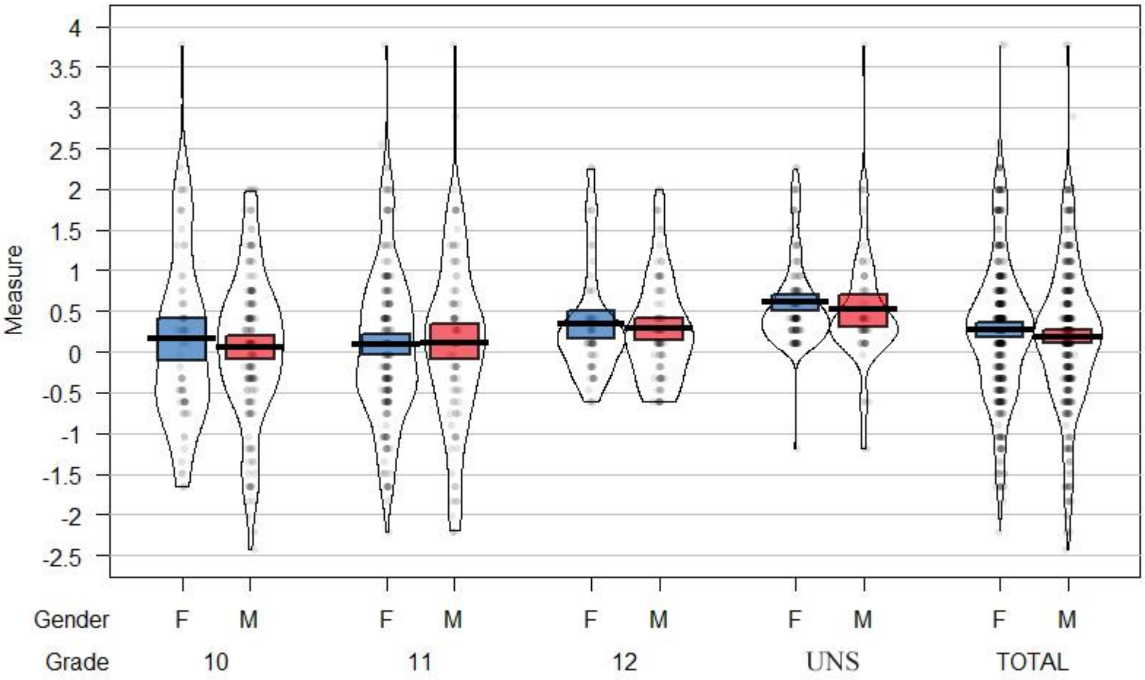

Figure 1. Pirate plot for comparing student measure (logit) based on gender and grade

In evaluating the difficulties of the inductive reasoning items, The classification in accordance with the LVI analysis is displayed in Table 5.

Table 5. The categorisation of inductive reasoning items difficulties

| Task | Difficulty level I, LVI ≥ Mean logit + 2SD | Difficulty level II, Mean logit + 2SD > LVI ≥ 1SD | Difficulty level III, 1SD > LVI ≥ Mean logit | Difficulty level IV, Mean logit > LVI ≥ -1SD | Difficulty level V, LVI < -1SD |
|---|---|---|---|---|---|
| FA | | | FA4, FA5, | FA1, FA2, FA6, FA7, FA8, FA9, FA10 | FA3 |
| FS | | | FS9 | FS1, FS2,FS5, FS6, FS7, FS8, FS10 | FS3, FS4 |
| NA | | NA6 | NA5,NA7, NA8, NA9, NA10 | NA1, NA2, NA3, NA4 | |
| NS | NS6, NS8 | NS5, NS7, NS9, NS10 | | NS1, NS2, NS3, NS4 | |

The results of LVP analysis, which resulted in four categories in relation to gender and grade are also presented in Table 6.

Table 6. The categorisation of student inductive reasoning abilities

| Demographics | Very high, LVP > Mean Logit + 2SD | High, Mean Logit + 2SD ≥ LVP > Mean Logit | Moderate, Mean Logit ≥ LVP > Mean Logit - 2SD | Low, LVP < Mean Logit - 2SD |
|---|---|---|---|---|
| Gender | | | | |
| Female | 10 | 237 | 191 | 10 |
| Male | 3 | 202 | 184 | 19 |
| Total | 13 | 439 | 375 | 29 |
| | | | | |
| 10th grade | 2 | 102 | 114 | 13 |
| 11th grade | 6 | 122 | 147 | 16 |
| 12th grade | 2 | 74 | 77 | 0 |
| Undergraduate student | 3 | 141 | 37 | 0 |
| Total | 13 | 439 | 375 | 29 |

## Conclusions, recommendations, limitations

This dissertation includes two cross-sectional studies from pilot and main study with five published studies, one systematic literature review and four empirical study. In the first study in Chapter 2 is systematic review that was conducted on how often students have misconceptions about science was used to inform some findings. These findings included the different instruments used to find these misconceptions, the subjects on which students frequently have misconceptions, and the benefits and drawbacks of each test instrument. Some test instruments are used in combination to generate insightful results that can be used to support accurate interpretations of student misconceptions. Both written and oral tools have benefits and drawbacks. The technique of analysis can be strengthened by performing an integrated combination and by removing any flaws in a single instrument. Most researchers prefer simple multiple-choice tests (32.23%) and multiple tier tests (33.06%). According to study 1, researchers discovered that biology, chemistry, and physics subjects frequently lead to misconceptions among students. Biology had 15 concepts, chemistry had 12 concepts, and physics had 33 concepts. The systematic review provided evidence that the nature of misconception is resistant and tenacious to change, which poses a challenge for the advancement of scientific knowledge in the future. Those who wish to conduct study or teach with these tools must take great care to employ the appropriate techniques. Study 1 recommends three main steps before conducting research on misconceptions, including (1) examining the idea that typically causes misconceptions in students, (2) selecting a diagnostic tool based on benefits and drawbacks, (3) using combination two or more instrument to enhance research quality.

After conducting systematic literature review, the investigation of instrument validity and reliability was measure in first empirical study (Study 1) as pilot study, and study 2 as main study with larger sample size performed to invest item difficulty pattern. Student misconceptions in science evaluation is presented in Study 3. Pilot study in study 2 confirmed that all the items in the developed instrument are valid and reliable covering student ability based on item-person. The 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. Study 2 explores Evaluating item difficulty patterns for assessing student misconceptions in science across science subjects with larger sample size. Study 2 confirms that all items in the developed two-tier multiple choices diagnostic test meet the valid and reliable criteria. The item difficulty level of items on various science concepts is not universally based on science topics, but they are connected or similar across science disciplines, especially in physics, biology, and chemistry. Researchers also found items in the science concept may have different difficulty levels based on gender and grade. An empirical study of students' misconception in science was presented in Study 3. Study 3 confirmed significant differences in student conception mean scores between all cohorts; however, post-hoc analysis for ANOVA 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.

Lastly, Study 4 informed the findings in assessing student inductive reasoning comprehensively using Rasch measurement approach. The adapted inductive reasoning test was shown to be valid and reliable in Indonesia and other countries, thus indicating this instrument can be employed in a wide range of cultural contexts. The items in the test are free of bias and only

NS6 had a moderate to large DIF. Even though females outperformed males in relation to inductive reasoning abilities, no significant gender differences were found among the grades. Significant differences were found among all the groups, with the exception of the 10th and 11th grades. The classification of the difficulty of items revealed a wide range of difficulty levels, where numeric items were more difficult than figural items. Most of the students were classified as having high or moderate abilities. in general, findings in this study provided initial information related to Indonesian students' inductive reasoning ability.

## Educational implication

. The results from this research can be used as foundation to develop student misconception in science and inductive reasoning in Indonesian curriculum whereby misconception tests can be used to evaluate student understanding, and the inductive reasoning test was often used for entrance test in higher education level and job carrier.

## Recommendations

General recommendations based on series empirical studies in this dissertation presented as below:

1. Teachers or educators have to aware what kind of topics distributing misconceptions in science subject. Therefore, they can improve the student understanding about science concept and science achievement.
2. Screening for student understanding in the end of learning activity was needed using proper instrument, we recommended teachers can use the two-tier multiple choice diagnostics test to identify student knowledge and reasoning in a particular science concept.
3. For future researchers, pilot study as study 2 need to conducted before main study in study 3 and study 4 to confirm instrument validity and reliability in instrument development stage.
4. Future researchers can map the overall item difficulty level of whole science concepts.
5. Time series data collection or longitudinal research design must be added to explore whether there is a change of item difficulty level with the racking method in the Rasch measurement. Racking analysis allows researchers to evaluate whether there is a change in the difficulty level of the item on the different testing times sequentially.
6. The investigation of the relations between students' science misconceptions and thinking skills such as inductive reasoning and science reasoning is needed using The complex model using Structural Equation Modelling (SEM), not only assessing separately.
7. Future studies to mapping students' inductive reasoning needs to conduct using a longitudinal research design and include mixed methods.

## Limitations

Some limitations based on series empirical studies in this dissertation presented as below:

1. Researchers did not develop items based on all scientific concepts studied in Indonesia. Items selected are based on concepts that distribute misconceptions in the previous research (Allen, 2014; American Association for the Advancement of Science (AAAS), 2012; Csapó, 1998; Soeharto, Csapó, et al., 2019).
2. All respondents were from West Kalimantan, one of the provinces in Indonesia, one must exercise caution in generalizing the results to all Indonesian students, although the Rasch analysis have demonstrated that the samples hold local independence.
3. Studies in this dissertation performed quantitative analysis only; a mix of quantitative and qualitative methods may provide more meaningful insights.

## References

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

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

Badaruddin, K., & Hawi, A. (2022). Assessment of Student Attitudes in the 2013 Curriculum: Its Implementation and Problems. *Webology*, *19*(1), 6408–6419.

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

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

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

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*, 399–412. https://doi.org/10.1111/bjop.12046

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

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.

Hasan, S. H. (2013). History Education in Curriculum 2013: A New Approach To Teaching History. *Historia: Jurnal Pendidik Dan Peneliti Sejarah*, *14*(1), 163. https://doi.org/10.17509/historia.v14i1.2023

Istikomah, F., Rochmad, R., & Winarti, E. R. (2017). Analysis of 7th Grade Students' Inductive Reasoning Skill in PBL-Bertema Model Towards Responsibility Character. *Unnes Journal of Mathematics Education*, *6*(3), 345–351. https://doi.org/10.15294/ujme.v6i3.17600

Keeley, P. (2012). Misunderstanding misconceptions. *Science Scope*, *35*(8), 12–13.

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*.

Leedy, P. D., & Ormrod, J. E. (2005). *Practical research: Planning and design* (Vol. 1). Pearson.

Linacre, John M. (2021). *Winsteps® Rasch measurement computer program User's Guide*. Winsteps.com.

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.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

Mubarokah, F. D., Mulyani, S., & Indriyanti, N. Y. (2018). Identifying students' misconceptions of acid-base concepts using a three-tier diagnostic test: A case of Indonesia and Thailand. *Journal of Turkish Science Education*, *15*(Special Issue), 51–58. https://doi.org/10.12973/tused.10256a

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

Phillips, N. D. (2017). Yarrr! The pirate's guide to R. In *APS Observer* (Vol. 30, Issue 3).

Prastowo, A., & Fitriyaningsih, F. (2020). Learning Material Changes as the Impact of the 2013 Curriculum Policy for the Primary School/Madrasah Ibtidaiyah. *Edukasia : Jurnal Penelitian Pendidikan Islam*, *15*(2), 251. https://doi.org/10.21043/edukasia.v15i2.7947

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. https://doi.org/10.36681/tused.2021.56

Siswono, T. Y. E., Hartono, S., & Kohar, A. W. (2020). Deductive or Inductive? Prospective Teachers' Preference of Proof Method on An Intermediate Proof Task. *Journal on Mathematics Education*, *11*(3), 417–438. https://doi.org/10.22342/jme.11.3.11846.417-438

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. (2021). Evaluation and development of students' misconception using diagnostic assessment in science across school grades: A Rasch measurement approach. *Journal of Turkish Science Education*, *18*, 351–370. https://doi.org/10.36681/tused.2021.78

Soeharto, S., & Csapó, B. (2021). Evaluating item difficulty patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts. *Heliyon*, *7*(11), e08352. https://doi.org/10.1016/j.heliyon.2021.e08352

Soeharto, S., & Csapó, B. (2022a). Exploring Indonesian student misconceptions in science concepts. *Heliyon*, *8*(9), e10720. https://doi.org/10.1016/j.heliyon.2022.e10720

Soeharto, S., & Csapó, B. (2022b). Assessing Indonesian student inductive reasoning: Rasch analysis. *Thinking Skills and Creativity*, *46*(November 2021), 101132. https://doi.org/10.1016/j.tsc.2022.101132

Soeharto, S., & Csapó, B. (2022c). Assessing Indonesian student inductive reasoning: Rasch analysis. *Thinking Skills and Creativity*, *46*, 101132. https://doi.org/10.1016/j.tsc.2022.101132

Soeharto, S., Csapő, 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). https://doi.org/10.15294/jpii.v8i2.18649

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.

Sukarelawan, M. I., Jumadi, J., Kuswanto, H., Soeharto, S., & Hikmah, F. N. (2021). Rasch Analysis to Evaluate the Psychometric Properties of Junior Metacognitive Awareness Inventory in the Indonesian Context. *Jurnal Pendidikan IPA Indonesia*, *10*(4), Article 4. https://doi.org/10.15294/jpii.v10i4.27114

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.

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.

Wickham, H. (2016). Data analysis. In *Ggplot2* (pp. 189–201). Springer.

Yin, R. K. (2018). *Case study research and applications: Design and methods*. Sage Books.

## Related publications in dissertation

Soeharto, S. (2021). Evaluation and development of students' misconception using diagnostic assessment in science across school grades: A Rasch measurement approach. Journal of Turkish Science Education, 18(3), 351-370. https://doi.org/10.36681/tused.2021.78

Soeharto, S., & Csapó, B. (2021). Evaluating item difficulty patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts. Heliyon, 7(11), e08352. https://doi.org/10.1016/j.heliyon.2021.e08352

Soeharto, S., & Csapó, B. (2022a). Exploring Indonesian student misconceptions in science concepts. Heliyon, 8(9), e10720. https://doi.org/10.1016/j.heliyon.2022.e10720

Soeharto, S., & Csapó, B. (2022b). Assessing Indonesian student inductive reasoning: Rasch analysis. Thinking Skills and Creativity, 46, 101132. https://doi.org/10.1016/j.tsc.2022.101132

Soeharto, S., Csapő, 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). https://doi.org/10.15294/jpii.v8i2.18649

## Unrelated publications and collaboration during PhD studies

Akhmetova, A., Imambayeva, G., & Csapó, B. (2022). Assessing and validating young Kazakhstanis' reading skills in English, the impact of classroom climate, and their engagement on reading skills. *Indonesian Journal of Applied Linguistics*, *12*(2), 280–292. https://doi.org/10.17509/ijal.v12i2.37321

Martono, Dewantara, J. A., & Soeharto. (2020). The Ability of Indonesian Language Education Students in Designing Lesson Plan through Teaching Practice in School. *Universal Journal of Educational Research*, *8*(11), 5489–5497. https://doi.org/10.13189/ujer.2020.081152

Sarimanah, E., Soeharto, S., Dewi, F. I., & Efendi, R. (2022). Investigating the relationship between students' reading performance, attitudes toward ICT, and economic ability. *Heliyon*, *8*(6), e09794. https://doi.org/10.1016/j.heliyon.2022.e09794

Soeharto, S., & Csapó, B. (2021). Building a House From Lego Blocks: Using Cross Cultural Validation to Develop the Constructed Motivation Questionnaire (CMQS) in Science. *Pedagogika*, *142*(2), Article 2. https://doi.org/10.15823/p.2021.142.12

Sukarelawan, M. I., Jumadi, J., Kuswanto, H., Soeharto, S., & Hikmah, F. N. (2021). Rasch Analysis to Evaluate the Psychometric Properties of Junior Metacognitive Awareness Inventory in the Indonesian Context. *Jurnal Pendidikan IPA Indonesia*, *10*(4), Article 4. https://doi.org/10.15294/jpii.v10i4.27114

## Conference papers

Soeharto, S., & Csapó, B. (2019). Students' Misconceptions and Diagnostic Assessment in Science. In: Varga, Aranka; Andl, Helga; Molnár-Kovács, Zsófia (eds.) Neveléstudomány – Horizontok és dialógusok. Absztraktkötet. : XIX. ONK. Pécs, 2019. november 7-9. Pécs, Hungary : Pécsi Tudományegyetem Bölcsészettudományi Kar Neveléstudományi Intézet, pp. 538-538.

Soeharto, S. (2019). Developing Three-Tier Diagnostic Test In Science To Assess Student Misconceptions In Science. Abstract book: The 13th Training and Practice International Conference on Educational Science. Kaposvár, Hungary: Kaposvár University Faculty of Pedagogy, pp. 202-202.

Soeharto, S. (2021). The Evaluation of Students` Inductive Reasoning and Its Role In Science Achievement Using Rasch Analysis. Abstract book: The 14th Training and Practice International Conference on Educational Science. Kaposvár, Hungary: Kaposvár University Faculty of Pedagogy, pp. 55-55.

Soeharto, S., & Csapó, B. (2021). Psychometric Evaluation in Developing E-Learning Readiness in Science Classroom (ELRSC) Questionnaire Using Rasch Analysis. Abstract book: ATEE-EDITE-ELTE online conference on 11 June: Research in Teacher Education – the next generation. Hungary: Budapest, 48-48.

Soeharto, S., & Csapó, B. (2021). Investigating the relationship between test anxiety and motivation in science learning. Abstract book: in JURE 2021: Education and Citizenship: Learning and Instruction and the Shaping of Futures. Online, pp. 24-24.

Soeharto, S., & Csapó, B. (2021). The diagnostic test evaluation and the student misconception development in science. Abstract book: in EARLI 2021: Education and Citizenship: Learning and Instruction and the Shaping of Futures. Online, pp. 217-217.

Soeharto, S., & Csapó, B. (2021). Investigating students' e-learning readiness in the science classroom. Abstract book: in EAPRIL 2021: the 15th annual EAPRIL Conference for Practitioner Research on Improving Learning. Online, pp. 42-42.

Soeharto, S. (2021). *Evaluating Students' E-Learning Training Needs in Science Classroom during COVID-19 Pandemic Era.* In: Molnár, Győnyvér and Tóth Edith; Dancs, Katinka (eds.) *CES 2021: 21st Conference on Educational Sciences: Education's responses to future challenges*. Programme and Abstracts. Szeged, Hungary: Szegedi Tudományegyetem, pp. 477-477.

Soeharto, S. (2021). *Investigating the Influence of Students' Inductive Reasoning on Science and Mathematics Achievement.* In: Molnár, Győnyvér and Tóth Edith; Dancs, Katinka (eds.) *CES 2021: 21st Conference on Educational Sciences: Education's responses to future challenges*. Programme and Abstracts. Szeged, Hungary: Szegedi Tudományegyetem, pp. 287-287.