Journal of Education and e-Learning Research
Vol. 12, No. 4, 636-647, 2025
ISSN(E) 2410-9991 / ISSN(P) 2518-0169
DOI: 10.20448/jeelr.v12i4.7849

Analysis of measurement of metacognitive skills and students' self-efficacy using the Rasch model

Novi Indah Earlyanti[1]
Amri Sandy[2]

[1]*Sekolah Tinggi Ilmu Kepolisian, Jakarta, Indonesia.*
Email: noviindahearlyanti@stik-ptik.ac.id
[2]*Universitas Matana, Tangerang Selatan, Indonesia.*
Email: amri.sandy@matanauniversity.ac.id

(Corresponding Author)

Abstract

This study aimed to develop an instrument to measure students' metacognitive skills and self-confidence, focusing on Statistics and Data Science courses at a Police Science College in Jakarta, Indonesia. Using a quantitative approach, the population included all students enrolled in the 2024–2025 academic year from three study programs. A total of 142 students were selected through random group sampling, with 45 students representing each program. Data were collected via a Google Form containing 25 multiple-choice items for metacognitive assessment and 25 Likert-scale statements for self-confidence measurement. Data analysis used Winstep, Jamovi, and ConQuest software, based on the Rasch model's assumptions: unidimensionality, local independence, parameter invariance, and log-linear properties. The findings showed that 73.7% of participants had intermediate to high ability levels, with raw scores between 21–24. No extreme scores (0 or 25) were found, suggesting a relatively homogeneous ability distribution. In contrast, self-confidence levels were higher across the sample. The combined average logit score for metacognitive skills and self-confidence was +3.08, indicating high performance and confidence among students. These results support the reliability of the developed instrument and provide insight into students' cognitive and affective readiness in quantitative courses within police science education.

Keywords: Instrument development, Measurement validity, Metacognitive skills, Police science education, Rasch model, Self-confidence.

Citation | Earlyanti, N. I., & Sandy, A. (2025). Analysis of measurement of metacognitive skills and students' self-efficacy using the Rasch model. *Journal of Education and E-Learning Research, 12*(4), 636–647. 10.20448/jeelr.v12i4.7849
History:
Received: 29 April 2025
Revised: 31 October 2025
Accepted: 28 November 2025
Published: 11 December 2025

Publisher: Asian Online Journal Publishing Group

Funding: This study received no specific financial support.
Institutional Review Board Statement: The Ethical Committee of the Sekolah Tinggi Ilmu Kepolisian, Indonesia has granted approval for this study on 21 June 2025 (Ref. No. 2025/05/21/1).
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: Both authors contributed equally to the conception and design of the study. Both authors have read and agreed to the published version of the manuscript.

Contents

Contribution of this paper to the literature

This study contributes to the literature on metacognitive skills and self-confidence. It also developed an instrument to measure students' metacognitive skills and self-confidence.

1. Introduction

Students with the skills to manage their knowledge tend to have higher self-confidence (Dwilaksana, 2020; Ferdian, Anwar, & Herdini, 2025; Sukmawati, Qadar, & Sulaeman, 2022). This is certainly beneficial in the long term for students, especially Police Science College students as future officers. For example, they are not easily stressed, think critically, evaluate situations objectively, create new solutions in solving complex problems, and are independent in developing appropriate and effective learning strategies. One form of formal learning or self-development outside the academic environment is managing, organizing, and evaluating knowledge well (Flavell, 1979; Schraw & Dennison, 1994).

Metacognitive skills, as part of the higher-order thinking skills (HOTS), involve conceptual abilities, organizing knowledge, and experience in decision-making regulation (Anderson & Krathwohl, 2001; Bloom, 1956; Zohar & Dori, 2003). One of the basic courses that is very important to study is Statistics and Data Science because it is the fundamental material for analyzing both basic and advanced research, including planning, developing theoretical frameworks, designing research instruments, collecting data, describing data, and analyzing the results of a study to support decision-making.

Therefore, integrating metacognitive skills and self-confidence into the syllabus of statistics courses, through both theory and practice, is an important reason to study. This approach not only helps students understand the steps involved in solving problems based on data but also instills procedures and work sequences for problem-solving and is accompanied by self-confidence in opening up open discussion spaces (Bandura, 1997; Ningsih, Ramalis, & Purwana, 2018).

As students who are educated to solve problems quickly and accurately, Police Science College students generally have sufficient fieldwork experience. This experience enriches their ability to analyze and solve problems in critical situations. Therefore, their ability and potential to solve problems effectively and efficiently are adequate; they are skilled in identifying, compiling, and simplifying problems in the form of tables, graphs, and images that are easy to understand, so they can be described and analyzed with confidence in delivering convincing data analysis results (Iswahyudi, 2012).

As a basic course, Statistics and Data Science play a role in supporting logical skills and the development of research data management, design, and analysis at both basic and advanced levels. Students with good metacognitive skills tend to organize problem-solving strategies better, identify errors, and improve their academic performance (Hattie & Timperley, 2007; Isfiani & Ekanara, 2022).

Some previous studies related to instruments in the education field have been conducted by Rafik and Sari (2023); Ilman, Putri, Julita, Ristanto, and Isfaeni (2024); Anriani and Gholobi (2024); Jarnudi, Hidayat, Syafrizal, Fahamzah, and Amrudin (2024); Rusdiyanto, Rubini, and Ardianto (2023), and Sugiharto, Cantika, and Werdhani (2023). In reality, and based on previous studies, the measurement quality of instruments for metacognition and self-confidence variables has not been widely assessed, especially in the field of Statistics, where both variables are integrated in the form of tests and non-tests. Another obstacle is the limited ability to evaluate the measurement of examination instruments or tests and other non-tests, such as in classical test theory (CTT) or the true score theory developed by Spearman (1904); DeMars (2010); Retnawati (2014), and Sumintono and Widhiarso (2013), which is formulated as follows.

$$X = T + e \qquad (1)$$

This means that the test taker's ability or raw score (X) can only be measured from the number of correct answers (T) and errors or measurement inaccuracy scores (e) that cannot be avoided in a test. It is clear that this information is too limited to describe X through T. Measurement experts developed a new measurement known as item response theory (IRT), also called modern test theory, which includes the Rasch model, 1PL model (One-parameter logistic model), 2PL model, 3PL model, and 4PL model. The following is the form of the Rasch model equation when adapted into the classical test theory formula.

$$X = P(Y_{pi} = 1|\theta, b) = \frac{1}{1+e^{-(\vartheta p - b_i)}} + \epsilon \qquad (2)$$

It is known that Equation 2 provides more information than Equation 1, with (P(Y)_{pi}=1); the probability of individual p answering a question i correctly. θ_{p}; the ability of individual p (metacognition); and b_{i} the level of difficulty of question i (statistics question); Y_{pi}, the answer score (1 = correct and 0 = incorrect). and ϵ the probability of data errors not explained by the model.

The Rasch model is more relevant in describing individual abilities by calculating the probability of a correct answer for each individual on each question, taking into account the interaction of individual abilities (θp) and the level of difficulty of the questions (b_i) (Van Zile-Tamsen, 2017). In addition, the Rasch model verifies that each question and individual follow the unidimensionality assumption, which is indicated by fit (infit and outfit), helping to evaluate whether the data fits the model or not. The reference criterion for the question item is said to be fit (infit) if it is between the values 0.5 < MNSQ < 1.5, while it is called the standard Z outfit if it is within -2.0 < ZSTD < +2.0 (Sari & Mahmudi, 2024).

The results of the Rasch model are presented on a logit scale, which allows the results to be directly compared across different populations and items, provided the data meet the model assumptions. The Rasch model can also accommodate different research instrument scale formats (e.g., dichotomous or polytomous), offering a broader range of outputs to explain (Wright, 1977, 1994; Yen, 1993). While ordinal data such as the Likert scale, the Rasch model approach uses the Rasch Rating Scale Model (RSM) or Partial Credit Model (PCM).

The results of Rasch analysis allow for fairer scores and local independence for the population because it considers individual abilities according to the difficulty level of the questions on the same scale (Wright, 1977).

$$P(Y_{pi} = \mathrm{k}|\theta, \mathrm{b}, \tau) = \frac{e^{\vartheta p - b_i - \tau_k}}{\sum_{i=0}^{m} e^{\vartheta p - b_i - \tau_k}} + \epsilon \qquad (3)$$

Explanation for per (3), P(Y_pi=k); The probability of individual p giving a response to category k: θp; individual p's ability (self-confidence); bi, the difficulty level of question i; τk, the threshold parameter for category k, which indicates the transition from one category to the next; m, the number of response categories. The PCM model is more flexible than RSM because it allows each question item to have different categories. The formula is similar, but the threshold (τk) is calculated separately for each question item. In other words, for example, question A has categories 1, 2, 3, and 4, and question B only has categories 1 and 2, which can be calculated simultaneously.

The following are the requirements and assumptions for the Rasch model to be fulfilled: (1) Unidimensionality, the instrument analyzed must measure one main ability or construct (for example, statistical metacognitive ability, self-confidence level, or understanding of a particular concept). It can be checked with factor analysis (Exploratory Factor Analysis or Confirmatory Factor Analysis) or with a Unidimensionality Index such as the eigenvalue from principal component analysis (PCA) on residual data. (2) Local Independence, meaning participants' answers to one question item do not depend on their answers to other questions. This can be checked with residual correlation analysis between items. (3) Parameter Invariance, where item difficulty and participant ability (person ability) must be independent of each other. It can be verified by examining fit statistics, namely Infit Mean Square (MNSQ) and Outfit MNSQ. (4) Log-Linear Properties, which involve using logarithmic equations to relate the probability of correct answers to the difference between participant ability (θ) and item difficulty (b). This property explains that the Rasch model provides consistent and mathematically measurable estimates (Sari, Pongsophon, Vongsangnak, Pimthong, & Pitiporntapin, 2022).

2. Research Methodology

This research was conducted at a Police Science College in Jakarta, Indonesia. The population consisted of all Police Science College students in Jakarta, Indonesia; specifically, students of class 82 who took the Statistics Data Science course in 2024. The population was students enrolled in this course, totaling 307 individuals. The sample was selected based on the composition of the study programs (Police Administration, Police Technology, and Police Law), representing 40 students. The test was conducted via Google Forms during the 13th and 14th meetings, assuming students had studied all the materials. The research instrument used multiple-choice questions with a time limit of 120 minutes. The questions covered all syllabuses of the Statistics and Data Science course concerning metacognition indicators, including knowledge, experience, and regulation (Table 1, Table 2, and Table 3).

Table 1. Framework of research instrument for sub-variable knowledge.

Dimension	Subdimension	Topics	Indicator	Sample of item
Knowledge ***Definition:*** *Awareness of knowledge and thought processes*	Declarative knowledge **Definition:** Knowledge of facts, information, or concepts that a person has.	- Measures in statistics - Frequency distribution	(1) Explains the definitions of mean, median, and mode. (2) Identify the types of data distribution (Normal and Surprise data).	What is data centralization size? a. The size used illustrates the variation of the dataset. b. The measure used determines the relationship between the two variables. c. **A measure that describes the average location of the data.** d. A measure that groups data into categories.
	Procedural knowledge **Definition:** Knowledge of how to do things, including problem-solving methods and strategies.	- Measures in statistics - Frequency distribution and statistical table	(1) Calculate the mean, median, and mode values of the data provided. (2) Compile a frequency distribution table from raw data.	The correct procedure for performing a normality test using statistical software is: a. Enter the data, select the t-test menu, and read the p-value. b. **Enter data, select the normality test menu, and read the results of graphs or statistics**. c. Calculate the mean and standard deviation, then make a histogram. d. Select data with outliers, then calculate the z-score value.
	Conditional knowledge **Definition:** Knowledge of when and why to use a particular strategy or method in the right context.	- Measures in statistics - Non-parametric hypotheses & statistics	(1) Choose when to use mean versus median in statistics reports. (2) Compile a frequency distribution table from raw data	For police study researchers, when should primary data be used instead of secondary data to analyze patterns of social conflict? a. When secondary data is available in large quantities. b. **When you want to get specific and up-to-date information from the public.** c. When research time is limited. d. When historical data is needed in the long term

This study aims to apply the Rasch model in evaluating the measurement instrument of students' metacognitive skills and self-confidence in Statistics and Data Science courses at a Police Science College in Jakarta, Indonesia. This study is expected to be useful in developing valid and reliable instruments and to provide new insights into the integration of the creation of Statistics and Data Science questions based on metacognitive variable indicators, testing them with modern test theory. In addition, the results of this study can be applied to the development of similar instruments in other courses, thus supporting the achievement of broader learning objectives and in-depth experiences for students.

However, the combination of metacognition variables and student self-confidence in learning has an important relationship because it indicates that what has been understood is aligned with the correct answer choices. However, not much research has explored these two variables, especially in the context of statistics and data science learning. Another aspect to be achieved is to provide important explanations for educators in designing more effective teaching strategies, designing high-level thinking questions (HOTS), and organizing knowledge experiences in one decision-making regulation based on valid and reliable data.

Table 2. Framework of research instrument for sub-variable experience.

Dimension	Subdimension	Topics	Indicator	Sample of item
Experience **Definition:** Awareness of the ongoing thought process in a given situation.	Reflection on the learning process **Definition:** The ability to reflect on how to think, understand, or complete a task during or after an activity.	- Simple regression - Hypothesis	(1) Evaluate strategies in solving simple regression problems. (2) Identify common errors in creating frequency distribution tables.	Police studies students collected data from two simulation groups. Ex.. A (With the involvement of community representatives), Ke. B (Without involving community representatives). {A, B} = {(85,70), (90, 65), (80, 60), (88, 75), (92,67)}. Hypothesis: H0: There was no difference in the average success rate between groups A and B. H1: There is a difference in the average success rate between groups A and B. The first step to take to test this hypothesis is: a. Calculate the correlation between the time of discussion and the success of conflict resolution. b. **Perform a t-test on two independent samples to compare the mean success.** c. Using regression analysis to predict success based on discussion time. d. Calculate the average score for each group and draw conclusions directly from the data.
	Feelings or intuition towards tasks **Definition:** Emotional awareness or intuitive perception of completing a specific task,	- Multiple regression - Hypothesis	(1) Reveals the difficulty of understanding the concept of multiple regression (2) Describe feelings towards hypothetical concepts	A police member collects data based on the security risk factor of the occurrence of conflicts and provocative actions of the parties involved (Variable X). Decisions are made based on intuition and data analysis by intervening early or waiting (Variable Y). To test the relationship between risk factors and police decisions, the appropriate method is: a. Test t in pairs b. Pearson correlation c. **Simple linear regression analysis.** d. Chi-square test.
	Awareness of the success or failure of the strategy **Definition:** Ability to recognize the strategy used.	- Measures in statistics - Non-parametric hypotheses & statistics	(1) Recognize effective strategies in analyzing data with non-parametric statistics. (2) Experience using data analysis tools.	One investigator found that the mode of certain crime data appeared at a very low level, while the mean and median were in the higher range. What are the implications for the hypothesis analysis being carried out? a. **The low mode indicates that the data are highly variable, so the hypothesis results may be unstable.** b. Focus on mode as it shows the most dominant trend. c. Ignore the mode because the mean and median are more important. d. Change the hypothesis test to focus solely on the mode.

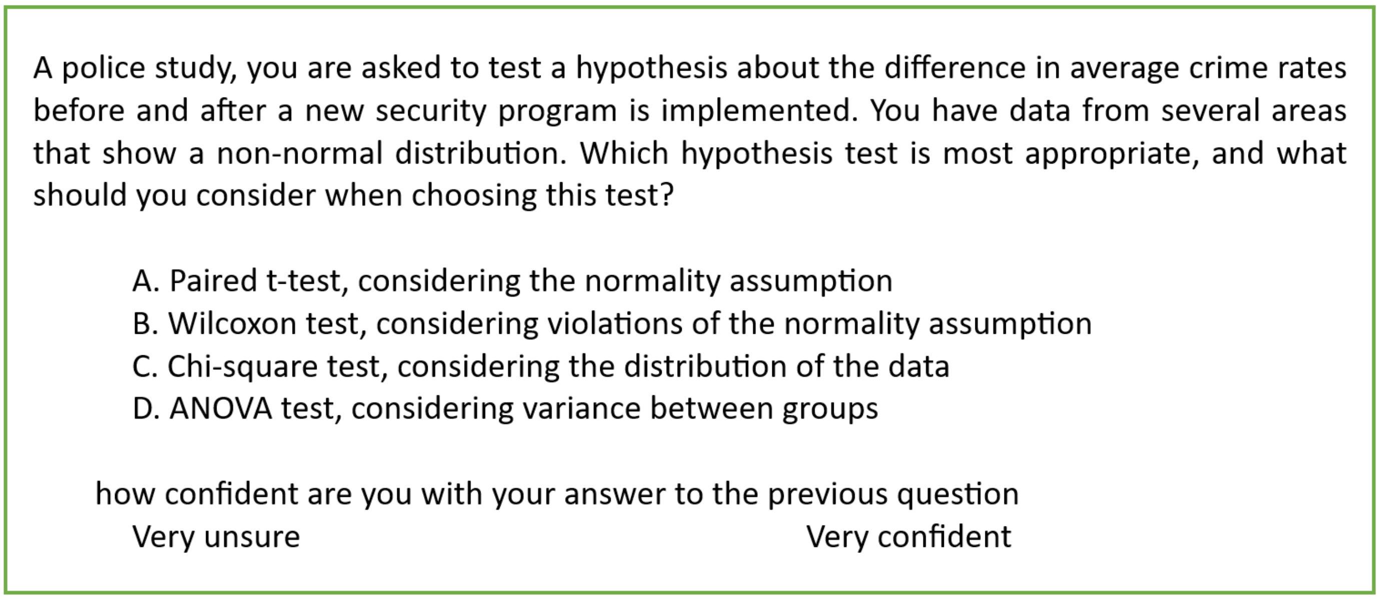
Table 3. Framework of research instruments for sub-variables of regulation.

Dimension	Subdimension	Topics	Indicator	Sample of item
Regulation **Definition:** Awareness to control and regulate thought processes while	*Planning* **Definition:** Setting goals, choosing strategies, and predicting outcomes before committing to tasks	- Simple regression - Parameteric name statistics	(1) Prepare a plan of steps in solving simple regression test questions. (2) Design learning strategies to understand non-parametric statistical concepts.	A police officer decided to measure the severity of the conflict in different regions. He divided the area into three zones (A, B, and C) and recorded the number of conflicts as follows: Zone A: 10, 12, 9 Zone B: 15, 14, 13 Zone C: 8, 7, 10 The officer planned an intervention by calculating the standard deviation of the number of conflicts in Zone B. What is the standard deviation?

Dimension	Subdimension	Topics	Indicator	Sample of item
studying or working				a. 0.47 b. 1.00 **c. 1.15** d. 1.24
	Monitoring **Definition:** Keeping an eye on progress during the learning process or completing tasks to ensure the right strategy.	- Multiple regression - Hypothesis	(1) Monitor the suitability of the steps in calculating the frequency distribution. (2) Evaluate whether the hypothesis testing carried out is in accordance with the data scenario.	A police officer monitors the effectiveness of conflict resolution strategies in the mining area. After the strategy was implemented for three months, he obtained the following data. Month 1: 20 conflict incidents Month 2: 15 conflict incidents Month 3: 10 conflict incidents The officer wanted to ensure that the downward trend in the number of incidents indicated the success of the strategy. To monitor this trend, the most appropriate analysis methods used are: a. Calculate the average number of conflicts per month. b. **Use a line chart to visualize the change in the number of conflicts each month.** c. Compare the conflict data of this region with the region without intervention. d. Using hypothesis tests to determine the success of the strategy.
	Evaluation **Definition:** Assess the results after the task is completed to determine success and identify necessary improvements.	- Frequency distribution - Multiple regression - Introduction to data science	(1) Assess whether the multiple regression calculation strategy provides correct and efficient results. (2) Analyze learning outcomes on the concept of descriptive science and statistical data, and design improvement steps if necessary.	A police chief has completed a training program to improve police members' negotiating skills in handling social conflicts. To evaluate the success of the program, officers used data on the success of negotiations before and after training: Before training: 60% of negotiations are successful. After training: 85% of negotiations are successful. The most appropriate statistical method to evaluate whether this increase is significant is: a. Test z for the proportion of two samples. b. **Test two independent samples.** c. Chi-square test for frequency distribution. d. Paired t-test.

2.1. Research Instrument

The measurement of the variables studied includes a discussion of Statistics and Data Science courses covering measurements in statistics, frequency distribution and statistical tables, hypothesis testing, simple regression, multiple regression, non-parametric statistics, and an introduction to data science. The questionnaire consists of two parts: (1) Metacognitive Skills Scale (True/False), and (2) Belief/Self-Confidence Scale (Likert). Declarative knowledge: 5 questions; procedural knowledge: 5 questions; conditional knowledge: 4 questions; experience: 6 questions (2 reflections, 2 intuitions, 2 success/failure strategies); regulation: 5 questions (1 planning, 2 monitoring, 2 evaluation). The Figure 1 illustrates an example of filling out a questionnaire in a Google form.



A police study, you are asked to test a hypothesis about the difference in average crime rates before and after a new security program is implemented. You have data from several areas that show a non-normal distribution. Which hypothesis test is most appropriate, and what should you consider when choosing this test?

A. Paired t-test, considering the normality assumption
B. Wilcoxon test, considering violations of the normality assumption
C. Chi-square test, considering the distribution of the data
D. ANOVA test, considering variance between groups

how confident are you with your answer to the previous question
Very unsure Very confident

Figure 1. Example of a quotation question from Google Form.

After the data from the Google form is saved in the Excel program, it is then analyzed using the Rasch model with Jamovi and Winsteps software. The metacognition and self-confidence skills test results are tabulated and cleaned using Excel, then saved and reanalyzed using the Jamovi program. The next data analysis process follows the four conditions and assumptions of the fulfillment of the Rasch model that have been explained previously (Unidimensionality, Local Independence, Parameter Invariance, and Loglinear Properties).

3. Results and Discussion

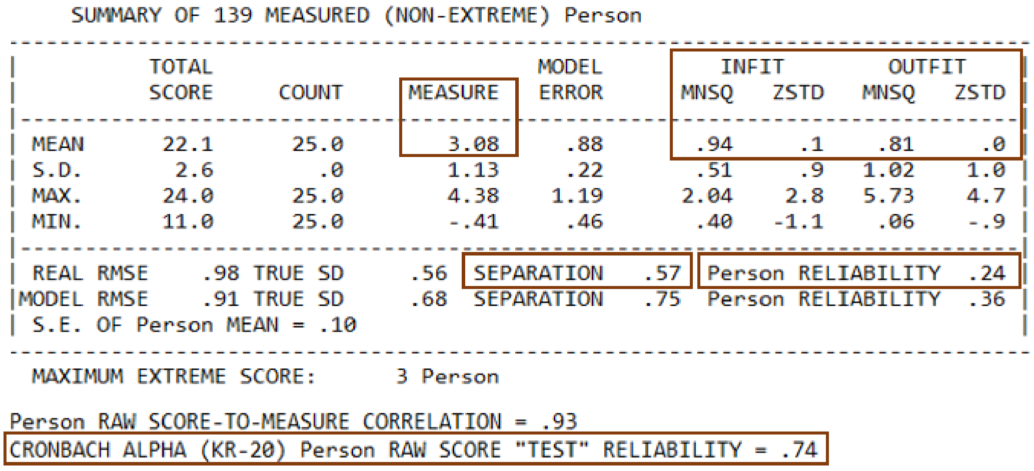
3.1. Metacognition

3.1.1. Item Reliability Test (Model Fit)

The results of the Rasch model analysis of the research instrument using Jamovi and Winsteps produced the same fit model; only the information from the Jamovi program was too simple, so the results from the Winsteps program will be displayed here. The summary statistics from Winsteps (see Figure 2, Figure 3, and Figure 4).

Based on Figure 2 and Figure 3, it is known that there are three respondents (students) who have perfect scores (EXTREME Person), so they were excluded from the analysis process. Furthermore, the average score of respondents in the Statistics and Data Science (Measure) course is 3.08 logits. As is known, the average logit value of more than 0.0 indicates that the respondent will answer more questions with a form similar to the question.

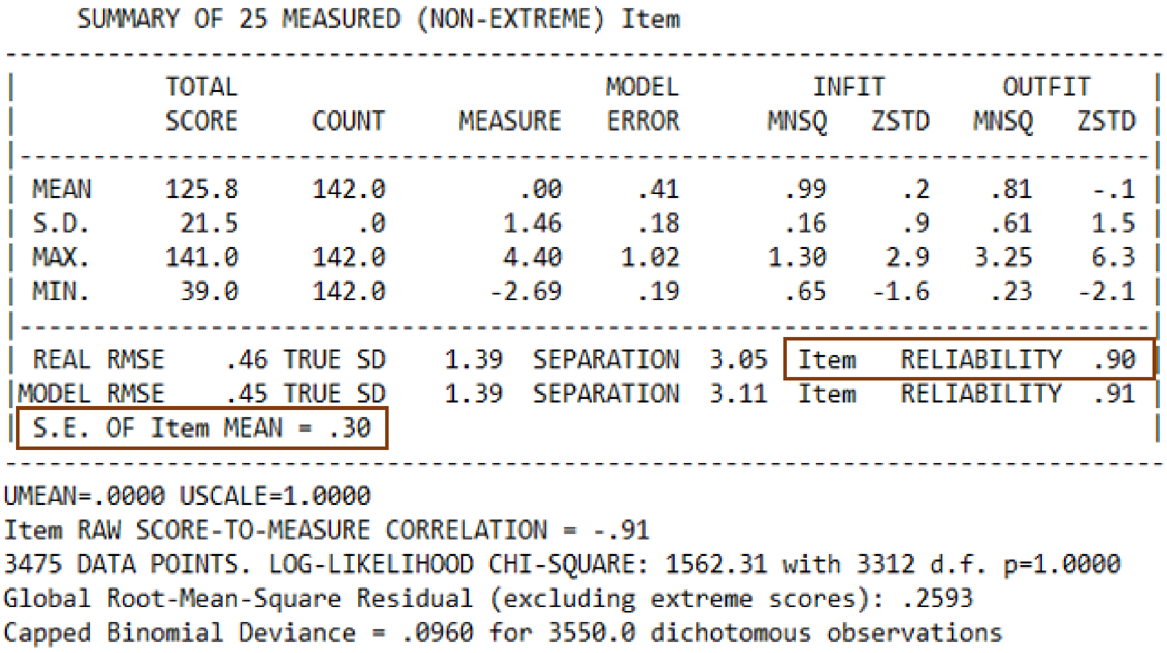
Furthermore, Cronbach's alpha value of 0.74 indicates that the overall interaction between person and item is very good, with the reference value meeting the reliable requirements above 0.7, which is categorized as good (Sumintono & Widhiarso, 2013). On the other hand, with the person reliability value of 0.24 and item reliability of 0.90, it can be concluded that the consistency of the respondents' answers is weak. However, the quality of the questions in the instrument is very good.



```
     SUMMARY OF 139 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|           TOTAL                         MODEL         INFIT        OUTFIT    |
|           SCORE      COUNT    MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       22.1       25.0       3.08     .88       .94     .1    .81     .0 |
| S.D.        2.6         .0       1.13     .22       .51     .9   1.02    1.0 |
| MAX.       24.0       25.0       4.38    1.19      2.04    2.8   5.73    4.7 |
| MIN.       11.0       25.0       -.41     .46       .40   -1.1    .06    -.9 |
|-----------------------------------------------------------------------------|
| REAL RMSE     .98 TRUE SD     .56  SEPARATION   .57  Person RELIABILITY  .24 |
|MODEL RMSE     .91 TRUE SD     .68  SEPARATION   .75  Person RELIABILITY  .36 |
| S.E. OF Person MEAN = .10                                                   |
-------------------------------------------------------------------------------
  MAXIMUM EXTREME SCORE:       3 Person

Person RAW SCORE-TO-MEASURE CORRELATION = .93
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .74
```

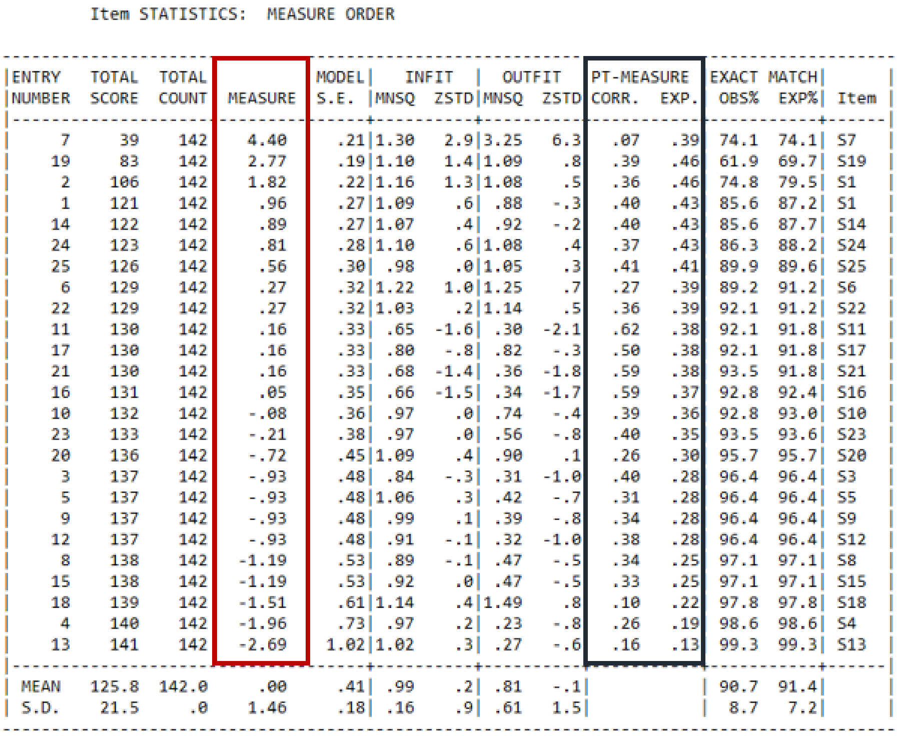
Figure 2. Summary statistics of person ability.

Based on the opinion of Boone, Staver, and Yale (2014), The test items can be considered valid if they meet three requirements: (1) Infit data is between the values 0.5 < MNSQ < 1.5. (2) Standard Z outfit data is between -2.0 < ZSTD < +2.0. (3) The point-measure correlation value is not negative and falls between 0.4 < PT Measure Corr < 0.85 (Sari & Mahmudi, 2024). Based on Figure 2 and Figure 3, the requirements for valid items are met according to the first and second requirements, although the third requirement has not all been met (see Figure 4). However, this strengthens the previous evidence that the consistency of respondents' answers is weak; in other words, some respondents answered by guessing. Referring to this information, it is quite useful or reasonable to improve test items with multiple meanings or interpretations (ambiguous and confusing for respondents).



```
     SUMMARY OF 25 MEASURED (NON-EXTREME) Item
-------------------------------------------------------------------------------
|           TOTAL                         MODEL         INFIT        OUTFIT    |
|           SCORE      COUNT    MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      125.8      142.0        .00     .41       .99     .2    .81    -.1 |
| S.D.       21.5         .0       1.46     .18       .16     .9    .61    1.5 |
| MAX.      141.0      142.0       4.40    1.02      1.30    2.9   3.25    6.3 |
| MIN.       39.0      142.0      -2.69     .19       .65   -1.6    .23   -2.1 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .46 TRUE SD    1.39  SEPARATION  3.05  Item   RELIABILITY  .90 |
|MODEL RMSE    .45 TRUE SD    1.39  SEPARATION  3.11  Item   RELIABILITY  .91 |
| S.E. OF Item MEAN = .30                                                     |
-------------------------------------------------------------------------------
UMEAN=.0000 USCALE=1.0000
Item RAW SCORE-TO-MEASURE CORRELATION = -.91
3475 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 1562.31 with 3312 d.f. p=1.0000
Global Root-Mean-Square Residual (excluding extreme scores): .2593
Capped Binomial Deviance = .0960 for 3550.0 dichotomous observations
```

Figure 3. Summary statistics of ability items.



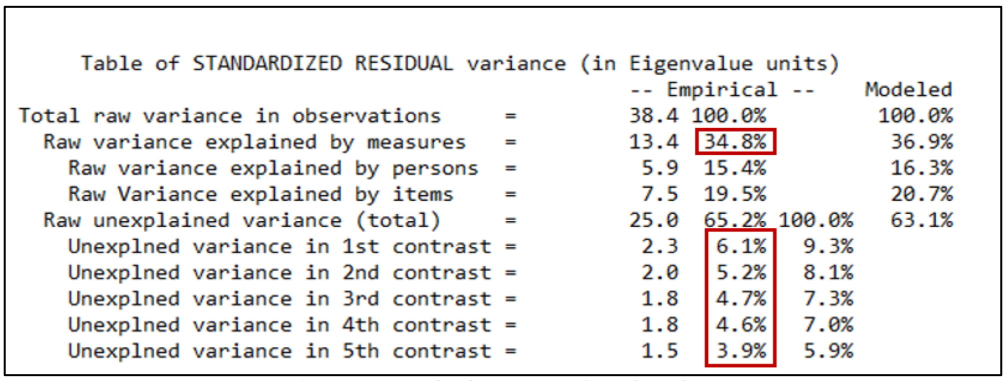
```
Item STATISTICS:  MEASURE ORDER

--------------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT    |  OUTFIT   |PT-MEASURE |EXACT  MATCH|        |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item  |
|------------------------------------+----------+-----------+-----------+------------+-------|
|     7     39    142    4.40     .21|1.30   2.9|3.25   6.3|  .07   .39| 74.1  74.1| S7    |
|    19     83    142    2.77     .19|1.10   1.4|1.09    .8|  .39   .46| 61.9  69.7| S19   |
|     2    106    142    1.82     .22|1.16   1.3|1.08    .5|  .36   .46| 74.8  79.5| S1    |
|     1    121    142     .96     .27|1.09    .6| .88   -.3|  .40   .43| 85.6  87.2| S1    |
|    14    122    142     .89     .27|1.07    .4| .92   -.2|  .40   .43| 85.6  87.7| S14   |
|    24    123    142     .81     .28|1.10    .6|1.08    .4|  .37   .43| 86.3  88.2| S24   |
|    25    126    142     .56     .30| .98    .0|1.05    .3|  .41   .41| 89.9  89.6| S25   |
|     6    129    142     .27     .32|1.22   1.0|1.25    .7|  .27   .39| 89.2  91.2| S6    |
|    22    129    142     .27     .32|1.03    .2|1.14    .5|  .36   .39| 92.1  91.2| S22   |
|    11    130    142     .16     .33| .65  -1.6| .30  -2.1|  .62   .38| 92.1  91.8| S11   |
|    17    130    142     .16     .33| .80   -.8| .82   -.3|  .50   .38| 92.1  91.8| S17   |
|    21    130    142     .16     .33| .68  -1.4| .36  -1.8|  .59   .38| 93.5  91.8| S21   |
|    16    131    142     .05     .35| .66  -1.5| .34  -1.7|  .59   .37| 92.8  92.4| S16   |
|    10    132    142    -.08     .36| .97    .0| .74   -.4|  .39   .36| 92.8  93.0| S10   |
|    23    133    142    -.21     .38| .97    .0| .56   -.8|  .40   .35| 93.5  93.6| S23   |
|    20    136    142    -.72     .45|1.09    .4| .90    .1|  .26   .30| 95.7  95.7| S20   |
|     3    137    142    -.93     .48| .84   -.3| .31  -1.0|  .40   .28| 96.4  96.4| S3    |
|     5    137    142    -.93     .48|1.06    .3| .42   -.7|  .31   .28| 96.4  96.4| S5    |
|     9    137    142    -.93     .48| .99    .1| .39   -.8|  .34   .28| 96.4  96.4| S9    |
|    12    137    142    -.93     .48| .91   -.1| .32  -1.0|  .38   .28| 96.4  96.4| S12   |
|     8    138    142   -1.19     .53| .89   -.1| .47   -.5|  .34   .25| 97.1  97.1| S8    |
|    15    138    142   -1.19     .53| .92    .0| .47   -.5|  .33   .25| 97.1  97.1| S15   |
|    18    139    142   -1.51     .61|1.14    .4|1.49    .8|  .10   .22| 97.8  97.8| S18   |
|     4    140    142   -1.96     .73| .97    .2| .23   -.8|  .26   .19| 98.6  98.6| S4    |
|    13    141    142   -2.69    1.02|1.02    .3| .27   -.6|  .16   .13| 99.3  99.3| S13   |
|------------------------------------+----------+-----------+-----------+------------+-------|
| MEAN   125.8  142.0     .00     .41| .99    .2| .81   -.1|           | 90.7  91.4|       |
| S.D.    21.5     .0    1.46     .18| .16    .9| .61   1.5|           |  8.7   7.2|       |
--------------------------------------------------------------------------------------------------
```

Figure 4. Point measure correlation.

Based on Figure 4, Information on the level of difficulty of the questions (logit) can be found by dividing the questions into two parts: difficult questions and easy questions. Difficult questions (S) include: Question number 7 or S7 (4.40 logit), S19 (2.77 logit), S2 (1.182 logit), S1 (0.96 logit), S14 (0.89 logit), S24 (0.81 logit), S25 (0.56 logit), S6 (0.27 logit), S22 (0.27 logit), S11 (0.16 logit), S17 (0.16 logit), S21 (0.16 logit), and S16 (0.05 logit). Easy questions consist of: S10 (-0.08 logit), S23 (-0.23 logit), S20 (-0.72 logit), S3 (-0.93 logit), S5 (-0.93 logit), S9 (-0.93 logit), S12 (-0.93 logit), S8 (-1.19 logit), S15 (-1.19 logit), S18 (-1.51 logit), S4 (-1.96 logit), and S13 (-2.68 logit). All questions can be considered as measurement tools for Statistics and Data Science metacognition, with some improvements in the sentences or adjustments to the questions based on the reliability information of the questions to prevent misinterpretation.

3.2. Unidimensionality of Question Items

The requirement for unidimensionality of question items refers to the majority of data variance that can explain one main dimension of the research measurement instrument. In other words, the instrument developed can measure what it is intended to measure, in this case, the construct of metacognition in Statistics and Data Science courses. Unidimensionality is based on the Rasch model, using Principal Component Analysis (PCA) of the residuals (Sumintono & Widhiarso, 2013). Figure 5 explains the fulfillment of the data unidimensionality requirement using Winsteps.



```
      Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                    -- Empirical --    Modeled
Total raw variance in observations     =         38.4 100.0%          100.0%
  Raw variance explained by measures   =         13.4  34.8%           36.9%
    Raw variance explained by persons  =          5.9  15.4%           16.3%
    Raw Variance explained by items    =          7.5  19.5%           20.7%
  Raw unexplained variance (total)     =         25.0  65.2% 100.0%    63.1%
    Unexplned variance in 1st contrast =          2.3   6.1%   9.3%
    Unexplned variance in 2nd contrast =          2.0   5.2%   8.1%
    Unexplned variance in 3rd contrast =          1.8   4.7%   7.3%
    Unexplned variance in 4th contrast =          1.8   4.6%   7.0%
    Unexplned variance in 5th contrast =          1.5   3.9%   5.9%
```

Figure 5. Standard variance of residual data.

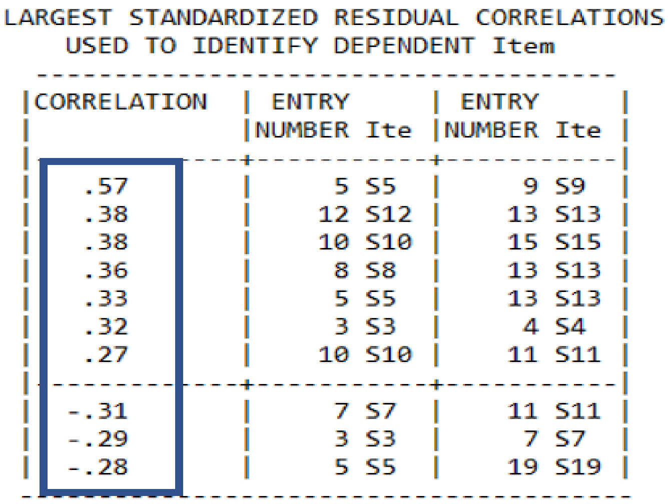
Based on the data Unidimensionality requirement, if the raw variance data measurement result is above 20% (dichotomy), it is considered to have met the Unidimensionality requirement. If the value is above 40% (polytomy), it is better, and above 60% indicates it is special (Sumintono & Widhiarso, 2013). Meanwhile, the data processing results with the Rasch model (Figure 5) show a raw variance of 34.8%, indicating that the data unidimensionality requirement is met. Likewise, for other requirements, the data variance that cannot be explained by the instrument ideally does not exceed 15% (Sumintono & Widhiarso, 2013). Furthermore, it can be explained from Figure 5 that the data variance that the instrument cannot explain is all below 10%. In other words, the data is quite ideal to meet the unidimensionality requirement. So, based on the unidimensionality information, the Rasch model data can assure that the data produced is valid, reliable, and relevant according to the constructs measured by the statistical metacognition and data science variables.

3.3. Local Independence Test of Question Items

Local independence is a component of Rasch model analysis; after accounting for the respondent's ability (or the main construct being measured), the responses to each item must be unconnected or independent of each other, depending solely on individual ability and not on other items that are already known from previous responses (Yen, 1993). If this assumption is violated, the answer to the other item influences the answer to one item.

Figure 6 explains that the local independence required by the Rasch model is around the residual correlation above 0.2 or below -0.2 on the logit scale. So, question numbers S5 and S9 are correlated quite highly with 0.59 logits. A high correlation (both positive and negative) indicates that there is a significant relationship between the residuals of the two items. This relationship is not explained by the main dimensions being measured. Respondents' answers to question number S5 affect the answers to question number S9, or these two items measure very similar aspects (content redundancy).

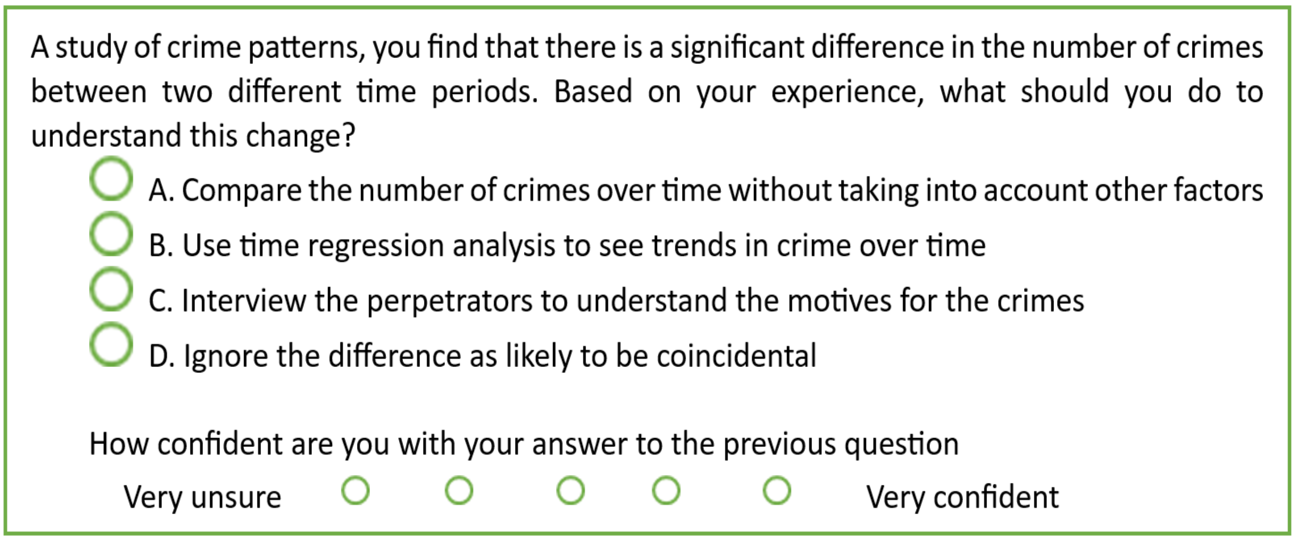
Likewise, the correlation of -0.31 between items S7 and S11 indicates a high negative correlation, suggesting that these two items have an inverse or conflicting relationship. If one item is answered correctly, the other is likely answered incorrectly, which may reflect bias in item construction or indicate that the two items measure conflicting concepts.



```
LARGEST STANDARDIZED RESIDUAL CORRELATIONS
   USED TO IDENTIFY DEPENDENT Item
  -----------------------------------------
 |CORRELATION  | ENTRY      | ENTRY      |
 |             |NUMBER Ite  |NUMBER Ite  |
 |-------------+------------+------------|
 |     .57     |     5 S5   |     9 S9   |
 |     .38     |    12 S12  |    13 S13  |
 |     .38     |    10 S10  |    15 S15  |
 |     .36     |     8 S8   |    13 S13  |
 |     .33     |     5 S5   |    13 S13  |
 |     .32     |     3 S3   |     4 S4   |
 |     .27     |    10 S10  |    11 S11  |
 |-------------+------------+------------|
 |    -.31     |     7 S7   |    11 S11  |
 |    -.29     |     3 S3   |     7 S7   |
 |    -.28     |     5 S5   |    19 S19  |
  -----------------------------------------
```

Figure 6. Local independence test.

To overcome local dependency, namely by fixing the question or deleting one of the questions with the most frequent chance of appearing, such as question numbers S3, S10, and S13. Likewise, there is a high negative correlation with question number S7. Here is an excerpt from question number S5, which can be seen in Figure 7, and the answer quote for question number 5 (S5) can be seen in Figure 8.



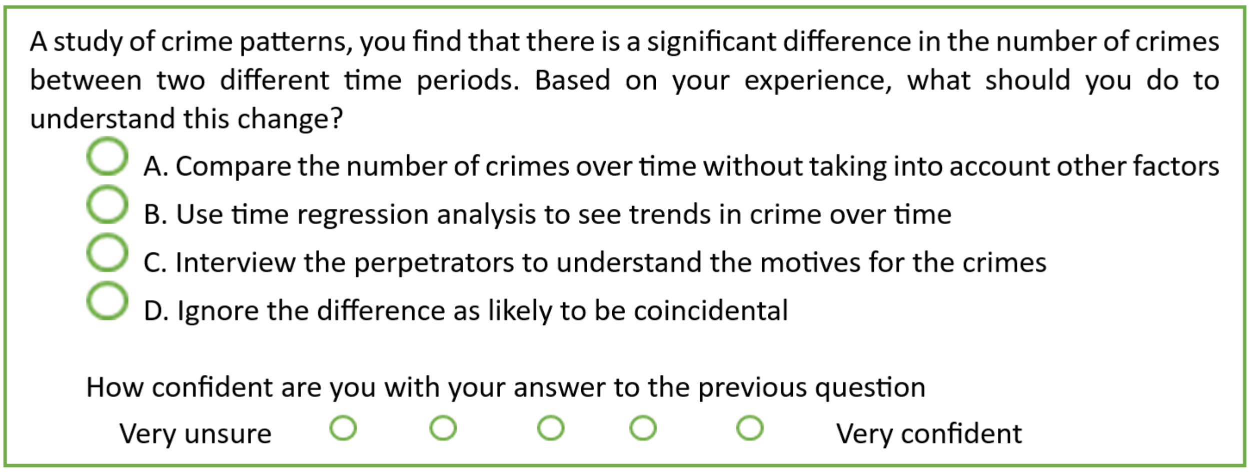
A study of crime patterns, you find that there is a significant difference in the number of crimes between two different time periods. Based on your experience, what should you do to understand this change?

○ A. Compare the number of crimes over time without taking into account other factors
○ B. Use time regression analysis to see trends in crime over time
○ C. Interview the perpetrators to understand the motives for the crimes
○ D. Ignore the difference as likely to be coincidental

How confident are you with your answer to the previous question

Very unsure ○ ○ ○ ○ ○ Very confident

Figure 7. Excerpt from question number 5 (S5) from Google form.



A study of crime patterns, you find that there is a significant difference in the number of crimes between two different time periods. Based on your experience, what should you do to understand this change?

○ A. Compare the number of crimes over time without taking into account other factors
○ B. Use time regression analysis to see trends in crime over time
○ C. Interview the perpetrators to understand the motives for the crimes
○ D. Ignore the difference as likely to be coincidental

How confident are you with your answer to the previous question

Very unsure ○ ○ ○ ○ ○ Very confident

Figure 8. Answer quote for question number 5 (S5) from Google form.

3.4. Item Parameter Invariance

Rasch model parameter invariance is used to explain that the item difficulty parameter remains consistent regardless of the characteristics of the respondent group; in addition, the individual's ability being measured is not influenced by the type of particular item given.

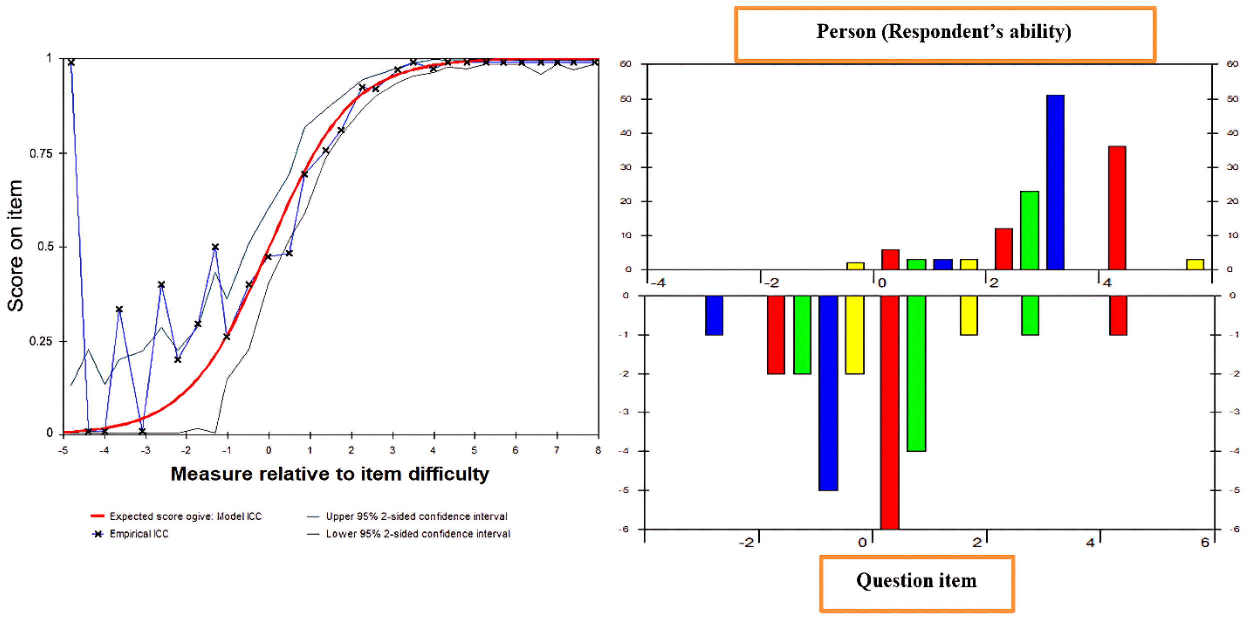
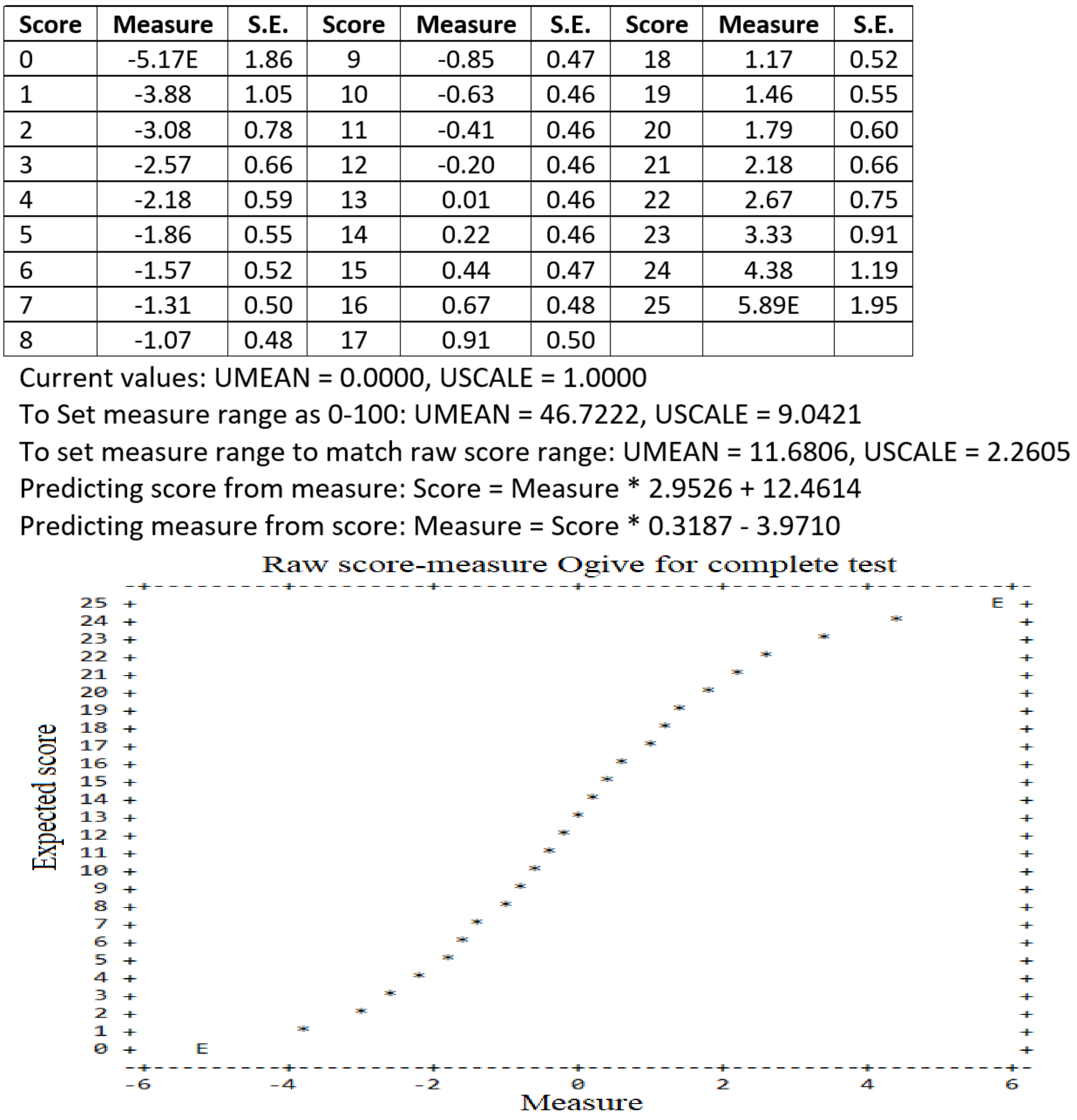


Figure 9. Answer patterns and difficulty levels of question items.

Based on Figure 9, it can be concluded that the metacognition measuring instrument can provide fair and unbiased results for all respondents following the ideal line pattern. The distribution of individual ability (Person ability) and item difficulty level (Item difficulty) shows a logical pattern and does not depend on certain attributes.



Score	Measure	S.E.	Score	Measure	S.E.	Score	Measure	S.E.
0	-5.17E	1.86	9	-0.85	0.47	18	1.17	0.52
1	-3.88	1.05	10	-0.63	0.46	19	1.46	0.55
2	-3.08	0.78	11	-0.41	0.46	20	1.79	0.60
3	-2.57	0.66	12	-0.20	0.46	21	2.18	0.66
4	-2.18	0.59	13	0.01	0.46	22	2.67	0.75
5	-1.86	0.55	14	0.22	0.46	23	3.33	0.91
6	-1.57	0.52	15	0.44	0.47	24	4.38	1.19
7	-1.31	0.50	16	0.67	0.48	25	5.89E	1.95
8	-1.07	0.48	17	0.91	0.50			

Current values: UMEAN = 0.0000, USCALE = 1.0000

To Set measure range as 0-100: UMEAN = 46.7222, USCALE = 9.0421

To set measure range to match raw score range: UMEAN = 11.6806, USCALE = 2.2605

Predicting score from measure: Score = Measure * 2.9526 + 12.4614

Predicting measure from score: Measure = Score * 0.3187 - 3.9710

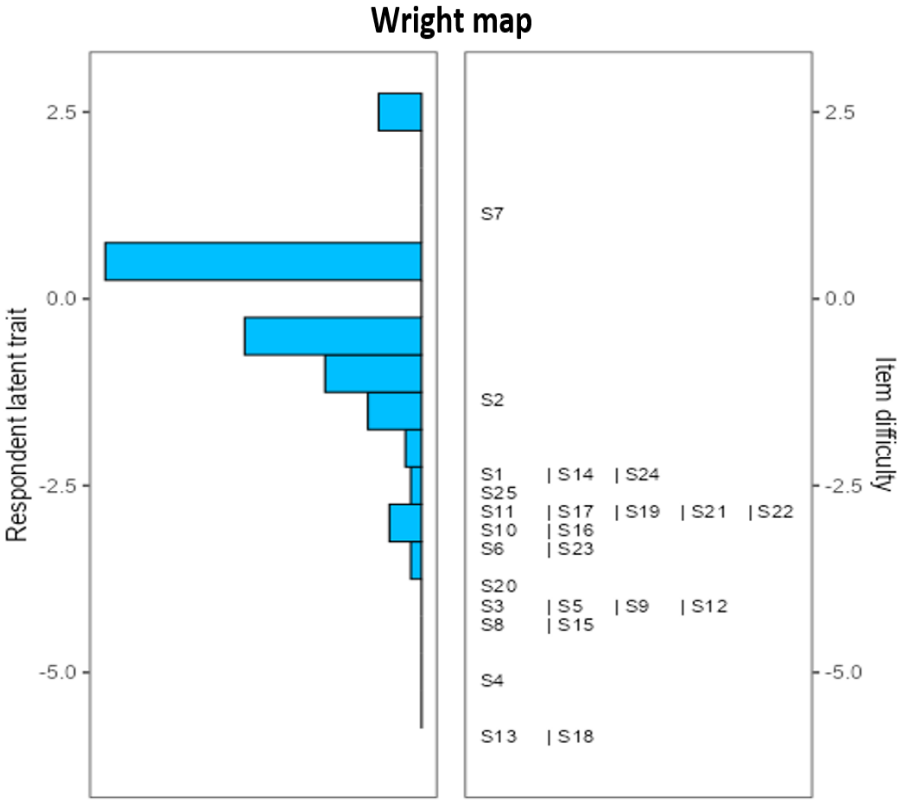


Figure 10. Distribution of test question answers and level of question item difficulty.

Based on Figure 10, the distribution of item scores and measures of participants can be explained according to the Rasch model. Item scores range from 0 to 25, with logit measures ranging from -5.17 (for a score of 0) to +5.89 (for a score of 25). Measures represent the level of participant ability in the Rasch model. Standard errors (S.E.) are larger at both ends of the distribution (extreme scores), indicating that the measure estimates in that area are less stable.

Most participants scored in the middle of the distribution, with scores of 21–24 covering 73.7% of participants. This indicates that most participants possess medium to high levels of ability. Extreme scores (0 or 25) are rare, reflecting the few participants with very low or very high abilities.

The equation shows the linear relationship between raw scores and logit measures:

(1) Predicted score based on measure: Score = Measure × 2.953 + 12.46.

(2) Predicted measure based on score: Measure = score × 0.319 − 3.97.

The ogive graph shows the log-linear relationship between raw scores and logit measures (See Figure 11. This indicates that the Rasch model fits the data.

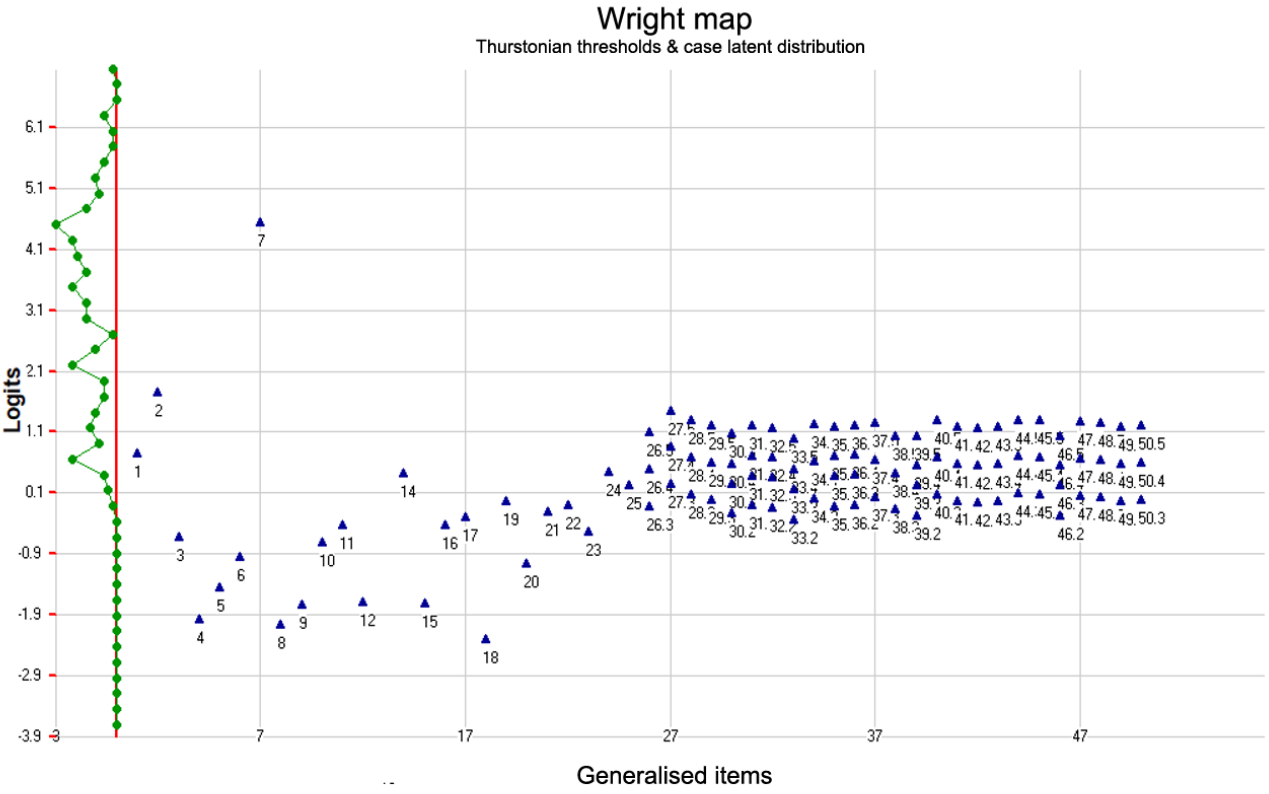


Figure 11. Distribution of students' self-confidence metacognition answers.

4. Conclusion and Suggestions

4.1. Conclusion

The analysis results indicate that the measurement of metacognition (Binary scale) and self-confidence (Likert scale) is reliable and meets the requirements based on Rasch model analysis. It is recommended to improve and clarify ambiguous and confusing questions for respondents by simplifying sentences or words accordingly. The distribution of participant score data and ability measures in the Rasch model shows that most participants have abilities at medium to high levels, with raw scores ranging from 21 to 24, covering 73.7% of the population. This suggests that participants have relatively homogeneous abilities at medium and high levels. Extreme scores (0 and 25) were not observed, indicating that only a few participants had very low or very high abilities.

The following equation can explain the linear relationship between raw scores and logit measures.

1. Score Prediction from Size: Score = Size × 2.953 + 12.46.
2. Size Prediction from Score: Size = score × 0.319 − 3.97.

In the self-confidence variable, analyzed as a separate or combined dimension, the measurement results show that the average level of student self-confidence logit is higher than the average metacognition logit. This indicates that Police Science College students tend to have greater self-confidence in decision-making. However, this has not been accompanied by a more mature plan of knowledge, experience, and organization of regulation to improve metacognition skills.

In multidimensional analysis, self-confidence can be considered as a separate dimension from metacognition, whereas in the rating scale model analysis, both can be combined into a single dimension. These results indicate that the measurement of metacognition and self-confidence for the two variables are interrelated and influence the process of making academic and non-academic decisions.

4.2. Suggestions

The application of Rasch model analysis is recommended using multidimensional analysis of the Rasch model with data that have different scales, such as Binary and Likert scales. Researchers should use a consistent data format and software that supports this analysis, such as Winsteps, Jamovi, and ConQuest. The use of the Rasch model can provide an expansion of meaning and insight into the relationship between dimensions compared to unidimensional analysis.

The need to validate data with different scales explains that both scales can be analyzed within a single model framework. Researchers can further explore the effect of different scales on model parameter estimates, especially regarding the size of the logit and standard error in each dimension.

Further researchers are advised to develop design instruments that involve different scales (Binary or Likert) so that they require more complex analysis. Therefore, initial testing on simulated data before application to real data is highly recommended to minimize potential errors.

Further researchers are advised to integrate a multidimensional approach with machine learning or data mining techniques to explore more complex patterns in the data, especially when the number of dimensions and items increases.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Anriani, N., & Gholobi, M. I. (2024). Developing an instrument to measure the effectiveness of the link and match learning and the availability of the practical laboratory facilities on the quality of vocational school graduates. *Jurnal Pendidikan Indonesia Gemilang, 4*(1), 1-7. https://doi.org/10.53889/jpig.v4i1.311

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York and London: Longmans.

Boone, W. J., Staver, R. J., & Yale, S. M. (2014). *Rasch analysis in the human sciences*. London: Springer.

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

Dwilaksana, C. (2020). *Have teachers become the key to education? YPKIK*. Jakarta: Police Science Study Development Foundation.

Ferdian, N. C., Anwar, L., & Herdini, H. (2025). Validity and practicality test of e-student worksheets using the ARCS (attention, relevance, confidence, and satisfaction) model on acid-base concepts. *International Journal of STEM Education for Sustainability, 5*(1), 132-149. https://doi.org/10.53889/ijses.v5i1.584

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Ilman, E. N., Putri, A., Julita, N. H., Ristanto, R. H., & Isfaeni, H. (2024). Development of a learning style diagnostic assessment instrument based on experiential learning theory. *International Journal of Biology Education Towards Sustainable Development, 4*(2), 83-100. https://doi.org/10.53889/ijbetsd.v4i2.542

Isfiani, I. R., & Ekanara, B. (2022). Metacognition profile on habits of mind in biology learning. *Jurnal Pendidikan Indonesia Gemilang, 2*(2), 95-104. https://doi.org/10.53889/jpig.v2i2.138

Iswahyudi, G. (2012). *Metacognitive activities in solving direct proof problems reviewed from gender and mathematical ability*. Paper presented at the Seminar Nasional Pendidikan Matematika UNS Surakarta. Surakarta.

Jarnudi, K., Hidayat, S., Syafrizal, S., Fahamzah, J., & Amrudin, A. (2024). Development of a language attitude instrument and its application to survey Islamic high school students' language attitudes towards English. *Jurnal Pendidikan Indonesia Gemilang, 4*(2), 135-149.

Ningsih, D. R., Ramalis, T. R., & Purwana, U. (2018). Development of critical thinking skills tests based on item response theory analysis. *WaPFi (Wahana Pendidikan Fisika), 3*(2), 45-50.

Rafik, M., & Sari, I. J. (2023). The urgency of developing computational thinking skills test instruments on biology subject for senior high school students. *International Journal of Biology Education Towards Sustainable Development, 3*(2), 94-103.

Retnawati, H. (2014). *Item response theory and its applications: For researchers, measurement and testing practitioners, graduate students*. Yogyakarta: Nuha Medika.

Rusdiyanto, D. R., Rubini, B., & Ardianto, D. (2023). Instrument for student attitudes assessment in science learning: A review and bibliometric analysis. *Jurnal Pendidikan Indonesia Gemilang, 3*(2), 209-221. https://doi.org/10.53889/jpig.v3i2.202

Sari, E. D. K., & Mahmudi, I. (2024). *Rasch modeling analysis in educational assessment. Analisis Menggunakan Aplikasi Winstep)* (1st ed.). Purwokerto: PT. Pena Persada Kerta UTama.

Sari, I. J., Pongsophon, P., Vongsangnak, W., Pimthong, P., & Pitiporntapin, S. (2022). The development of molecular genetics concept test for senior high school students using Rasch analysis. *International Journal of Evaluation and Research in Education, 11*(4), 1687-1695. https://doi.org/10.11591/ijere.v11i4.21846

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*(4), 460-475. https://doi.org/10.1006/ceps.1994.1033

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology, 15*(2), 201–292. https://doi.org/10.2307/1412107

Sugiharto, A., Cantika, W., & Werdhani, R. A. (2023). Validity and reliability of the medical student e-learning management system readiness scale. *Jurnal Pendidikan Indonesia Gemilang*, *3*(2), 254-266. https://doi.org/10.53889/jpig.v3i2.237

Sukmawati, S., Qadar, R., & Sulaeman, N. F. (2022). High school students' motivation towards physics learning during COVID-19 outbreak. *International Journal of STEM Education for Sustainability*, *2*(1), 105-132. https://doi.org/10.53889/ijses.v2i1.15

Sumintono, B., & Widhiarso, W. (2013). *Applications of the Rasch model: For social sciences research.* Jakarta: Trim Komunikata Publishing House.

Van Zile-Tamsen, C. (2017). Using Rasch analysis to inform rating scale development. *Research in Higher Education*, *58*(8), 922-933. https://doi.org/10.1007/s11162-017-9448-0

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116. https://doi.org/10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zohar, A., & Dori, Y. J. (2003). Higher order thinking skills and low-achieving students: Are they mutually exclusive? *Journal of the Learning Sciences*, *12*(2), 145-181. https://doi.org/10.1207/S15327809JLS1202_1