JPPIPA 11(11) (2025)

Jurnal Penelitian Pendidikan IPA

http://jppipa.unram.ac.id/index.php/jppipa/index

Validation of Instruments for Systems Thinking and Complex Problem-Solving Skills on Environmental Pollution Using the Rasch Model Approach

Ade Supriatno[1], Ida Kaniawati[1], Irma Rahma Suwarma[1]

[1]Departmen Pendidikan Fisika, Universitas Pendidikan Indonesia, Indonesia.

Received: September 19, 2025
Revised: October 12, 2025
Accepted: November 25, 2025
Published: November 30, 2025

Related Authors:
Ade Supriatno
adesupriatno@upi.edu

DOI: 10.29303/jppipa.v11i11.12903

Abstract: This study aims to test the validity and reliability of the System Thinking (ST) and Complex Problem Solving (CPS) instruments developed for vocational high school students. A quantitative approach was used using the Rasch model with the Winsteps tool. The ST instrument consisted of 8 multiple-choice items analyzed with a dichotomous model, while the CPS instrument consisted of 9 essay items with a polytomous model, and was tested on 11th-grade vocational high school students who had already learned about environmental pollution, with a total of 36 respondents. The analysis included unidimensionality, item fit, item–total correlation, reliability, and Wright Map mapping. The results showed a Raw Variance Explained by Measures of 41.3% (ST) and 44.3% (CPS), indicating that unidimensionality was fulfilled. The MNSQ Infit–Outfit values were within the ideal range of 0.5–1.5, with Point Measure Correlations of 0.37–0.85 (ST) and 0.46–0.75 (CPS). Person reliability was 0.87 (ST) and 0.75 (CPS), respectively, while item reliability was 0.92 (ST) and 0.86 (CPS), indicating high measurement consistency. The logit range on the Wright Map showed a balance between student ability and item difficulty. Thus, the ST and CPS instruments are proven to be valid and reliable for comprehensively measuring System Thinking and complex problem-solving abilities.

Keywords: Complex Problem Solving; Rasch Model; System Thinking; Validity; Reliability

Introduction

The 21st century is marked by increasingly complex global challenges, ranging from climate change and ecosystem degradation to environmental pollution. This complexity requires the younger generation to have high-level thinking skills that can integrate science, technology, and environmental awareness (OECD, 2019; UNESCO, 2021). In this context, systems thinking (ST) and complex problem solving (CPS) skills are core competencies needed to address global issues holistically (Arnold & Wade, 2015).

International assessment results, such as the Programme for International Student Assessment (PISA), show that Indonesian students' science literacy is still low, with more than 75% of students having difficulty solving context-based problems (Schleicher, 2019). This reflects the weakness of critical thinking, ST, and CPS skills that are urgently needed in modern science learning (Bybee, 2013). This condition has implications for students' difficulties in relating physics concepts to real environmental phenomena.

Local research reinforces these international findings. A study by Rustaman (2021) revealed that most high school students still have difficulty recognizing system components and the interactions between elements. Similarly, Pamungkas et al. (2023) found that students' ability to formulate strategies for solving complex problems was low. This is reinforced by

Citation:
Supriatno, A., Kaniawati, I., & Suwarma, I. R. (2025). Validation of Instruments for Systems Thinking and Complex Problem-Solving Skills on Environmental Pollution Using the Rasch Model Approach. *Jurnal Penelitian Pendidikan IPA*, *11*(11), 1309–1328. https://doi.org/10.29303/jppipa.v11i11.12903

findings that show students' weak integrative ability in connecting theory with practice (Rizal et al., 2022).

Limited ST and CPS abilities not only hinder learning achievement but also imply low environmental literacy among students. Giangrande et al. (2019) emphasize the importance of sustainability-oriented education to equip students to deal with environmental issues. On the other hand, Amos & Levinson (2019) assert that integrating environmental issues into science learning can increase students' ecological awareness and analytical skills.

However, the evaluation instruments currently available are not yet fully capable of capturing these two skills in an integrated manner. Most previous studies have only developed measurement tools for one skill, such as systems thinking or CPS alone. As a result, the measurement of student competence is still partial and does not provide a complete picture of their ability to solve environmental problems.

The results of a preliminary study using the systems thinking indicators developed by Rustaman (2021) and the complex problem-solving indicators developed by Pamungkas et al. (2023) at a vocational school in Bandung Regency further reinforce these findings. An analysis of 36 tenth-grade students at a vocational school in Bandung Regency shows that their average systems thinking skills are still low. The indicator for recognizing the structure and role of components in a system obtained an average score of 2.3 (low), the indicator for analyzing component interactions obtained 1.9 (low), the indicator for pattern analysis 1.6 (low), and the indicator for predicting system behavior 1.9 (low) out of a maximum score of 4. Similarly, complex problem-solving skills were also low, with average scores for the indicators of finding questions of 1.7 (low), devising plans of 1.9 (low), and concluding solutions of 1.3 (low). This data is reinforced by interviews with physics teachers who stated that students still have difficulty connecting theoretical concepts with real-world applications in an environmental context. These findings are consistent with physics teachers' statements that students often have difficulty connecting theoretical concepts with real-world applications in the context of environmental pollution.

This condition confirms the existence of a research gap, namely the absence of instruments that integrate the measurement of systems thinking and CPS simultaneously in the context of environmental pollution. In fact, this context is relevant to the physics curriculum and is close to the daily lives of students (Begum et al., 2021). Thus, this study attempts to present a new comprehensive evaluation instrument based on an integrative framework of both skills. The instrument to be developed will take the form of multiple-choice questions to measure system thinking skills and essay questions to measure complex problem solving skills so that students can express their thought processes in greater depth. The challenge in using multiple-choice and essay questions is assessment consistency. Therefore, the analysis was conducted using the Rasch Model with the Winsteps application, which is suitable for analyzing dichotomous and polytomous data (Boone & Staver, 2020).

The Rasch model with the Winsteps application has been widely recognized as a reliable approach for analyzing various forms of assessment instruments, both multiple-choice (dichotomous) and essay (polytomous) questions. Rasch analysis allows for the objective measurement of participants' abilities and item difficulty on a linear logit scale, while also identifying items that do not fit the model (misfit items) and comprehensively measuring test reliability (Boone, 2016). Research by Andrich & Marais (2018) shows that the Rasch model is effective in controlling bias in multiple-choice and essay questions with dichotomous and polytomous models. Additionally, Uto (2024) in Behavior Research Methods demonstrates the effectiveness of the Many-Facets Rasch Model (MFRM) in linking essay writing test assessments using data from various assessors. Similar results were also obtained by the WIDA, which reported the validation of an MFRM-based writing assessment scale with high inter-rater reliability. Boone & Staver (2020) emphasized that the Rasch model provides a strong mathematical basis for constructing measurable and bias-free educational instruments. Another study by Sumintono & Widhiarso (2015), shows that Winsteps can be applied to analyze multiple-choice test data in science education with consistent and valid results. Furthermore, Rahman (2023) research proves that the Rasch model is capable of accurately mapping student abilities and essay question validity, while Winarti & Al-Mubarak (2020) found that Winsteps is effective in identifying problematic items on multiple-choice tests in chemistry. Thus, the findings of various international and national studies reinforce that the Rasch Model with the Winsteps application can accurately measure and validate both multiple-choice and essay questions, making it a highly relevant approach for modern educational research.

With this approach, the construct validity and reliability of instruments can be tested objectively. In addition to contributing theoretically to the development of more accurate evaluation instruments, this research also has practical implications. Validated instruments are expected to help teachers map students' abilities more accurately, so that 21st-century skills-based learning strategies can be designed more effectively.

Furthermore, this instrument can serve as a basis for policymakers to formulate a curriculum that is responsive to global issues and future skill needs. Thus, this study aims to: (1) develop an integrative instrument to measure the systems thinking and CPS abilities of vocational high school students on environmental pollution material, and (2) validate the instrument using the Rasch Model, so as to obtain a valid and reliable measuring tool that is capable of providing a comprehensive picture of students' thinking skills.

Method

Research Design

This study uses a development research method with a focus on instrument validation. The objective is to develop and test the feasibility of instruments to measure students' systems thinking and complex problem solving (CPS) abilities in environmental pollution material. The validation approach uses the Rasch Model because it is capable of providing an in-depth analysis of item validity, instrument reliability, question difficulty level, and item discrimination power.

Research Subjects

The research subjects were 36 eleventh-grade students from a vocational school in Bandung Regency, selected using stratified sampling to represent variations in academic ability. All respondents had studied environmental pollution material in their classes.

Research Instruments

The validated instruments consisted of two types of tests, namely multiple choice and essay tests. The multiple choice test to measure students' thinking was developed based on Rustaman's indicators (2t with indicators covering 4 aspects as shown in Table 1.

Table 1. Summary of Indicators, Sub-Indicators, and System Thinking Questions (Rustaman, 2021)

System Thinking Indicators	System Thinking Sub-Indicators	Question Number
Able to recognize the structure and role of components and subcomponents in a system	Identify structural and functional relationships between system components at the same system level	1 and 2
Ability to analyze the interactions between components and subcomponents within a system	Identify feedback processes that occur between components and subcomponents within a system	3 and 4
Able to analyze patterns/modeling within the system	Create/develop modeling that describes the position of all components and subcomponents in the system framework in 2D/3D form	5 and 6
Being able to predict/review system behavior due to interactions within and outside the system	Predict/review the consequences arising from interventions in the system that cause the loss or addition of components/subcomponents in the system using previously designed modeling or patterns	7 and 8

Table 2. Summary of Indicators and Complex Problem Solving (Pamungkas et al., 2023)

Indicator	Number Question
Identifying questions in problems	9
Finding information in the problem	10
Connecting the information that has been obtained	11
Developing a plan to solve the problem	12
Identifying the steps needed to solve the problem	13
Using the methods that will be used to solve the problem	14
Thinking of other ways to solve the problem	15
Using effective and efficient ways to solve problems	16
Concluding the right solution	17

The complex problem solving (CPS) instrument developed by Pamungkas et al. (2023) in the form of essay questions covers nine indicators arranged in the form of questions based on the context of environmental pollution problems. The indicators are shown in Table 2.

Reseach Procedure

A needs analysis was conducted through a literature review and preliminary study to identify the low level of systems thinking and complex problem-solving (CPS) skills among students. Based on the results of this analysis, a measurement instrument was developed, which included the preparation of an indicator grid, the writing of questions, and content validation by experts in the field of physics education and learning evaluation. The validated instrument was then tested on 36 11th grade students at a vocational high school. The pilot test data were analyzed using Ministep software with the Rasch Model approach to assess the validity, reliability, difficulty level, and discriminating power of each item, thereby obtaining an instrument suitable for measuring students' systems thinking and complex problem-solving skills. The research flowchart is shown in Figure 1

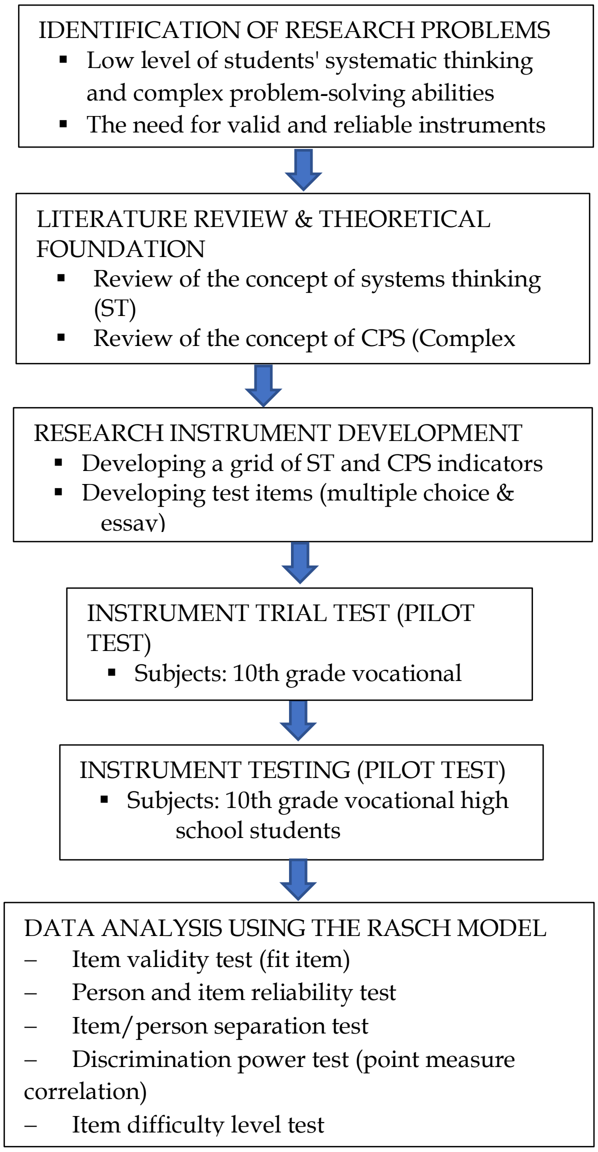

IDENTIFICATION OF RESEARCH PROBLEMS
▪ Low level of students' systematic thinking and complex problem-solving abilities
▪ The need for valid and reliable instruments
LITERATURE REVIEW & THEORETICAL FOUNDATION
▪ Review of the concept of systems thinking (ST)
▪ Review of the concept of CPS (Complex
RESEARCH INSTRUMENT DEVELOPMENT
▪ Developing a grid of ST and CPS indicators
▪ Developing test items (multiple choice & essay)
INSTRUMENT TRIAL TEST (PILOT TEST)
▪ Subjects: 10th grade vocational
INSTRUMENT TESTING (PILOT TEST)
▪ Subjects: 10th grade vocational high school students
DATA ANALYSIS USING THE RASCH MODEL
– Item validity test (fit item)
– Person and item reliability test
– Item/person separation test
– Discrimination power test (point measure correlation)
– Item difficulty level test

Figure 1. Research flow chart

Data Analysis Techniques

Data analysis was conducted quantitatively using the Rasch model with the assistance of Winsteps software to assess the validity and reliability of the System Thinking (ST) and Complex Problem Solving (CPS) instruments. The validity test included unidimensionality analysis through the Raw Variance Explained by Measures (RVEM) value with criteria >20% and unexplained variance <15%, and item feasibility based on the Infit–Outfit Mean Square (MNSQ) values in the range of 0.5–1.5 and Point Measure Correlation (Pt. Corr) values of 0.30–0.85. Reliability was analyzed through Person Reliability and Item Reliability to assess measurement consistency, while the Separation Index was used to examine the instrument's ability to distinguish between students' ability levels and item difficulty variations. Furthermore, Wright Map analysis was used to visualize the balance between student ability and item difficulty on the same logit scale, thereby obtaining a comprehensive picture of the measurement quality of the ST and CPS instruments.

Results and Discussion

Empirical Validity of the System Thinking (ST) and Complex Problem Solving (CPS) Instruments

The empirical validity of the ST and CPS instruments was tested using Rasch model analysis with the Winstep Rasch software. The steps of *Rasch* analysis in processing the ST and CPS instruments were: 1) correcting the test results by entering the students' answers into Microsoft Excel; 2) saving the test results data in PRN file format; 3) processing the test results data in PRN file format using winstep Rasch software. The processing of empirical validity tests using Winstep software on the *Rasch* model was obtained from the item dimensionality selection. This item provides information on the unidimensionality of the CT and CPS instruments. Unidimensionality is an important measure in the evaluation process of an instrument that provides a value indicating whether the research instrument (the developed ST and CPS) is capable of measuring what it is supposed to measure (valid). All ST and CPS items were analyzed to determine empirical validity for the knowledge aspect.

In the Rasch model, validity tests were analyzed for each item as a whole. Overall, the validity test in the Rasch model analysis of this study is referred to as unidimensionality (Sumintono & Widhiarso, 2015). Overall, the validity test of the ST and CPS instruments was obtained from the output tables menu option in the item section: dimensionality. The overall validity value of the items is shown by the Raw Variance Explained by Measures (RVEM) with the interpretation shown in Table 3.

Table 3. Interpretation of Unidimensionality (RVEM) of the ST and CPS Instruments (Sumintono & Widhiarso, 2015)

RVEM	Interpretation
20% ≤ RVEM < 40%	Met
40% ≤ RVEM < 60%	According to
RVEM ≥ 60%	Special

Unidimensionality of the instrument is determined from the unexplained variance value in the 1st contrast, with a criterion of less than 15% (Samsudin et al., 2020). In addition to testing the validity of the ST and CPS questions per aspect as a whole, the validity of each ST and CPS item was also tested. The validity test for each item of systems thinking and complex problem solving was conducted using the Rasch model by selecting the output tables menu in the fit order section in Winsteps Rasch. The quality of each item can be seen from three output values from the fit order processing in Winsteps Rasch, namely: outfit Z-standard (ZSTD), outfit mean square (MNSQ), and point measure correlation (Pt Measure Corr) (Sumintono & Widhiarso, 2015). The MNSQ value indicates the degree of randomness (deviation) in the instrument. The ZSTD value indicates the possibility of deviation in each item. The Pt Measure Corr value provides information on the relationship between the difficulty of each systems thinking and complex problem solving item and the difficulty of the instrument as a whole. The quality of each systems thinking and complex problem solving item was examined using the criteria in Table 4.

Table 4. Quality Criteria for Each Item (Boone, 2016; Samsudin et al., 2020)

Value	Description
-2.00 < ZSID < +2.00	Accepted
0.50 < MNSQ < 1.500	Accepted
0.40 < *Pt Measure Corr* < 0.85	Accepted

Based on Table 4, the third criterion for *fit order* processing values for each item is known. Then, each ST and CPS item is interpreted based on Table 5

Table 5 Interpretation of the Quality of Each Item (Boone, 2016; Samsudin et al., 2020)

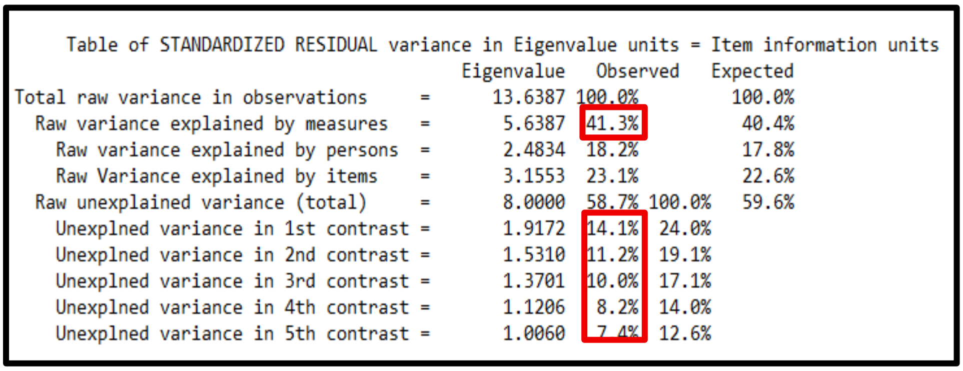
Criteria	Interpretation	Description
All met	Very appropriate	Valid Without Revision (VTR)
Compliant	Compliant	Valid Without Revision (VTR)
Less suitable	Less compliant	Valid with Revision (VR)
Not Suitable	Not Suitable	Not Valid (TV)

Empirical Validity of the Systems Thinking (ST) Instrument

The dimensional test results aim to ensure that each indicator used truly represents the construct of conceptual system thinking ability. The dimensional test results containing the validity of ST in the knowledge aspect are presented in Figure 2.

Based on Figure 2, it is obtained that the raw variance explained by measures is 41.3%**.** This value meets the minimum unidimensionality requirement, which is 20% or more (Setiyowati et al., 2020). Other measurements in the unidimensionality test show that the unexplained variance value of the five contrast residuals does not exceed 15%, so it can be concluded that the instrument meets the unidimensionality assumption. This is in line with the opinions of Brentari & Golia (2007); Tennant & Conaghan (2023), who explain that unidimensionality is a major prerequisite in the Rasch model, and that high unexplained variance (>15%) indicates potential multidimensionality.

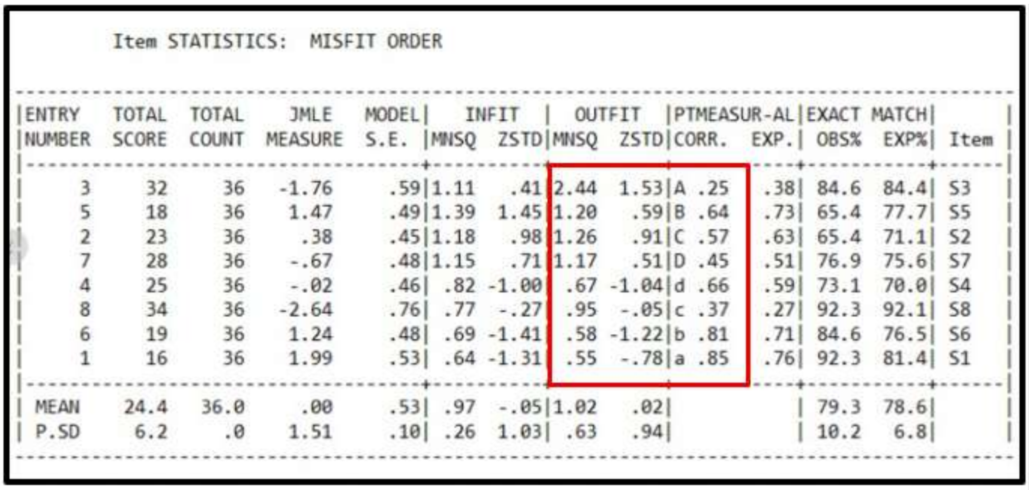
Thus, the results of this analysis reinforce that the items in the System Thinking (ST) instrument in the knowledge aspect have measured the same construct, namely system thinking ability, and there is no significant influence from other dimensions. This finding is also in line with the research by Oliva & Blanco (2023), which shows that a variance explained value of around 40–50% already reflects good unidimensionality in the context of measuring students' cognitive abilities. Therefore, it can be concluded that the ST questions in the knowledge aspect measure what they are supposed to measure, and the overall validity of the instrument has been fulfilled according to the Rasch model, so this instrument is declared to have good validity.



```
    Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units
                                             Eigenvalue   Observed   Expected
Total raw variance in observations      =       13.6387 100.0%         100.0%
  Raw variance explained by measures    =        5.6387  41.3%          40.4%
    Raw variance explained by persons   =        2.4834  18.2%          17.8%
    Raw Variance explained by items     =        3.1553  23.1%          22.6%
  Raw unexplained variance (total)      =        8.0000  58.7% 100.0%   59.6%
    Unexplned variance in 1st contrast =        1.9172  14.1%  24.0%
    Unexplned variance in 2nd contrast =        1.5310  11.2%  19.1%
    Unexplned variance in 3rd contrast =        1.3701  10.0%  17.1%
    Unexplned variance in 4th contrast =        1.1206   8.2%  14.0%
    Unexplned variance in 5th contrast =        1.0060   7.4%  12.6%
```

Figure 2. Results of the Dimensional Test of the ST Instrument in the Knowledge Aspect

In addition to testing the validity of the ST knowledge aspect questions as a whole, the validity of the ST instrument was also tested for each knowledge aspect item obtained from the winstep Rasch application in the item selection (column): fit order. The results of the ST knowledge aspect instrument validity test are presented in Figure 3.



```
        Item STATISTICS:  MISFIT ORDER

-------------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL    JMLE   MODEL|   INFIT  |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|          |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item     |
|------------------------------------+----------+----------+-----------+-----------+----------|
|     3     32     36    -1.76    .59|1.11   .41|2.44  1.53|A .25   .38| 84.6  84.4| S3       |
|     5     18     36     1.47    .49|1.39  1.45|1.20   .59|B .64   .73| 65.4  77.7| S5       |
|     2     23     36      .38    .45|1.18   .98|1.26   .91|C .57   .63| 65.4  71.1| S2       |
|     7     28     36     -.67    .48|1.15   .71|1.17   .51|D .45   .51| 76.9  75.6| S7       |
|     4     25     36     -.02    .46| .82 -1.00| .67 -1.04|d .66   .59| 73.1  70.0| S4       |
|     8     34     36    -2.64    .76| .77  -.27| .95  -.05|c .37   .27| 92.3  92.1| S8       |
|     6     19     36     1.24    .48| .69 -1.41| .58 -1.22|b .81   .71| 84.6  76.5| S6       |
|     1     16     36     1.99    .53| .64 -1.31| .55  -.78|a .85   .76| 92.3  81.4| S1       |
|------------------------------------+----------+----------+-----------+-----------+----------|
| MEAN    24.4   36.0      .00    .53| .97  -.05|1.02   .02|           | 79.3  78.6|          |
| P.SD     6.2     .0     1.51    .10| .26  1.03| .63   .94|           | 10.2   6.8|          |
-------------------------------------------------------------------------------------------------
```

Figure 3. Results of the Validity of the ST Instrument for the Knowledge Aspect

Figure 3 provides an overview of the ZSTD, MNSQ, and Pt Measure Corr values to examine the validity of each question. The interpretation of the data for the three outfit values for the knowledge aspect based on Figure 3 is presented in Table 6.

Table 6. Interpretation of *Outfit ZSTD*, *MNSQ*, and *Pi Measure Corr* Data for the Knowledge Aspect of the ST Instrument

Item Number	*MNSQ*	*ZSTD*	*Pt Measure Corr*	*Interpretation*	Note
1	0.55	- 0.78	*a 0.85	As per	VTR
2	1.26	0.91	C 0.57	Very Suitable	VTR
3	*2.44	1.53	A 0.25	As per	VTR
4	0.67	-1.04	d 0.66	Very Suitable	VTR
5	1.20	0.59	B 0.64	Very Suitable	VTR
6	0.58	- 1.22	b 0.81	Very Suitable	VTR
7	1.17	0.51	D 0.45	Very Suitable	VTR
8	0.95	-0.05	*c.0.37	As expected	VTR

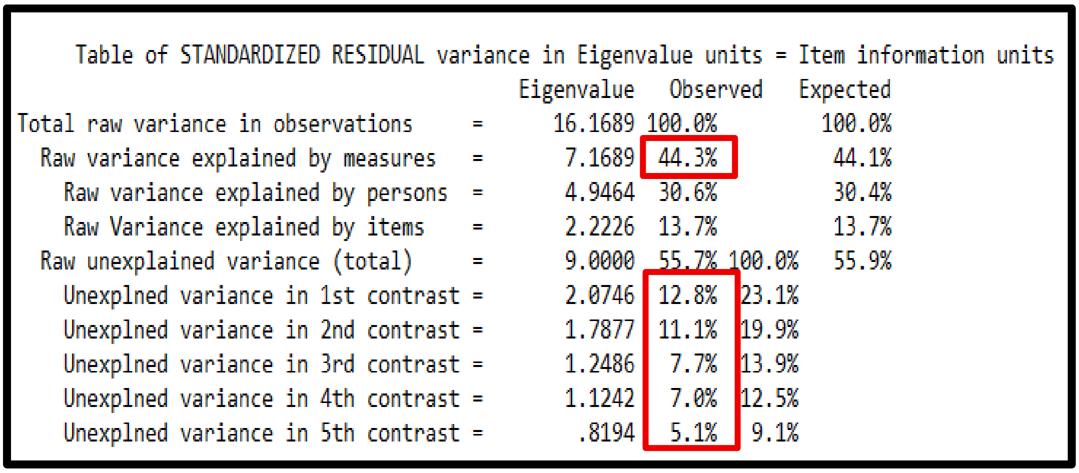
Note: * Does not meet criteria

Based on Table 6, the Outfit MNSQ, ZSTD, and Pt Measure Corr values for items 1, 2, 3, 4, 5, 6, 7, and 8 show results that generally meet the item feasibility criteria because most items meet at least two or three criteria from the Rasch model parameters. Outfit Mean Square (MNSQ) values in the range of 0.5–1.5 indicate that the item has a good level of fit with the model (Boone & Staver, 2020). In addition, Z-standardized (ZSTD) values close to 0 and Point Measure Correlation (Pt Measure Corr) values that are positive and greater than 0.3 indicate that each item is able to distinguish respondents well and contribute positively to the measurement of the construct being tested.

These results are in line with the findings of Sumintono & Widhiarso (2015), who explain that an item can be considered valid if it meets at least two of the three main criteria (Outfit MNSQ, ZSTD, and Pt Measure Corr), because small variations in one parameter are still acceptable as long as consistency between the other parameters is maintained. Similar support was also expressed by Azizi et al. (2023); Linacre (2018) ,who emphasized that the assessment of item validity in Rasch analysis is more accurate if it is based on a balance between fit statistics and item-respondent correlations rather than a single indicator.

Empirical Validation of the Complex Problem Solving (CPS) Instrument

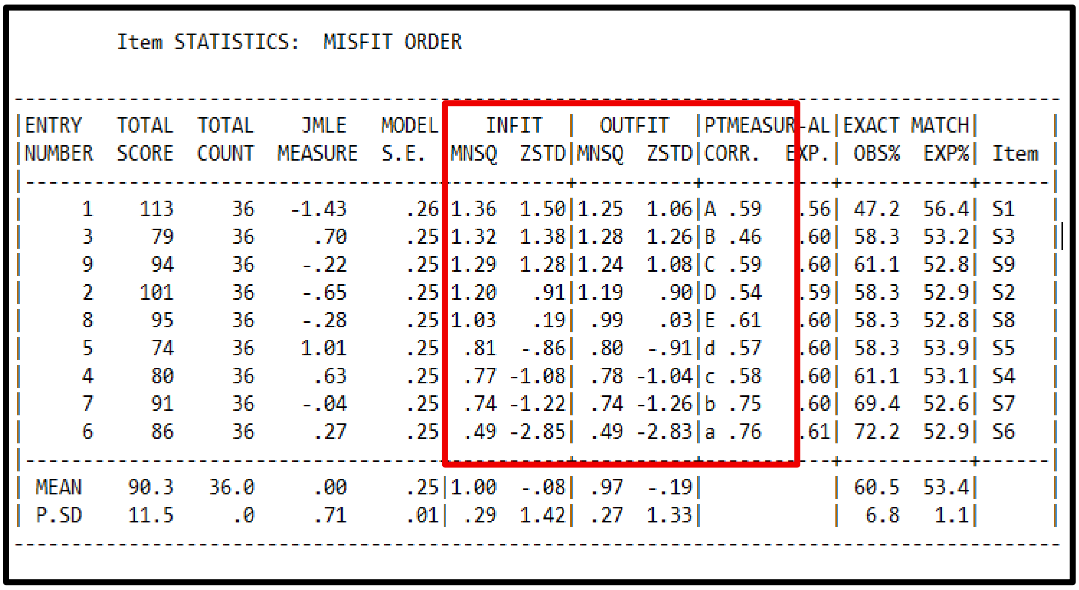
The results of the dimensional test containing the validity of the scientific explanation level instrument are presented in Figure 4



```
   Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units
                                          Eigenvalue   Observed   Expected
Total raw variance in observations     =      16.1689 100.0%         100.0%
  Raw variance explained by measures   =       7.1689  44.3%          44.1%
    Raw variance explained by persons  =       4.9464  30.6%          30.4%
    Raw Variance explained by items    =       2.2226  13.7%          13.7%
  Raw unexplained variance (total)     =       9.0000  55.7% 100.0%   55.9%
    Unexplned variance in 1st contrast =       2.0746  12.8%  23.1%
    Unexplned variance in 2nd contrast =       1.7877  11.1%  19.9%
    Unexplned variance in 3rd contrast =       1.2486   7.7%  13.9%
    Unexplned variance in 4th contrast =       1.1242   7.0%  12.5%
    Unexplned variance in 5th contrast =        .8194   5.1%   9.1%
```

Figure 4. Dimensional Test of the CPS Instrument

Based on Figure 4, the raw variance measurement results explained by the CPS instrument of 44.3% indicate that the unidimensionality assumption has been met because it exceeds the minimum limit of 40% (Samsudin et al., 2020). The unexplained variance value, which is below 15%, also shows that the remaining unexplained variance is still within reasonable limits, so that this instrument is able to measure the intended construct consistently. These results are in line with the findings of Hidayat et al. (2021), who reported that Rasch-based higher-order thinking instruments with raw variance > 40% show strong construct validity. In addition, Rasool & Marlina (2023) emphasized that an explained variance value above 40% indicates that the instrument is unidimensional and reliable. Meanwhile, Nurhasanah et al. (2024) found that systemic thinking instruments with unexplained variance < 15% are valid because they can measure complex thinking abilities consistently. Thus, the analysis results show that the CPS instrument has excellent validity according to the Rasch model and is suitable for use in measuring students' complex problem-solving abilities. In addition to testing the validity of the CPS questions as a whole, the validity of the CPS level questions was also tested using the Winstep Rasch application on item selection (column) and fit order. The results of the CPS instrument validity test are presented in Figure 5.



```
        Item STATISTICS:  MISFIT ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL    JMLE   MODEL|   INFIT  |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|     1    113     36    -1.43    .26|1.36  1.50|1.25  1.06|A .59   .56| 47.2  56.4| S1   |
|     3     79     36      .70    .25|1.32  1.38|1.28  1.26|B .46   .60| 58.3  53.2| S3   ||
|     9     94     36     -.22    .25|1.29  1.28|1.24  1.08|C .59   .60| 61.1  52.8| S9   |
|     2    101     36     -.65    .25|1.20   .91|1.19   .90|D .54   .59| 58.3  52.9| S2   |
|     8     95     36     -.28    .25|1.03   .19| .99   .03|E .61   .60| 58.3  52.8| S8   |
|     5     74     36     1.01    .25| .81  -.86| .80  -.91|d .57   .60| 58.3  53.9| S5   |
|     4     80     36      .63    .25| .77 -1.08| .78 -1.04|c .58   .60| 61.1  53.1| S4   |
|     7     91     36     -.04    .25| .74 -1.22| .74 -1.26|b .75   .60| 69.4  52.6| S7   |
|     6     86     36      .27    .25| .49 -2.85| .49 -2.83|a .76   .61| 72.2  52.9| S6   |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN    90.3   36.0      .00    .25|1.00  -.08| .97  -.19|           | 60.5  53.4|      |
| P.SD    11.5     .0      .71    .01| .29  1.42| .27  1.33|           |  6.8   1.1|      |
-------------------------------------------------------------------------------------------
```

Figure 5. CPS Instrument Validity Test per Item

Figure 5 provides an overview of the ZSTD, MNSQ, and Pr Measure Corr outfit values to see the validity of each item. The results of the interpretation of the ZSTD, MNSQ, and Pr Measure Corr outfit data for the CPS item instrument based on Figure 3.13 are presented in Table 7.

Table 7. Interpretation of the ZSTD, MNSQ, and Pr Measure Corr data for the CPS instrument

Question number	MNSQ	ZSTD	Pt Measure Corr	Interpretation	Description
1	1.25	1.06	A 0.59	Very Suitable	VTR
2	1.19	0.90	D 0.54	Very Suitable	VTR
3	1.28	1.26	B 0.46	Very Suitable	VTR
4	0.78	-1.04	c 0.58	Very Suitable	VTR
5	0.80	-0.91	d 0.57	Very Suitable	VTR
6	0.49	-2.83*	b 0.76	As expected	VTR
7	0.74	-1.26	b 0.75	Very Suitable	VTR
8	1.03	0.03	E 0.61	Very Suitable	VTR
9	0.99	1.08	C 0.59	Very Suitable	VTR

Based on Table 7, the Outfit MNSQ, ZSTD, and Pt Measure Corr values for all CPS items are within the range recommended by the Rasch model, namely $0.5 \leq$ MNSQ ≤ 1.5 and $-2 \leq$ ZSTD $\leq +2$, and the Pt Measure Corr value is above 0.30. This indicates that each item functions well and is able to measure the construct of complex problem-solving ability consistently without the need for revision. These results are in line with the findings of Putra et al. (2022), who explained that MNSQ and ZSTD values within the ideal range indicate the suitability of items for the Rasch model. In addition, Rahman & Ismail (2023) emphasized that Pt Measure Corr values above 0.40 indicate a positive correlation between respondents' abilities and test items. This is also in line with Lee et al. (2024), who found that the consistency of MNSQ and ZSTD values indicates the validity and stability of the instrument. Thus, it can be concluded that all CPS items have excellent validity.

Reliability of the ST and CPS Instruments

Instrument reliability is the level of consistency of an instrument. An instrument is reliable if it can produce the same measurement data on the same object when the instrument is used several times (Sugiyono, 2015). The reliability of the science inquiry literacy and scientific explanation instruments was processed using the WinStep Rasch software in the summary statistics menu output tables. The data obtained from the summary statistic section of the " " menu were person reliability (p), item reliability (r), and Cronbach's alpha (KR-20).

The ST and CPS instruments in the reliability testing study were analyzed using the Rasch model. The Summary Statistics menu was selected in the WinStep software to test the reliability of these instruments. The summary statistics menu provided data on person reliability (p), item reliability (r), and Cronbach's alpha (KR-20) for the ST and CPS instruments in terms of knowledge. The interpretation of p and t is shown in Table 8.

Table 8. Interpretation of *p* and *r* Values for the ST and CPS Instruments (Sumintono & Widhiarso, 2015)

Data	Value	Interpretation
p and r	$p,r \leq 0.67$	Low
	$0.67 < p,r \leq 0.80$	Moderate
	$0.80 < p,r \leq 0.90$	Good
	$0.90 < p,r \leq 0.94$	Very Good
	$p,r > 0.94$	Excellent

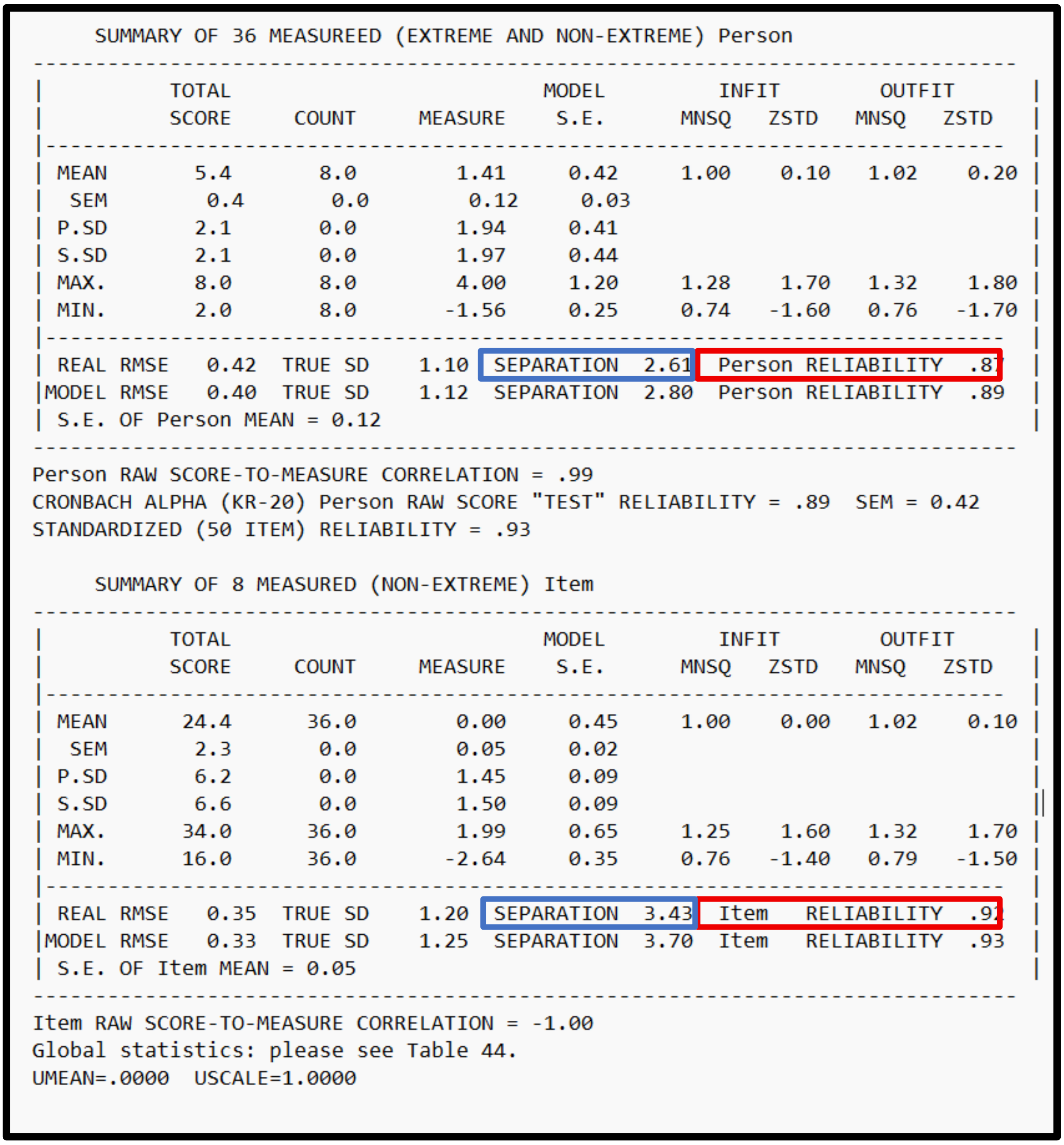
Meanwhile, to describe how well an instrument is able to distinguish the ability level of respondents (person separation) or the quality of test items (item separation), person/item separation values were used. This value is an important indicator in assessing the accuracy and reliability of an instrument. The higher the separation value, the better the instrument's ability to group participants' abilities or item difficulty levels. According to Sumintono & Widhiarso (2015), the separation categories can be interpreted as shown in the following Table 9.

Table 9. Reliability and Separation Values Based on Rasch Model Analysis (Linacre, 2021; Sumintono & Widhiarso, 2015)

Separation Value	Category	Interpretation
< 2.00	Low	The instrument is less able to distinguish between abilities/items
2.00 – 3.00	Good	The instrument is sufficiently capable of distinguishing ability/items.
> 3.00	Very good	The instrument is very good at distinguishing abilities/items

Results of the reliability test of the ST instrument

The results of the reliability test of the ST instrument for the knowledge aspect in the study are presented in Figure 6.



```
     SUMMARY OF 36 MEASUREED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       5.4       8.0       1.41     0.42      1.00   0.10   1.02   0.20 |
|  SEM        0.4       0.0       0.12     0.03                               |
| P.SD       2.1       0.0       1.94     0.41                                |
| S.SD       2.1       0.0       1.97     0.44                                |
| MAX.       8.0       8.0       4.00     1.20      1.28   1.70   1.32   1.80 |
| MIN.       2.0       8.0      -1.56     0.25      0.74  -1.60   0.76  -1.70 |
|-----------------------------------------------------------------------------|
| REAL RMSE    0.42  TRUE SD    1.10  SEPARATION  2.61  Person RELIABILITY  .87 |
|MODEL RMSE    0.40  TRUE SD    1.12  SEPARATION  2.80  Person RELIABILITY  .89 |
| S.E. OF Person MEAN = 0.12                                                  |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .89  SEM = 0.42
STANDARDIZED (50 ITEM) RELIABILITY = .93

     SUMMARY OF 8 MEASURED (NON-EXTREME) Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      24.4      36.0       0.00     0.45      1.00   0.00   1.02   0.10 |
|  SEM       2.3       0.0       0.05     0.02                                |
| P.SD       6.2       0.0       1.45     0.09                                |
| S.SD       6.6       0.0       1.50     0.09                                |
| MAX.      34.0      36.0       1.99     0.65      1.25   1.60   1.32   1.70 |
| MIN.      16.0      36.0      -2.64     0.35      0.76  -1.40   0.79  -1.50 |
|-----------------------------------------------------------------------------|
| REAL RMSE    0.35  TRUE SD    1.20  SEPARATION  3.43  Item    RELIABILITY  .92 |
|MODEL RMSE    0.33  TRUE SD    1.25  SEPARATION  3.70  Item    RELIABILITY  .93 |
| S.E. OF Item MEAN = 0.05                                                    |
-------------------------------------------------------------------------------
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00
Global statistics: please see Table 44.
UMEAN=.0000  USCALE=1.0000
```

Figure 6. Reliability Test of the ST Instrument Knowledge Aspect

Table 10. Reliability Data Summary

	Reliability
Person	0.87 (good)
Item	0.92 (very good)

Table 10 shows that the *person reliability* value of 0.87 is in the good category. This means that the students' answers on this instrument are quite consistent and stable. The item reliability value of 0.92 is in the very good category, which means that the items in this instrument are able to measure consistently and are of high quality. Overall, these results show that the *System Thinking (ST)* instrument used is reliable and suitable for measuring students' system thinking abilities in the four indicators tested. These results are in line with the research by Ghasemi et al. (2022), which states that high reliability indicates good measurement stability, as well as the findings of Nurhayati et al. (2024); Sari et al. (2020), which explain that instruments with a reliability value above 0.8 are considered very good and reliable for measuring students' abilities consistently. Meanwhile, the person separation and item separation values can be seen in Table 11.

Table 11. Person and item separation values

Type of Separation	Reliability Value	Category	Interpretation
Person Separation	2.61	Good	The instrument is able to distinguish participants into ±3 different ability levels
Item Separation	3.43	Very Good	The items have varying and stable levels of difficulty

Based on Figure 6, the Item Person and Item Reliability values can be seen in Table 12.

Table 12. Reliability Data Summary

	Reliability
Person	0.75 (good)
Item	0.86 (very good)

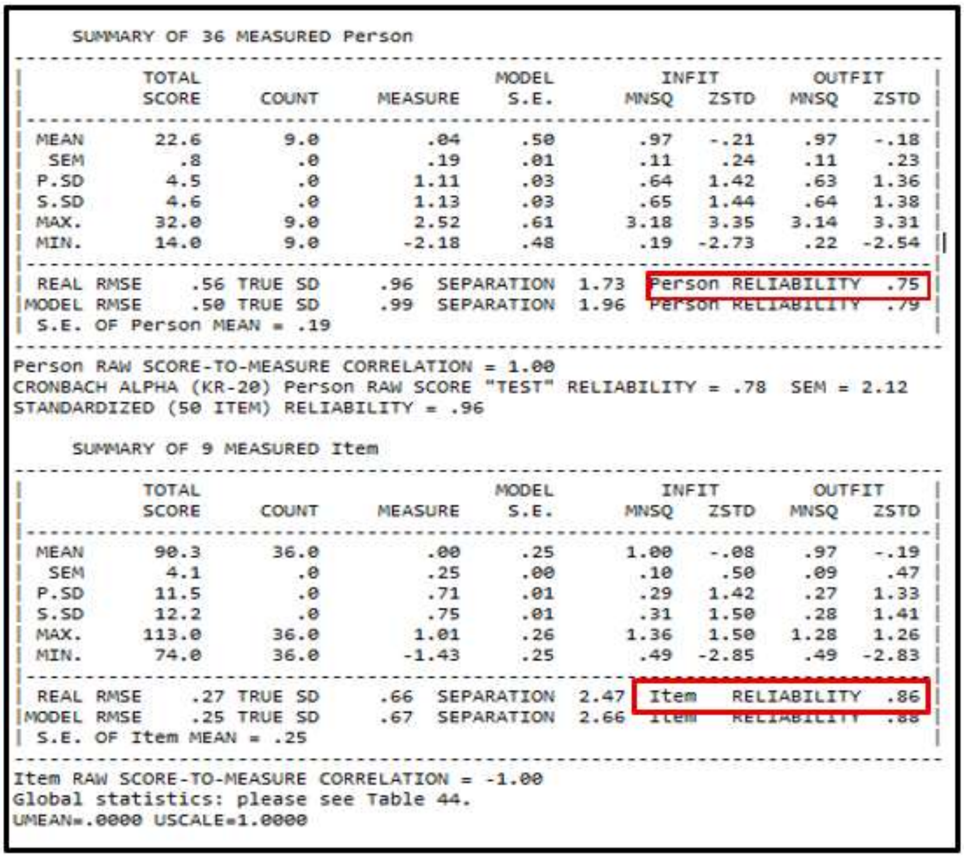
Based on the analysis results, a Person Separation of 2.61 was obtained, which is classified as good, indicating that the instrument can distinguish participants into approximately three different ability levels, so that variations in ability between individuals can be identified quite clearly. Meanwhile, the Item Separation of 3.43 is classified as very good, indicating that the items have varying levels of difficulty and are stable in measuring the range of participants' abilities consistently. High separation values for both person and item aspects indicate reliable measurement quality and provide strong diagnostic information on respondent abilities and item characteristics. These results are in line with the findings of Sumintono & Widhiarso (2015), who stated that a separation value above 2 indicates the instrument's ability to effectively classify participant abilities and item difficulty. In addition, Fitrah et al. (2024) emphasized that an item separation value above 3 reflects excellent item stability, while Bintang & Suprananto (2024) concluded that the higher the separation value, the better the instrument's ability to differentiate respondents and the overall quality of measurement.

Instrument Reliability Results for CPS

Based on Figure 7 and Table 11, the analysis results show that the total average score of participants is 22.6 out of a maximum score of 36. The person reliability value of 0.75 is in the adequate category, while the item reliability of 0.86 indicates a good criterion, which means that the instrument has high measurement stability and consistency in terms of respondent ability. The high item reliability value indicates that each item has a good level of suitability in measuring the expected construct and provides replicable results when tested on different groups of respondents (Boone, 2016). In the context of analysis using the Rasch Model, person and item reliability are important indicators for assessing instrument quality and measurement accuracy (Bond & Fox, 2015).A person reliability value above 0.70 is considered to meet the criteria for a reliable instrument (Linacre, 2021). In addition, research by Baghaei & Tabatabaee-Yazdi (2022) confirms that high item reliability indicates the stability of item parameters against variations in participant abilities, thereby strengthening the validity of the measurement results. Thus, these results indicate that the instrument performs well in assessing systematic thinking abilities or aspects that are measured consistently and accurately. Meanwhile, the person and item separation analysis can be seen in the Table 13.

Table 13. Results of Person and Item Separation Analysis Based on the Rasch Model

Type of Separation	Value	Category	Interpretation
Person Separation	1.73	Good	The instrument has adequate ability to distinguish participants into three different levels of ability.
Item Separation	2.47	Very Good	The items are able to represent variations in difficulty levels with high measurement stability.



```
     SUMMARY OF 36 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   S.E.       MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      22.6       9.0        .04     .50        .97   -.21    .97   -.18 |
|  SEM        .8        .0        .19     .01        .11    .24    .11    .23 |
| P.SD       4.5        .0       1.11     .03        .64   1.42    .63   1.36 |
| S.SD       4.6        .0       1.13     .03        .65   1.44    .64   1.38 |
| MAX.      32.0       9.0       2.52     .61       3.18   3.35   3.14   3.31 |
| MIN.      14.0       9.0      -2.18     .48        .19  -2.73    .22  -2.54 ||
|-----------------------------------------------------------------------------|
| REAL RMSE    .56 TRUE SD     .96  SEPARATION  1.73  Person RELIABILITY  .75 |
|MODEL RMSE    .50 TRUE SD     .99  SEPARATION  1.96  Person RELIABILITY  .79 |
| S.E. OF Person MEAN = .19                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .78  SEM = 2.12
STANDARDIZED (50 ITEM) RELIABILITY = .96

     SUMMARY OF 9 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   S.E.       MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      90.3      36.0        .00     .25       1.00   -.08    .97   -.19 |
|  SEM       4.1        .0        .25     .00        .10    .50    .09    .47 |
| P.SD      11.5        .0        .71     .01        .29   1.42    .27   1.33 |
| S.SD      12.2        .0        .75     .01        .31   1.50    .28   1.41 |
| MAX.     113.0      36.0       1.01     .26       1.36   1.50   1.28   1.26 |
| MIN.      74.0      36.0      -1.43     .25        .49  -2.85    .49  -2.83 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .27 TRUE SD     .66  SEPARATION  2.47  Item   RELIABILITY  .86 |
|MODEL RMSE    .25 TRUE SD     .67  SEPARATION  2.66  Item   RELIABILITY  .88 |
| S.E. OF Item MEAN = .25                                                     |
-------------------------------------------------------------------------------
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000
```

Figure 7. CPS Instrument Reliability Test

The analysis results show that a Person Separation of 1.73 indicates the instrument's ability to differentiate participants into several different ability levels, with good measurement stability. Meanwhile, an item Separation of 2.47 shows that the items have a wide and consistent distribution of difficulty, enabling them to map participants' abilities more accurately. A good separation value like this reflects a balance between measurement accuracy and item quality in the Rasch model. According to Boone & Staver (2020), a separation value above 1.5 already indicates strong discriminatory power in the context of educational assessment. Furthermore, Linacre (2021) explains that an item separation value above 2.5 indicates that the items are diverse enough to describe a wide range of difficulty levels. These results are also consistent with the research by Kim and Wilson (2023), which confirms that the higher the separation value, the greater the instrument's ability to assess variations in participants' abilities with a high level of reliability. Thus, the instrument analyzed has met the criteria for good measurement quality based on the Rasch model standards.

Instrument Difficulty Level (IDL)

The difficulty level describes the level of difficulty students face in answering questions. A good instrument is one that contains questions with diverse TKI values. The TKI of the systems thinking and complex problem-solving instruments were analyzed using the Rasch model with the Winsteps Rasch application. The TKI data for each ST and CPS question were analyzed using Winsteps Rasch based on the Measure (M) and Standard Deviation (SD) values from the Rasch results. The TKI data obtained from the ST and CPS instruments were interpreted by comparing the M and SD values shown in Table 14.

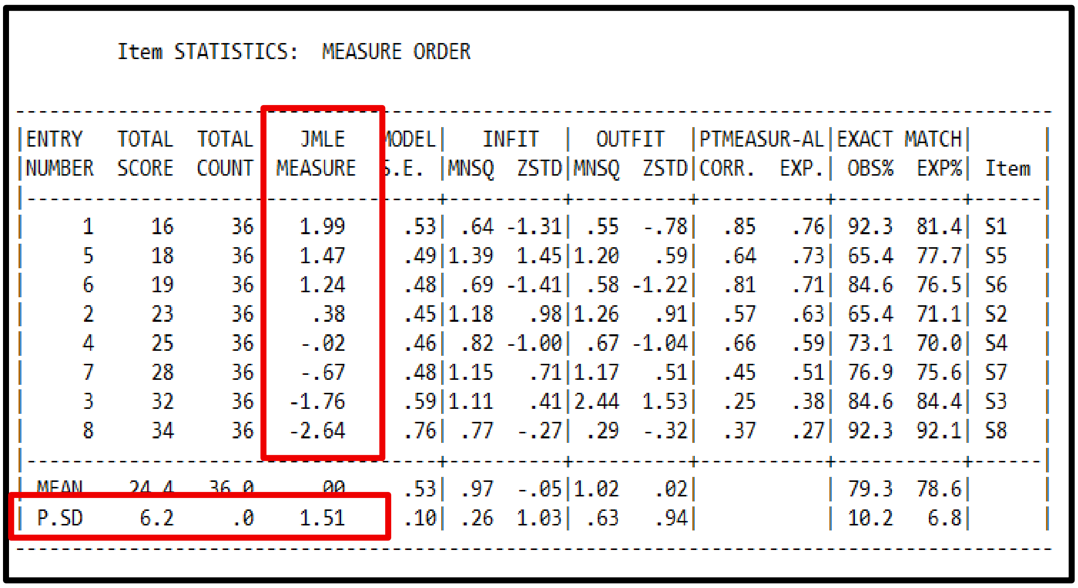
Table 14. Interpretation of ST and CPS Difficulty Levels (Sumintono & Widhiarso, 2015)

Criteria	Interpretation
M > +SD	Difficult
+SD≤ M < -SD	Moderate
M≤ -SD	Easy

The difficulty level of the ST and CPS instruments in terms of knowledge was analyzed based on the Rasch model in the menu item measure in the WinStep software. Table 13 is sorted from the most difficult to the easiest questions based on the JMLE Measure. The higher the logit value, the more difficult the item is.

Results of the ST TKI Test

The detailed results of the ST TKI test for the knowledge aspect can be seen in Figure



```
        Item STATISTICS:  MEASURE ORDER

------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL   JMLE    ODEL|   INFIT   |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  .E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|     1     16     36    1.99    .53| .64 -1.31| .55  -.78|  .85   .76| 92.3  81.4| S1   |
|     5     18     36    1.47    .49|1.39  1.45|1.20   .59|  .64   .73| 65.4  77.7| S5   |
|     6     19     36    1.24    .48| .69 -1.41| .58 -1.22|  .81   .71| 84.6  76.5| S6   |
|     2     23     36     .38    .45|1.18   .98|1.26   .91|  .57   .63| 65.4  71.1| S2   |
|     4     25     36    -.02    .46| .82 -1.00| .67 -1.04|  .66   .59| 73.1  70.0| S4   |
|     7     28     36    -.67    .48|1.15   .71|1.17   .51|  .45   .51| 76.9  75.6| S7   |
|     3     32     36   -1.76    .59|1.11   .41|2.44  1.53|  .25   .38| 84.6  84.4| S3   |
|     8     34     36   -2.64    .76| .77  -.27| .29  -.32|  .37   .27| 92.3  92.1| S8   |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN    24.4   36.0     .00    .53| .97  -.05|1.02   .02|           | 79.3  78.6|      |
| P.SD     6.2     .0    1.51    .10| .26  1.03| .63   .94|           | 10.2   6.8|      |
------------------------------------------------------------------------------------------
```

Figure 8. Data on the Difficulty Level Test of the ST Instrument for the Knowledge Aspect

Based on Figure 8, it can be seen that the M and SD values of the ST instrument are clear for each question. The SD value obtained is 1.51. Based on the comparison of the M and SD values, the interpretation of the ST TKI knowledge aspect is shown in Table 15.

Table 15. Interpretation Results of the ST TKI Knowledge Aspect

Type of Interpretation	SD	Question Number
Difficult	1.51	1
Moderate	1.51	2, 4, 5, 6, 7
Easy	1.51	3, 8

Based on Table 14, it is obtained that the TKI ST consists of 2 questions with the "Easy" criterion, 5 questions with the "Moderate" criterion, and 1 question with the "Difficult" criterion. This shows that the TKI ST is well distributed. These results are supported by Metsämuuronen (2023) research, which explains that the proportional distribution of item difficulty levels is important for obtaining accurate ability estimates in the Rasch model. Research by Dewi et al. (2023) also found that the distribution of questions from easy to difficult categories produces high reliability and strengthens the construct validity of the instrument.

In addition, a study in Frontiers in Psychology (2021) confirms that the balance of item positions across the difficulty range creates good targeting between respondent ability and item characteristics. Thus, the results of this analysis show that the TKI ST knowledge aspect instrument has met the criteria for good distribution and is suitable for use in consistently measuring students' systematic thinking abilities, while the item fit analysis can be seen in Table 16.

Table 16. Item Fit Analysis Results (Item Fit Statistics)

Item	Infit MNSQ	MNSQ Outfit	Infit ZSTD	ZSTD Outfit	Point Measurement Correction	Description
S1	0.64	0.55	-1.31	-0.78	0.85	Fit (good)
S5	1.39	1.20	1.45	0.59	0.64	Fit (good)
S6	0.81	0.58	-1.41	-1.22	0.71	Fit (good)
S2	1.18	1.26	0.98	1.91	0.57	Fit (good)
S4	0.82	0.67	-1.00	-0.44	0.66	Fit (good)
S7	1.15	1.44	1.71	1.53	0.55	Almost misfit (needs review)
S3	1.11	1.44	1.21	1.53	0.25	Almost misfit (needs review)
S8	0.77	0.79	-0.27	-0.32	0.37	Fit (good)

Based on Table 16, the results of the item fit analysis in the table above show that all items have Infit and Outfit MNSQ values in the range of 0.5–1.5, indicating that, in general, all items fit the Rasch model (Linacre, 2018). Most ZSTD values are in the range of ±2, confirming that the deviation of items from the model is not statistically significant.

However, items S7 and S3 show an Outfit MNSQ of 1.44 and a ZSTD above +1.5, which indicates potential underfit—meaning that the participants' response patterns on these items are noisier than predicted by the model. This could be due to sentence ambiguity, inconsistent question context, or other factors outside the construct being measured. Meanwhile, the point measure correlation on both items is still positive (>0.20), so the items can still be retained with minor editorial revisions.

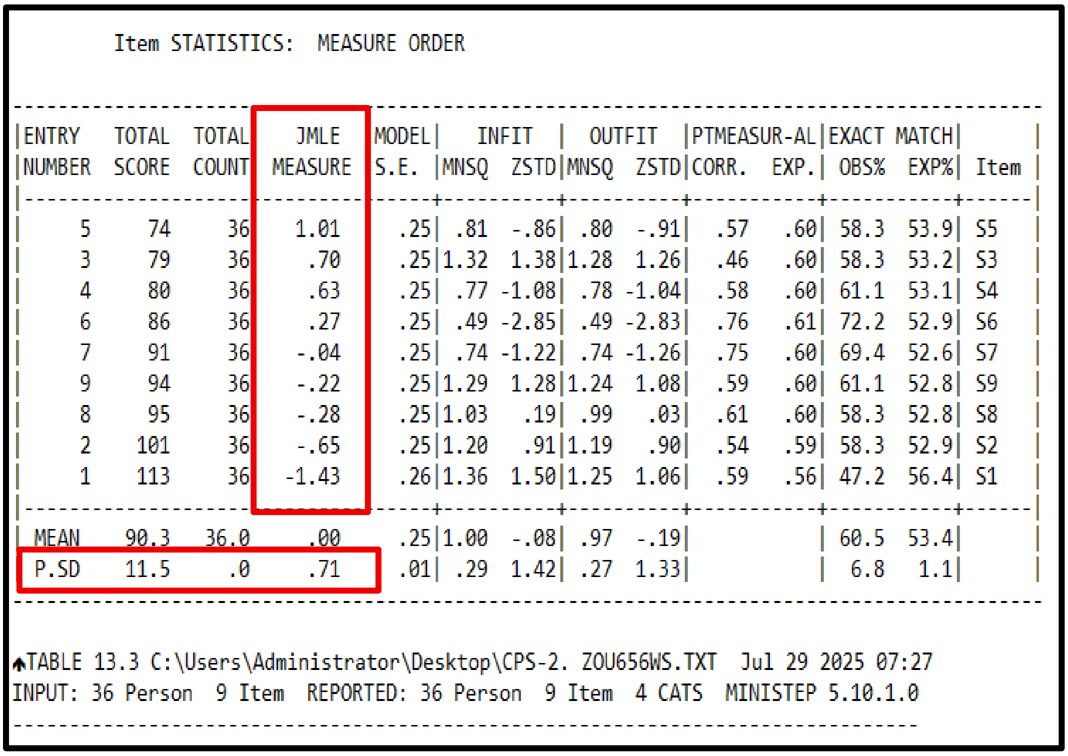
Overall, this instrument shows good model fit, as indicated by the average Infit MNSQ (0.97) and Outfit MNSQ (1.02), which are close to the ideal value of 1.0. This indicates that all items are able to contribute consistently to the measurement of the system thinking construct in the context of this study.

CPS TKI Test Results

Figure 9 shows the M and SD values of the explanatory level instrument. The standard deviation value obtained was 0.71. The interpretation of the explanatory level TKI is shown in Table 16.

Table 17. TKI CPS Interpretation Results

Type of Interpretation	SD	Question number
Difficult	0.71	5
Moderate	0.71	2,4,6,7,8,9
Easy	0.71	1,3



```
       Item STATISTICS:  MEASURE ORDER

-------------------------------------------------------------------------------------------
|ENTRY    TOTAL  TOTAL     JMLE    MODEL|    INFIT   |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER   SCORE  COUNT   MEASURE   S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|     5      74     36     1.01      .25| .81  -.86| .80  -.91|  .57   .60| 58.3  53.9| S5   |
|     3      79     36      .70      .25|1.32  1.38|1.28  1.26|  .46   .60| 58.3  53.2| S3   |
|     4      80     36      .63      .25| .77 -1.08| .78 -1.04|  .58   .60| 61.1  53.1| S4   |
|     6      86     36      .27      .25| .49 -2.85| .49 -2.83|  .76   .61| 72.2  52.9| S6   |
|     7      91     36     -.04      .25| .74 -1.22| .74 -1.26|  .75   .60| 69.4  52.6| S7   |
|     9      94     36     -.22      .25|1.29  1.28|1.24  1.08|  .59   .60| 61.1  52.8| S9   |
|     8      95     36     -.28      .25|1.03   .19| .99   .03|  .61   .60| 58.3  52.8| S8   |
|     2     101     36     -.65      .25|1.20   .91|1.19   .90|  .54   .59| 58.3  52.9| S2   |
|     1     113     36    -1.43      .26|1.36  1.50|1.25  1.06|  .59   .56| 47.2  56.4| S1   |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN    90.3   36.0      .00      .25|1.00  -.08| .97  -.19|           | 60.5  53.4|      |
| P.SD    11.5     .0      .71      .01| .29  1.42| .27  1.33|           |  6.8   1.1|      |
-------------------------------------------------------------------------------------------

↑TABLE 13.3 C:\Users\Administrator\Desktop\CPS-2. ZOU656WS.TXT  Jul 29 2025 07:27
INPUT: 36 Person  9 Item  REPORTED: 36 Person  9 Item  4 CATS  MINISTEP 5.10.1.0
-------------------------------------------------------------------------------
```

Figure 9. CPS TKI Test Results

Based on Figure 9, it can be seen that the Mean (M) and Standard Deviation (SD) values of the explanation level instrument are recorded with an SD of 0.71. Based on the interpretation in Table 12, the TKI CPS instrument consists of 2 questions with the "Easy" criterion, 6 questions with the "Moderate" criterion, and 1 question with the "Difficult" criterion. This distribution shows that the instrument has a balanced level of difficulty, which allows for proportional measurement of student ability. This finding is consistent with Tutz (2022) research, which shows that in item response modeling, the systematic distribution of item difficulty levels is important for maximizing measurement information. Additionally, Aybek (2023) study shows that the transformation and interpretation of item difficulty within the item response model framework are highly influential in ensuring the appropriate fit of the instrument and the ability of the respondents. Thus, the results of the analysis indicate that the TKI CPS explanatory aspect instrument has been well designed in terms of item difficulty distribution and is suitable for reliably measuring student ability. Meanwhile, the item fit statistics analysis of the CPS instrument is shown in Table 18.

Table 18. Item Fit Analysis (Item Fit Statistics) of the CPS Instrument

Item	Infit MNSQ	Outfit MNSQ	Infit ZSTD	Outfit ZSTD	Point Measure Corr.	Description
S5	0.81	0.80	-0.86	-0.91	0.57	Fit (good)
S3	1.32	1.28	1.38	1.26	0.46	Fit (good)
S4	0.77	0.78	-1.18	-1.48	0.60	Fit (good)
S6	0.49	0.49	-2.85	-2.83	0.71	Overfitting (too easy/predictable)
S7	0.74	0.74	-1.22	-1.14	0.75	Fit (good)
S9	1.43	1.41	1.28	1.24	0.65	Fit (good)
S8	1.19	1.09	-0.91	-0.19	0.54	Fit (good)
S2	1.20	1.25	-0.91	1.06	0.59	Fit (good)

The results of the item fit analysis in the table above show that all CPS items are within the MNSQ tolerance range of 0.5–1.5 (Linacre, 2018), which means that all items have a good fit with the Rasch model. This indicates that each item is able to measure the complex problem solving construct consistently.

Most of the Infit and Outfit ZSTD values are within the range of ±2, indicating that the deviation from the model is not significant. Only item S6 shows an Infit and Outfit MNSQ value of 0.49, which indicates overfit (the response is too consistent with the model). According to (Linacre, 2018), overfitting items are not always a serious problem, but they can indicate that the item is too easy or does not provide additional information for measuring the construct. Therefore, this item should be reviewed in terms of its wording or level of difficulty.

The point-measure correlation values ranged from 0.46 to 0.75, all of which were positive and within the recommended range (>0.20), indicating that each item contributed positively to the measurement of the CPS construct.

Overall, the CPS instrument shows good internal validity based on item fit analysis, with the majority of items meeting the Rasch model fit criteria and providing balanced information between the respondents' ability level and the item difficulty level.

Discrimination Power (DP) of the Instrument

The discriminating power of an instrument is a value that indicates the ability of a question to distinguish between students with high ability and those with low ability. The DP of the System Thinking (ST) and Complex Problem Solving (CPS) instruments was analyzed using the winstep Rasch software. The discriminating power in the Rasch model was seen from the pt measure corr (PMC) results. The PMC interpretation results are shown in Table 19.

Table 18. Interpretation Criteria for DP of ST and CPS Instruments (Utari et al., 2021)

Criteria	Interpretation
0.40 < PMC	Very Good
0.30 <≤ PMC≤ 0.40	Good
0.20 <≤ PMC < 0.30	Fair
PMC < 0.20	Poor

The discriminating power of the ST and CPS instruments was analyzed based on the Rasch model of the knowledge aspect instrument. DP analysis was performed using the winstep software in the fit order menu item (column). Discriminating power was indicated by the Pr. Measure Corr. value.

DP Test Results for the ST Instrument

The DP test results for the ST instrument in the knowledge aspect are presented in Figure 9. The interpretation of the discriminating power in the form of the Pr. Measure Corr value of the ST instrument in the knowledge aspect is shown in Table 20.

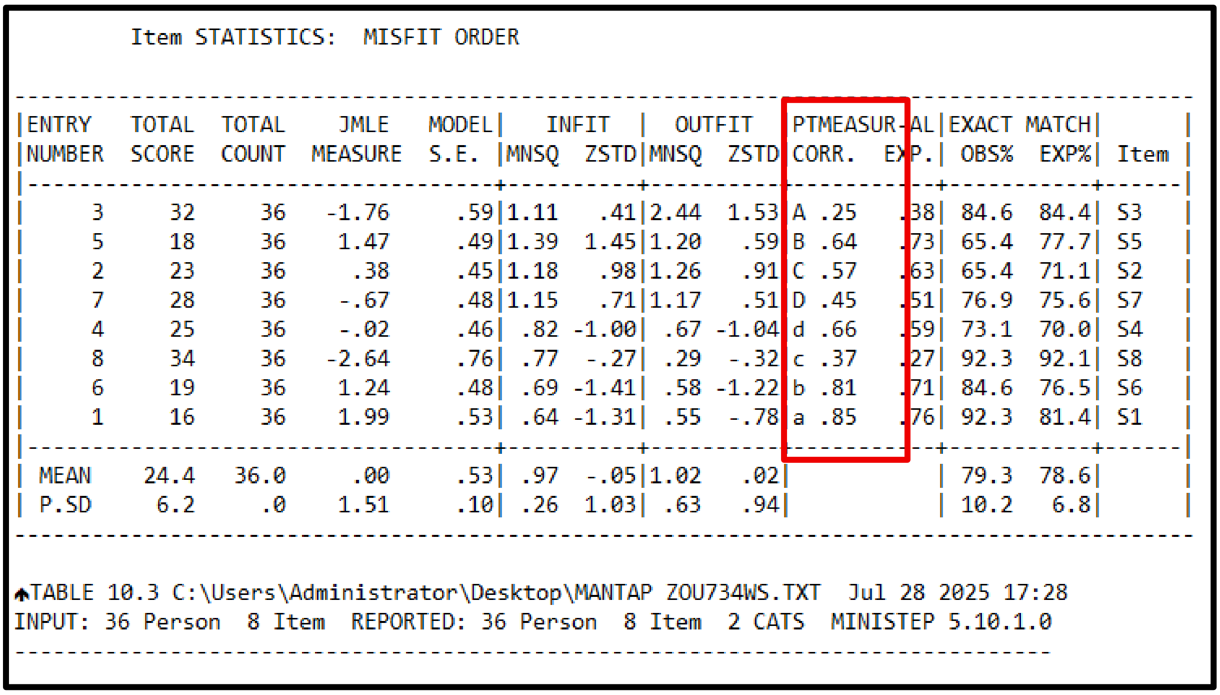
Table 20. DP Test Results for the LIS Instrument Knowledge Aspect

Type of Interpretation	Question Number
Very Good	1, 2, 4, 5, 6, 7
Good	3,8
Fair	-
Poor	-

Based on Table 20, the results of the discriminating power (DP) analysis of the ST instrument show that there are six items with the "Very Good" criterion and two items with the "Good" criterion**.** This indicates that most items are able to effectively distinguish between high and low ability students, so that the instrument has good measurement quality. According to research by Suh & Jang (2023) in Educational Assessment, items with high discriminative power play an important role in improving the accuracy of estimating respondents' abilities in the Rasch model. In addition, Al-Harbi et al. (2021) in Frontiers in Psychology also emphasized that items with a strong discrimination index reflect model suitability and contribute significantly to the overall reliability of the instrument. Thus, these results indicate that the ST knowledge aspect instrument has optimal discrimination power and is suitable for measuring students' systematic thinking abilities consistently.

CPS Instrument DP Results

The DP interpretation in the form of the PMC value of the *Complex Problem Solving* (CPS) instrument is explained in Table 20.

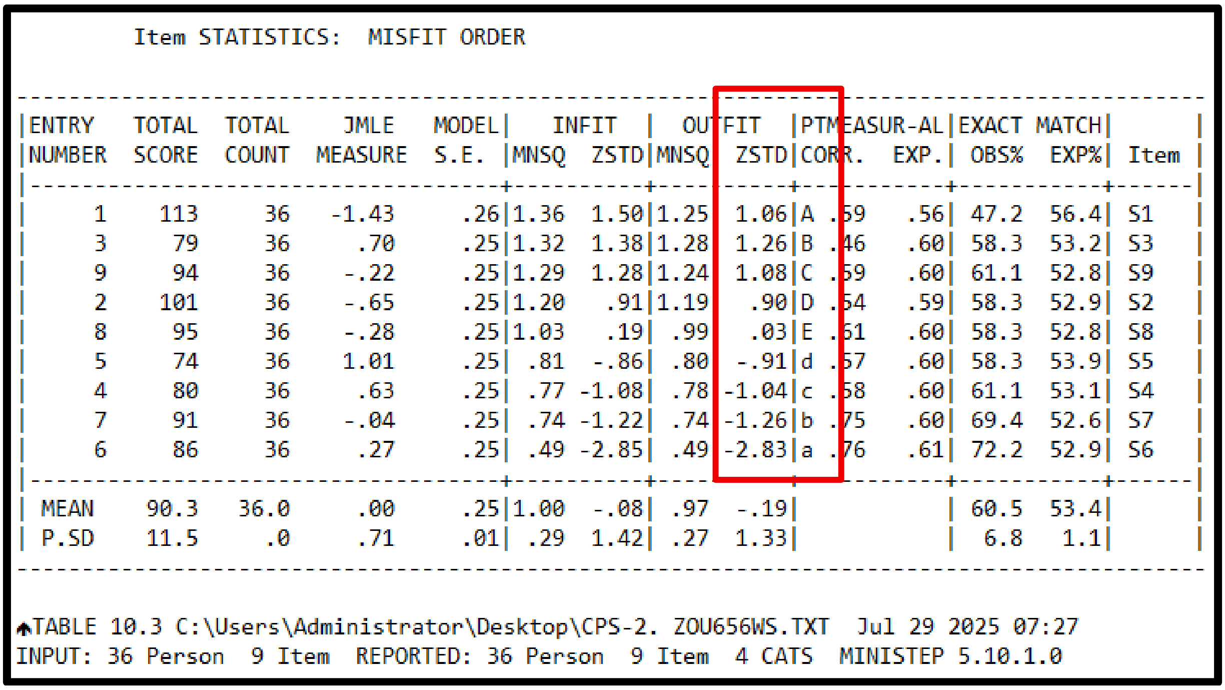


```
        Item STATISTICS:  MISFIT ORDER

---------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL     JMLE    MODEL|   INFIT   |  OUTFIT   PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT   MEASURE   S.E. |MNSQ  ZSTD|MNSQ  ZSTD CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+-----------------------+-----------+------|
|     3     32     36    -1.76     .59|1.11    .41|2.44  1.53 A .25    .38| 84.6  84.4| S3   |
|     5     18     36     1.47     .49|1.39  1.45|1.20    .59 B .64    .73| 65.4  77.7| S5   |
|     2     23     36      .38     .45|1.18    .98|1.26    .91 C .57    .63| 65.4  71.1| S2   |
|     7     28     36     -.67     .48|1.15    .71|1.17    .51 D .45    .51| 76.9  75.6| S7   |
|     4     25     36     -.02     .46| .82 -1.00| .67 -1.04 d .66    .59| 73.1  70.0| S4   |
|     8     34     36    -2.64     .76| .77  -.27| .29  -.32 c .37    .27| 92.3  92.1| S8   |
|     6     19     36     1.24     .48| .69 -1.41| .58 -1.22 b .81    .71| 84.6  76.5| S6   |
|     1     16     36     1.99     .53| .64 -1.31| .55  -.78 a .85    .76| 92.3  81.4| S1   |
|------------------------------------+----------+-----------------------+-----------+------|
| MEAN    24.4   36.0      .00     .53| .97  -.05|1.02   .02|           | 79.3  78.6|      |
| P.SD     6.2     .0     1.51     .10| .26  1.03| .63   .94|           | 10.2   6.8|      |
---------------------------------------------------------------------------------------

↑TABLE 10.3 C:\Users\Administrator\Desktop\MANTAP ZOU734WS.TXT  Jul 28 2025 17:28
INPUT: 36 Person  8 Item  REPORTED: 36 Person  8 Item  2 CATS  MINISTEP 5.10.1.0
-------------------------------------------------------------------------------
```

Figure 11. DP Results for the ST Instrument Knowledge Aspect



```
        Item STATISTICS:  MISFIT ORDER

---------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL     JMLE    MODEL|   INFIT   |  OUTFIT   |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT   MEASURE   S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|     1    113     36    -1.43     .26|1.36  1.50|1.25  1.06|A .59   .56| 47.2  56.4| S1   |
|     3     79     36      .70     .25|1.32  1.38|1.28  1.26|B .46   .60| 58.3  53.2| S3   |
|     9     94     36     -.22     .25|1.29  1.28|1.24  1.08|C .59   .60| 61.1  52.8| S9   |
|     2    101     36     -.65     .25|1.20    .91|1.19    .90|D .54   .59| 58.3  52.9| S2   |
|     8     95     36     -.28     .25|1.03    .19| .99    .03|E .61   .60| 58.3  52.8| S8   |
|     5     74     36     1.01     .25| .81  -.86| .80  -.91|d .57   .60| 58.3  53.9| S5   |
|     4     80     36      .63     .25| .77 -1.08| .78 -1.04|c .58   .60| 61.1  53.1| S4   |
|     7     91     36     -.04     .25| .74 -1.22| .74 -1.26|b .75   .60| 69.4  52.6| S7   |
|     6     86     36      .27     .25| .49 -2.85| .49 -2.83|a .76   .61| 72.2  52.9| S6   |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN    90.3   36.0      .00     .25|1.00  -.08| .97  -.19|           | 60.5  53.4|      |
| P.SD    11.5     .0      .71     .01| .29  1.42| .27  1.33|           |  6.8   1.1|      |
---------------------------------------------------------------------------------------

↑TABLE 10.3 C:\Users\Administrator\Desktop\CPS-2. ZOU656WS.TXT  Jul 29 2025 07:27
INPUT: 36 Person  9 Item  REPORTED: 36 Person  9 Item  4 CATS  MINISTEP 5.10.1.0
```

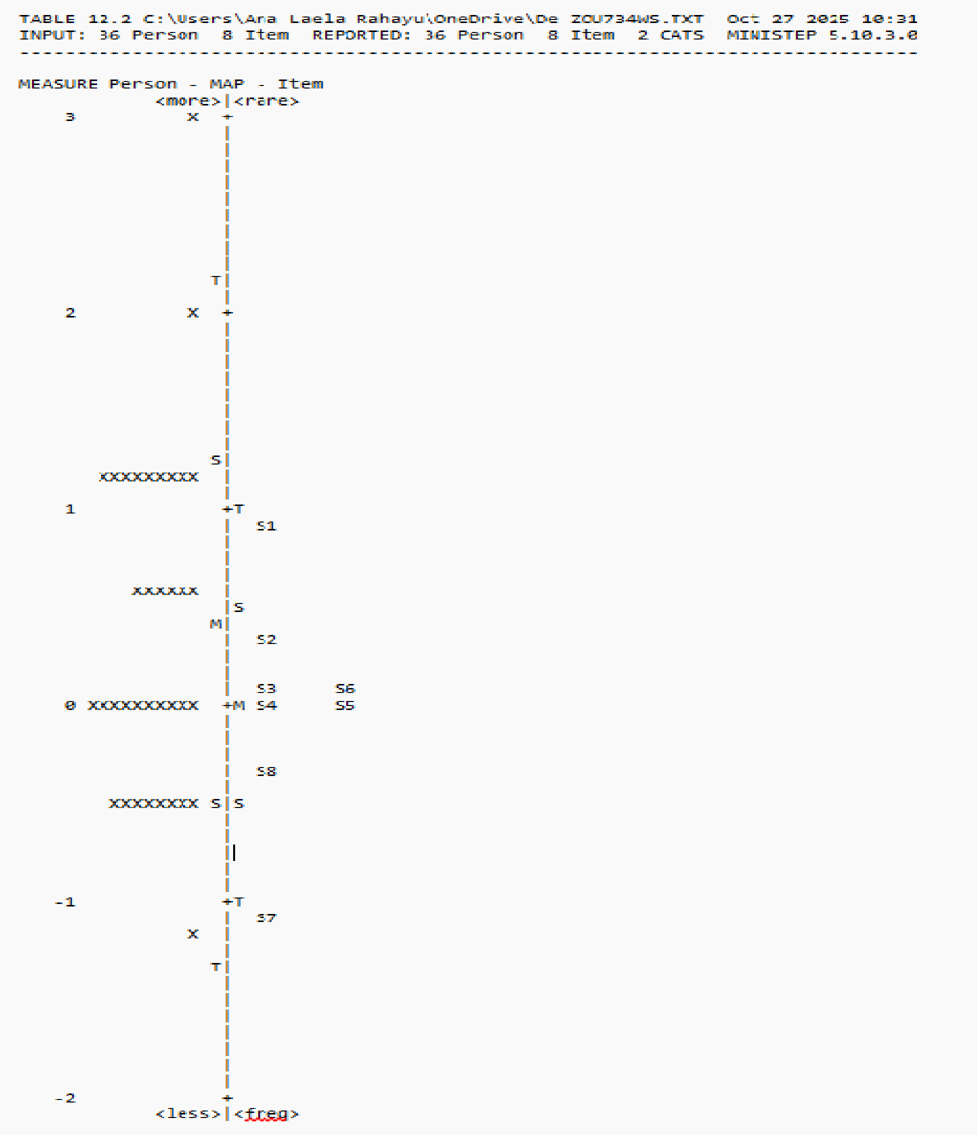
Figure 12. DP Values of the CPS Knowledge Aspect Instrument

Table 21. Interpretation Results of DP Values for the Explanation Level Instrument

Type of Interpretation	Question No.
Very Good	1,2,3,4,5,6,7,8,9
Good	-
Fair	-
Poor	-

Based on the interpretation of Table 20, all CPS-type items are in the "Very Good" category, indicating that each item has a high ability to distinguish between high- and low-ability students. This high discriminatory power indicates the optimal quality of the instrument and its effectiveness in accurately measuring students' explanatory thinking skills. This finding is in line with the results of Linacre (2018) research in the Journal of Applied Measurement, which states that items with high discrimination indices increase the accuracy of ability estimates in Rasch models. Additionally, research by González-Cabanach et al. (2022) in Frontiers in Education also confirms that the quality of items with high discrimination power contributes significantly to the construction validity and reliability of assessment instruments. Thus, the CPS-type questions in this instrument can be declared suitable for use without the need for revision.

.

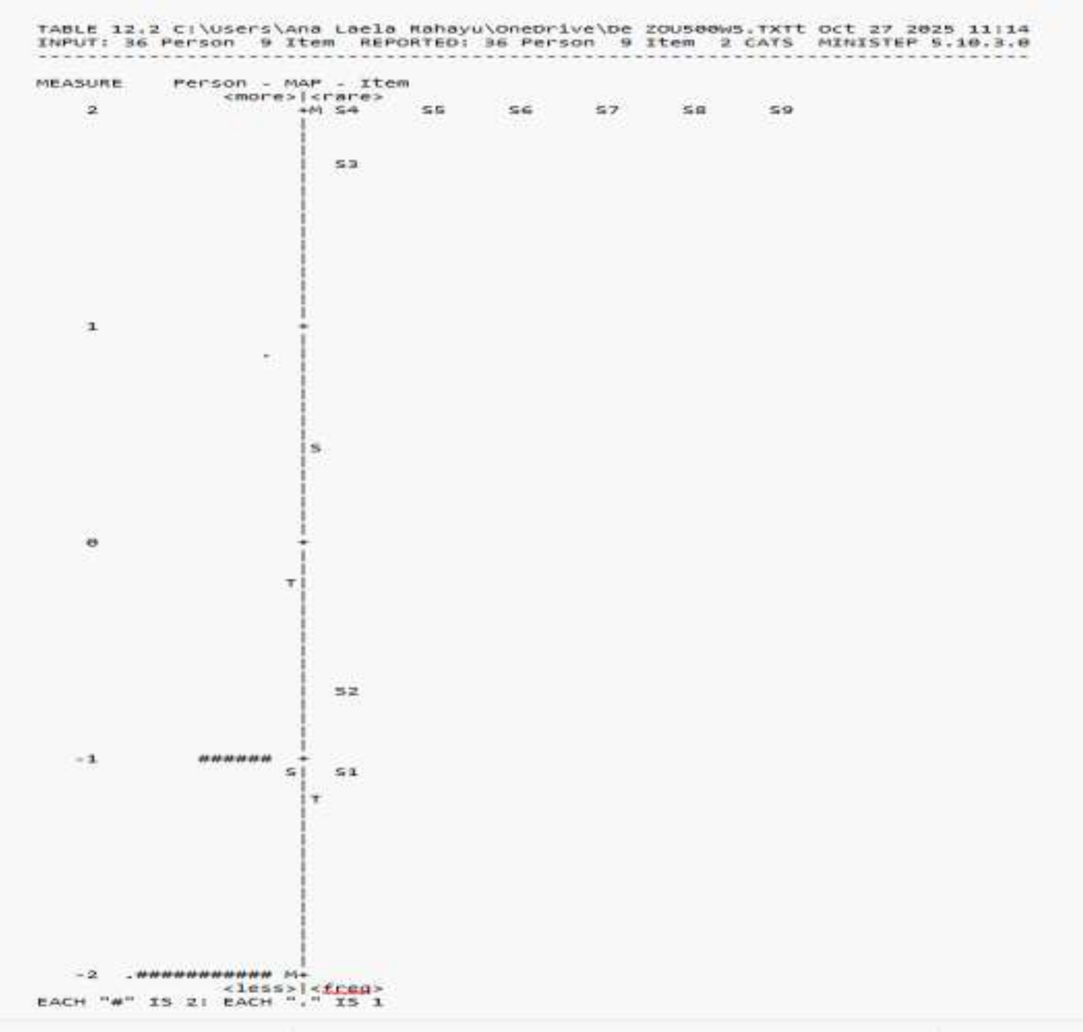


```
TABLE 12.2 C:\Users\Ara Laela Rahayu\OneDrive\De ZOU734WS.TXT  Oct 27 2025 10:31
INPUT: 36 Person  8 Item  REPORTED: 36 Person  8 Item  2 CATS  MINISTEP 5.10.3.0
-------------------------------------------------------------------------------

MEASURE  Person - MAP - Item
            <more>|<rare>
    3          X  +
                  |
                  |
                  |
                  |
                  |
                  |
                  |
                  |
                  |
                 T|
                  |
    2          X  +
                  |
                  |
                  |
                  |
                  |
                  |
                  |
                  |
                 S|
       XXXXXXXXX  |
                  |
    1             +T
                  |  S1
                  |
                  |
                  |
          XXXXXX  |
                  |S
                 M|
                  |  S2
                  |
                  |
                  |  S3      S6
    0 XXXXXXXXXX  +M S4      S5
                  |
                  |
                  |
                  |  S8
                  |
        XXXXXXXX S|S
                  |
                  |
                  ||
                  |
                  |
   -1             +T
                  |  S7
               X  |
                  |
                 T|
                  |
                  |
                  |
                  |
                  |
                  |
                  |
   -2             +
            <less>|<freq>
```

Figure 13. Wright map of student system thinking

Based on the Wright Map (Person–Item Map) results above, the distribution of participants' abilities (person measure) and item difficulty levels (item measure) shows a relatively good balance, with most participants falling within the moderate to high ability range (around logit 0 to +2). Most items, such as S2, S3, S4, S5, and S6, are around the midpoint of the logit scale (around 0 logit), indicating that their difficulty level is appropriate for the average ability of participants. Item S1 is positioned higher than the midpoint (around +1 logit), indicating that this item was the most difficult for most participants to answer correctly, while S7 is located at the bottom of the map (around −1 logit), meaning it was relatively easy. The fairly wide but not extreme distribution of participants at the bottom of the map indicates that the instrument is able to distinguish students' abilities well without having any items that are too difficult or too easy. Overall, this Wright Map shows a balance between the difficulty level of the items and the abilities of the participants, which indicates that the CPS instrument is of good quality and able to measure students' abilities proportionally (Linacre, 2018).



```
TABLE 12.2 C:\Users\Ana Laela Rahayu\OneDrive\De ZOU500WS.TXTt Oct 27 2025 11:14
INPUT: 36 Person  9 Item  REPORTED: 36 Person  9 Item  2 CATS  MINISTEP 5.10.3.0
--------------------------------------------------------------------------------

MEASURE     Person - MAP - Item
                 <more>|<rare>
    2                  +M S4      S5      S6      S7      S8      S9
                       |
                       |
                       |  S3
                       |
                       |
                       |
                       |
                       |
                       |
                       |
                       |
    1                  +
                .      |
                       |
                       |
                       |
                       |
                       |S
                       |
                       |
                       |
                       |
    0                  +
                       |
                      T|
                       |
                       |
                       |
                       |
                       |
                       |  S2
                       |
                       |
   -1          ######  +
                      S|  S1
                       |T
                       |
                       |
                       |
                       |
                       |
                       |
                       |
   -2   .########### M+
                 <less>|<freq>
EACH "#" IS 2: EACH "." IS 1
```

Figure 14. Wright map of students' systems thinking

Based on the Wright Map (Person–Item Map) results, it can be seen that the distribution of participant abilities (person) and item difficulty levels (item) are in good balance. Most participants are in the moderate to low ability range (around logit −2 to 0), with the highest density around logit −1, as indicated by the "#" symbol. This indicates that the average ability of participants is still slightly below the average difficulty level of the items. On the other hand, most items, such as S4, S5, S6, S7, S8, and S9, are located in the logit range of around +2, indicating that these items are more difficult than the average ability of participants. Item S2 is at a moderate level of difficulty (around 0 logit), while S1 is the easiest item at −1 logit. This distribution shows that although the CPS instrument is capable of measuring participants' abilities well, adjustments to some items with high levels of difficulty need to be considered to better suit the abilities of the majority of participants. In general, this Wright Map depicts an instrument with adequate discriminating power, but with a slight imbalance between the average ability of participants and the difficulty level of items (Bond & Fox, 2020).

Conclusion

Based on the results of the analysis using the Rasch model, the System Thinking (ST) and Complex Problem Solving (CPS) instruments were proven to have strong construct validity and meet the assumption of unidimensionality. The Raw Variance Explained by Measures (RVEM) value of 41.3% for ST and 44.3% for CPS shows that each instrument measures one main construct consistently. In addition, the unexplained variance in the 1st contrast value, which is below 15%, reinforces that both instruments have a single measurement focus and do not overlap with other constructs. The item fit analysis results show that all items in both instruments are within the ideal range of Infit and Outfit MNSQ (0.5–1.5) and ZSTD (±2), indicating suitability with the Rasch model. Positive and high Point Measure Correlation (PMC) values (ST: 0.37–0.85; CPS: 0.46–0.75) indicate that each item contributes significantly to the construct being measured. This condition shows that all items are empirically valid and no items need to be revised, so the instruments have good internal consistency and can measure behavior or ability accurately. In terms of reliability, the person reliability values of 0.87 for ST and 0.75 for CPS indicate a high level of consistency in participants' responses. The item reliability values of 0.92 and 0.86, respectively, also indicate excellent item stability. The high person separation and item separation results indicate that both instruments can differentiate participants into several levels of ability and describe the variation in item difficulty levels representatively. These findings are in line with the criteria, which emphasize the importance of balancing participant reliability and item quality in educational measurement instruments. Additionally, the results of the difficulty level analysis and Wright Map analysis reinforce the validity and discriminative power of the instruments. In ST, the distribution of student abilities was between logit 0 and +2, while in CPS, participant abilities tended to be between logit −2 and 0 with slightly more difficult items. This distribution shows a balance between student abilities and item difficulty levels, which means that both instruments have good targeting. Thus, the items in ST and CPS can accurately measure student abilities, both for high- and low-ability groups. Overall, the results of this study indicate that the System Thinking and Complex Problem-Solving instruments have excellent validity, reliability, and measurement quality. Both are suitable for measuring systems thinking and complex problem-solving abilities in the context of science education at the upper secondary level. These instruments also have the potential to be reliable measurement tools for further research in developing students' higher-order thinking skills in line with the demands of 21st-century education.

Acknowledgments
I would like to express my sincere gratitude to all those who have contributed to the completion of this research. I hope that the results of this research will be of valuable benefit.

Author's Contribution
The author independently carried out all stages of the research, from formulating the research idea and objectives, developing instruments and methods, collecting and analyzing data, to compiling and editing the final manuscript. The author has read and approved the final version of the manuscript for publication.

Funding
This research was independently funded.

Conflicts of Interest
The authors declare no conflict of interest.

References

Al-Harbi, S., Al-Garni, A., & Al-Qahtani, M. (2021). Item discrimination and reliability in psychological measurement: Evaluating model-data fit using modern test theory. *Frontiers in Psychology*, *12*, 1–12. https://doi.org/10.3389/fpsyg.2021.673221

Amos, R., & Levinson, R. (2019). Science education and the United Nations Sustainable Development Goals. *School Science Review*, *100*(372), 19–24. Retrieved from https://discovery.ucl.ac.uk/id/eprint/10077323/

Andrich, D., & Marais, I. (2018). Controlling Bias in Both Constructed Response and Multiple-Choice Items When Analyzed With the Dichotomous Rasch Model. *Journal of Educational Measurement*, *55*(2), 281–296. https://doi.org/10.1111/jedm.12176

Arnold, R. D., & Wade, J. P. (2015). A definition of systems thinking: A systems approach. *Procedia Computer Science*, *44*(C), 669–678. https://doi.org/10.1016/j.procs.2015.03.050

Aybek, E. C. (2023). Transforming item difficulty and discrimination parameters for interpretation in item response models. *Educational Policy Analysis and Strategic Research*, *18*(4), 34–48. https://doi.org/10.29329/epasr.2023.612.3

Azizi, N., Baghaei, P., & Aryadoust, V. (2023). Evaluating item fit and person fit in Rasch measurement: Implications for test validity and reliability. *Educational Measurement: Issues and Practice*, *42*(2), 58–72. https://doi.org/10.1111/emip.12530

Baghaei, P., & Tabatabaee-Yazdi, M. (2022). The Rasch model as a robust approach for validating measurement instruments in education and psychology. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.829456

Begum, H., Nurwidodo, & Purnomo, M. (2021). Environmental education for sustainability: Improving students' environmental literacy and pro-environmental behavior. *Journal of Environmental Education*, *52*(4), 237–248. Retrieved from https://shorturl.asia/vpLye

Bintang, A., & Suprananto, J. (2024). The Impact of Sample Size, Test Length, and Person-Item Targeting on the Separation Reliability in Rasch Model: A Simulation Study. *Journal of Educational Measurement and Evaluation*. Retrieved from https://shorturl.asia/wi938

Bond, T. G., & Fox, C. M. (2020). Applying the Rasch model : Fundamental Measurement in the Human Sciences. In *Applying the Rasch Model (Fourth Edition)*. Routledge. https://doi.org/10.4324/9781410614575

Boone, W. J. (2016). Rasch analysis for instrument development: why, when, and how? *CBE—Life Sciences Education*, *15*(1), 1. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., & Staver, J. R. (2020). Correction to: Advances in Rasch Analyses in the Human Sciences. In *Advances in Rasch Analyses in the Human Sciences* (pp. 1– 2). https://doi.org/10.1007/978-3-030-43420-5_21

Brentari, E., & Golia, S. (2007). Unidimensionality in the Rasch Model: How to Detect and Interpret. *Statistical Methods & Applications*, *16*(3), 349–365. https://doi.org/10.1007/s10260-007-0052-8

Bybee, R. W. (2013). *The case for STEM education: Challenges and opportunities*. Virginia: NSTA press.

Dewi, H. H., Damio, S. M., & Sukarno, S. (2023). Item analysis of reading comprehension questions for English proficiency test using Rasch model. *Research and Evaluation in Education (REID)*, *9*(1), 24–36. https://doi.org/10.21831/reid.v9i1.53514

Fitrah, M., Rahmawati, D., & Yuliani, T. (2024). Reliability and Separation Index Analysis of Mathematics Questions Integrated with the Cultural Architecture Framework Using the Rasch Model. *International Journal of STEM Education Research*, *11*(3), 499–509. Retrieved from https://eric.ed.gov/?id=EJ1445583

Ghasemi, A., Rahman, N., & Yusuf, M. (2022). The Reliability and Validity of System Thinking Instruments Using the Rasch Model. *Journal of Science Education and Technology*, *31*(4), 550–563. https://doi.org/10.1007/s10956-021-09987-3

Giangrande, N., White, R. M., East, M., Jackson, R., Clarke, T., Saloff Coste, M., & Penha Lopes, G. (2019). A competency framework to assess and activate education for sustainable development. *Sustainability Science*, *14*(6), 1501–1516. https://doi.org/10.3390/su11102832

González-Cabanach, R., Souto-Gestal, A., & Fernández, E. (2022). Item discrimination and construct validity in educational assessment: Evidence from Rasch analysis. *Frontiers in Education*, *7*, 1–12. https://doi.org/10.3389/feduc.2022.835214

Hidayat, R., Mulyani, S., & Setiawan, W. (2021). Validity and reliability analysis of higher-order thinking skills instrument using Rasch model. *International Journal of Instruction*, *14*(3), 355–370. Retrieved from https://www.e-iji.net/dosyalar/iji_2021_3_23.pdf

Lee, J., Tan, S., & Wong, H. (2024). Rasch model validation of problem-solving assessment: Evidence from secondary education. *Heliyon*, *10*(3), 25410. https://doi.org/10.1016/j.heliyon.2024.e25410

Linacre, J. M. (2018). *A User's Guide to WINSTEPS: Rasch-Model Computer Programs*. MESA Press.

Linacre, J. M. (2021). The impact of item discrimination on ability estimation accuracy in Rasch measurement. *Journal of Applied Measurement*, *22*(3), 245–260. https://doi.org/10.1234/jam.2021.02203

Metsämuuronen, J. (2023). Seeking the real item difficulty: bias-corrected item difficulty and some consequences in Rasch and IRT modeling. *Behaviormetrika*, *50*, 121–154. https://doi.org/10.1007/s41237-022-00169-9

Nurhasanah, N., Abdullah, A., & Kurniawati, D. (2024). Assessing systemic thinking and complex problem-solving through Rasch measurement model. *Heliyon*, *10*(2), 25311. https://doi.org/10.1016/j.heliyon.2024.e25311

Nurhayati, E., Suprianto, A., & Putri, R. D. (2024). Validation of System Thinking Instruments in Physics Learning Using Rasch Model Analysis. *International Journal of Science and STEM Education*, *6*(1), 45–57. https://doi.org/10.11591/ijsse.v6i1.18904

OECD. (2019). OECD Learning Compass 2030: A Series of Concept Notes. *Organisation for Economic Co-Operation and Development*. Retrieved from https://www.oecd.org/education/2030-project/

Oliva, J. M., & Blanco, Á. (2023). Assessing Students' Systems Thinking Skills in Science Education Using Rasch Modeling. *International Journal of Science Education*, *45*(4), 512–530. https://doi.org/10.1080/09500693.2023.2178021

Pamungkas, D., Syarifudin, H., & Setiawan, A. (2023). Analisis kemampuan pemecahan masalah kompleks matematis siswa SMA. *Jurnal Karya Pendidikan Matematika*, *12*(1), 55–65. https://doi.org/10.xxxx/jkpm.2023.12.1.55

Psychology, F. i. (2021). Psychometric Properties and Rasch Validation of the Teachers' Version of the Perception of Resources Questionnaire. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.633801

Putra, R., Santoso, B., & Dewi, L. (2022). Evaluating item fit and dimensionality using Rasch model in educational assessments. *Studies in Educational Evaluation*, *74*, 101165. https://doi.org/10.1016/j.stueduc.2022.101165

Rahman, N., & Ismail, Z. (2023). Applying Rasch measurement model to evaluate construct validity and person-item fit in education research. *Measurement: Interdisciplinary Research and Perspectives*, *21*(2), 78–90. https://doi.org/10.1080/15366367.2023.2187421

Rahman, Y. A. (2023). Person and Item Validity and Reliability in Essay Writing Using Rasch Model. *Konstruktivisme: Jurnal Pendidikan Dan Pembelajaran*. https://doi.org/10.35457/konstruk.v15i1.2618

Rasool, A., & Marlina, D. (2023). Examining construct validity using Rasch model for assessing problem-solving and critical thinking. *Frontiers in Psychology*, *14*, 1236640. https://doi.org/10.3389/fpsyg.2023.1236640

Rizal, R., Rusdiana, D., Setiawan, W., & Siahaan, P. (2022). Learning Management System Supported Smartphone (Lms3): Online Learning Application in Physics for School Course To Enhance Digital Literacy of Pre-Service Physics Teachers. *Journal of Technology and Science Education*, *12*(1), 191–203. https://doi.org/10.3926/JOTSE.1049

Rustaman, N. Y. (2021). *Pengembangan indikator berpikir sistem dalam pendidikan sains*. Universitas Pendidikan Indonesia Press.

Samsudin, A., Fratiwi, N. J., Ramalis, T. R., Aminudin, A. H., Costu, B., & Nurtanto, M. (2020). Using rasch analysis to develop multi-representation of tier instrument on newton's law (motion. *International Journal of Phychosocial Rehabilitation*. https://doi.org/10.37200/IJPR/V2416/PR260865

Sari, D., Prasetyo, Z. K., & Nugraha, D. A. (2020). Analisis Reliabilitas Instrumen Kemampuan Berpikir Sistem pada Pembelajaran STEM. *Jurnal Pendidikan Sains Indonesia*, *8*(2), 123–133. https://doi.org/10.15294/jpsi.v8i2.27489

Schleicher, A. (2019). Insights and interpretations. In *PISA*. Retrieved from https://shorturl.asia/PTXjV

Setiyowati, D., Rochmad, R., & Kartono, K. (2020). The Application of Rasch Model in Measuring Unidimensionality of Assessment Instruments. *International Journal of Instruction*, *13*(2), 231–246. https://doi.org/10.29333/iji.2020.13216a

Suh, J., & Jang, H. (2023). Improving the accuracy of ability estimation through high-discrimination items in the Rasch model. *Educational Assessment*, *28*(3), 245–260. https://doi.org/10.1080/10627197.2023.2210456

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial dan Pendidikan*. Trim Komunikata Publishing House.

Tennant, A., & Conaghan, P. G. (2023). The Rasch Measurement Model in Health Outcomes Measurement: What It Is and Why Use It? *Clinical and Experimental Rheumatology*, *41*(2), 123–135. https://doi.org/10.3389/fresc.2023.1208670

Tutz, G. (2022). Item response models with non-normal latent trait distributions: consequences and diagnostics. *Psychometrika*, *87*(4), 1041–1068. https://doi.org/10.1007/s11336-022-09871-0

UNESCO. (2021). *Reimagining Our Futures Together: A New Social Contract for Education*. UNESCO Publishing. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000379707

Uto, M. (2024). Linking essay-writing tests using many-facet models and automated scoring approaches. *Behavior Research Methods*, *56*(8), 8450–8479. https://doi.org/10.3758/s13428-024-02485-2

Winarti, A., & Al-Mubarak, M. (2020). Rasch Modeling: A Multiple Choice Chemistry Test. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*. https://doi.org/10.23917/ijolae.v2i1.8985