GAWI
Journal of Action Research
Vol. 6 No. 1 2026

Exploring Computational Thinking Skills in Pre-Service Science Teachers: A Rasch Model Analysis

Riskawati[1]*, Nurul Mutmainnah Herman[1], Dirgah Kaso Sanusi[1], Ihfa Indira Nurnaifah Idris[1], Abdurrahman[1], Hendra B.[2], and Sitti Rahma Yunus[3]
[1]Physics Education/Faculty of Mathematics and Science Education,
Universitas Negeri Makassar, Makassar, Indonesia
[2]Science Education, Faculty of Education, IAI Yapnas Jeneponto, Jeneponto, Indonesia
[3]School of Education, The University of Queensland, Australia
*riskawati@unm.ac.id

Received	Accepted	Published
15/05/2026	10/06/2026	11/06/2026

DOI
10.59329/gawi.v6i1.286

ABSTRACT
Computational Thinking (CT) has been increasingly recognized as a key competency in science education, especially in preparing future educators to foster 21st-century skills such as problem-solving and digital literacy. However, limited CT instruction in undergraduate programs has resulted in inadequate readiness among pre-service science teachers, particularly in contexts like Indonesia. This study addresses the gap in the evaluation of CT skills among pre-service teachers by using robust psychometric approaches. This article investigates the psychometric properties of a CT assessment tool, examines the CT ability levels of pre-service science teachers, and evaluates the presence of gender-based item bias. The study employed a quantitative, cross-sectional design utilizing Rasch model analysis to assess instrument reliability, item fit, and differential item functioning (DIF) across gender. A total of 419 pre-service science teachers from two universities in Indonesia were selected using stratified random sampling to ensure proportional representation across academic levels. Data collection and analysis: Data were collected using the 28-item Computational Thinking Test (CTt) and analyzed using Winsteps software for Rasch modeling, including item/person reliability, separation, fit statistics, Wright maps, and DIF analysis. Results indicated high item reliability and good model fit, though person reliability was relatively low, suggesting limited ability discrimination. Several items exhibited gender-based DIF, with some favoring males and others favoring females. The CTt shows strong potential for measuring CT, though refinement is needed to improve its ability to differentiate between skill levels and ensure gender fairness. CT is increasingly recognized as an essential competency in science education to support problem-solving, analytical reasoning, and 21st-century digital literacy skills. However, empirical evidence regarding the measurement quality of CT instruments among pre-service science teachers in Indonesia remains limited. This study aimed to evaluate the psychometric properties of the CTt, describe the distribution of CT skills among pre-service science teachers, and examine gender-based item bias using the Rasch model. A quantitative cross-sectional design was employed involving 419 pre-service science teachers from two Indonesian universities selected through stratified random sampling. Data were analyzed using Winsteps software to examine reliability, separation index, item fit, Wright map distribution, and DIF. The results showed that the CTt demonstrated high item reliability and good model fit, indicating that the instrument is appropriate for measuring CT skills. However, person reliability was relatively low, suggesting limited sensitivity in distinguishing individual ability levels. Wright map analysis indicated that most respondents were distributed at a moderate ability level. In addition, DIF analysis identified several items with potential gender bias, favoring both male and female respondents. Overall, the CTt shows strong potential as an instrument for measuring CT skills, although further refinement is needed to improve measurement sensitivity and ensure gender fairness. These findings contribute to the development of Rasch model–based CT assessment in pre-service science teacher education.
Keywords: Computational thinking; Gender bias; Item analysis; Pre-service science teachers

How to cite:
Riskawati, R., Herman, N. M., Sanusi, D. K., Idris, I. I. N., Abdurrahman, A., B. H., & Yunus, S. R. (2026). Exploring computational thinking skills in pre-service science teachers: A rasch model analysis. *Gawi: Journal of Action Research*, *6*(1), 1-14.

INTRODUCTION

Computational Thinking (CT) has emerged as a fundamental cognitive competency in modern education, particularly in science instruction. Defined as a problem-solving process involving decomposition, pattern recognition, abstraction, and algorithmic thinking (Liu et al., 2023; Wing, 2008), CT supports the development of analytical reasoning crucial in the 21st century. The Next Generation Science Standards (NGSS) emphasize CT integration to enhance scientific inquiry and learning outcomes (Aminger et al., 2021; Lee et al., 2020). Within teacher education, especially for pre-service science teachers, CT is increasingly regarded as vital for equipping future educators with the capacity to foster students' problem-solving and digital literacy skills (Connolly et al., 2021; Peters-Burton et al., 2022).

Despite global recognition, the implementation of CT in teacher education remains inconsistent. Many undergraduate science programs, particularly in physics, offer limited exposure to computational instruction (Caballero & Merner, 2018; Lyon & J. Magana, 2020). Pre-service teachers often lack self-efficacy and pedagogical strategies to apply CT effectively in classroom settings (Boulden et al., 2021). In Indonesia, although CT has been formally integrated into the national curriculum (Marifah, 2022), barriers such as limited teacher training, students' mathematical abilities, and lack of evaluation frameworks hinder its effectiveness (Ismarmiaty et al., 2022; Kumala et al., 2023). This underscores the urgent need to assess and strengthen CT competencies in teacher preparation.

The main problem addressed in this study is the insufficient integration of computational thinking into the pedagogical practices of pre-service science teachers. Despite its prominence in educational standards and curricula, CT remains poorly implemented due to gaps in training, instructional design, and self-efficacy. As a general solution, this study proposes using Rasch analysis to measure CT skills more precisely, identify proficiency levels, and provide insights for developing targeted instructional strategies.

Rasch modeling is increasingly employed in educational research to evaluate latent traits such as problem-solving and cognitive competencies with precision. It provides valid, reliable, and bias-controlled insights into how individuals engage with specific tasks (Chan et al., 2021). By applying Rasch analysis, educators and researchers can classify CT skill levels and examine the distribution of competency across various demographic and instructional variables. This methodology supports the development of more effective, data-driven CT instruction tailored to individual needs.

Several pedagogical interventions have proven successful in enhancing CT among pre-service teachers. For example, Peel et al. (2022) advocate for unplugged, hands-on science activities, while Odden & Burk (2020) highlight the role of computational essays in strengthening CT application. Research by Dong et al. (2023) and Verawati et al. (2023) demonstrates that integrating CT with constructivist learning approaches significantly improves engagement and comprehension. These instructional strategies, when supported by robust assessment frameworks like Rasch modeling, can bridge the gap between theoretical understanding and practical application of CT.

Recent studies show that while pre-service teachers are aware of CT's importance, many struggle with its classroom implementation due to inadequate computational training and unclear pedagogical pathways (Caballero & Merner, 2018; Lyon & J. Magana, 2020). Furthermore, the integration of CT in STEM education often overlooks the psychological dimensions of teaching, such as self-efficacy and motivation (Boulden et al., 2021; Connolly et al., 2021). Without addressing these factors, instructional reforms may fall short of equipping teachers to promote CT effectively.

In Indonesia, the challenge is compounded by systemic issues such as uneven teacher competency, lack of assessment tools, and contextual barriers in curriculum execution (Ismarmiaty et al., 2022; Prastowo & Fitriyaningsih, 2020). Although efforts to embed CT in the Informatics curriculum have begun, studies report that most educators are not equipped with practical knowledge to implement or evaluate CT (Wardani et al., 2023). These findings reveal a critical gap in both teacher preparation and evaluation frameworks, necessitating empirical studies that combine psychometric analysis and educational intervention.

This study aims to evaluate the computational thinking skills of pre-service science teachers using Rasch analysis to ensure an objective and valid assessment. The novelty lies in applying this psychometric model within the context of teacher education to classify proficiency levels and identify influencing factors. The scope includes measuring CT competencies, assessing the reliability of the instrument, and uncovering correlations with pedagogical preparedness. The study offers evidence-based insights for enhancing CT instruction, with implications for curriculum design and teacher training reform, particularly in Indonesia's science education landscape. Our research questions for this study were as follows: RQ1: Is the instrument reliable and valid according to the Rasch measurement model? RQ2: Are there any significant gender-based item biases identified through DIF analysis?

METHODS

Research Design

This study employed a quantitative, cross-sectional research design to examine the CT skills of 419 pre-service science teachers. A quantitative approach was adopted to address the research questions through the administration of cognitive tests specifically developed to assess students' CT abilities. The resulting data were numerical, derived from the total scores of correct responses to individual test items. To collect these data, a diagnostic tool—referred to as the CTt—was utilized. The responses obtained were analyzed using the Rasch measurement model, allowing for a more precise estimation of students' abilities and the quality of the test items. The demographic profile of the participants is presented in Table 1, providing contextual insight into the sample characteristics and supporting the interpretation of the findings.

Table 1 The demographic data of the pre-service science teachers

Gender	Amount	Percentage (%)
Male	78	18.61
Female	341	81.39
Total	419	100

Table 1 shows that a total of 419 pre-service science teachers from two universities in Indonesia participated in this study. Participants were selected using a stratified random sampling technique to ensure proportional representation across academic levels (Neuman, 2014). The selection aimed to assess their CT abilities. All respondents voluntarily joined the study, provided informed consent, and were assured of confidentiality. A coding system was used to anonymize participant identities (e.g., "M" for male, "F" for female). The CT instrument used was originally developed for school-age students in Spain but was applied to university students based on the assumption that they would perform better due to their higher academic level. However, many participants reported confusion during the test, and their scores were unexpectedly low. This outcome suggests a lack of CT exposure in Indonesia's earlier education stages, highlighting a mismatch between instrument expectations and respondents' preparedness (Ghani et al., 2022; Syafril et al., 2022; Tsakeni, 2021).

Instrument

The CTt underwent a rigorous content validation process through expert judgment, resulting in a refined version comprising 28 items (Aminger et al., 2021). The CTt is designed to assess students' developmental levels in CT, based on an operational definition that emphasizes the

ability to formulate and solve problems using core computing principles and programming logic. These principles include basic sequences, loops, iteration, conditionals, functions, and variables. The computational thinking test indicators are listed in Table 2.

Table 2 Indicator of computational thinking test

No.	Indicator	Number of Items
1	Basic Directions and Sequences (BDS)	1-4
2	Loops - Repeat Times (RT)	5-8
3	Loops - Repeat Until Condition is Met (RUC)	9-12
4	Simple Conditionals (If)	13-16
5	Complex Conditionals (If/Else)	17-20
6	While Conditionals (WC)	21-24
7	Simple Functions (SF)	25-28

The instrument is a multiple-choice test consisting of four answer options, with only one correct answer, and is intended to be completed within approximately 45 minutes. Each item in the CTt targets one or more of seven computational concepts, arranged by increasing levels of difficulty. Items are presented in one of two interactive environments, “The Maze” or “The Canvas”—which are commonly employed in programming education platforms such as Code.org. Visual elements, such as directional arrows or coding blocks, may appear as part of the answer choices, and the items vary in complexity depending on the presence or absence of nested computational structures. The cognitive demands required to answer the items include sequencing, completion, and debugging tasks. The CTt is administered collectively and delivered online, making it accessible via both mobile and desktop electronic devices. Preliminary psychometric evidence from the instrument's administration to a sample of 400 Spanish students has been reported (González, 2015)). Examples of finalized CTt items, translated into English, are provided in Figure 1, along with detailed specifications for each item.

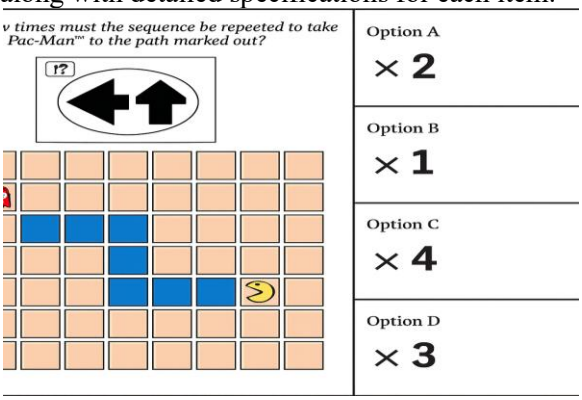

Figure 1 Item Example of instrument

Collection Procedure

This study was conducted over a three-day period at two different universities in Indonesia. Data were collected through the administration of the CTt, which was distributed to students via Google Forms during scheduled computer-based learning sessions. The test was completed using laptops in a computer laboratory setting, with an allocated duration of 45 minutes. Within the Google Form, students were instructed to provide demographic information, including gender, school name, and grade level. Prior to beginning the CTt test, the researcher delivered a brief orientation session, during which students received an explanation regarding the objectives and format of the test. Additionally, three example questions were provided to familiarize students with the structure and characteristics of the items they would encounter. Upon completing the test, students were instructed to press the 'Submit' button to record their responses. The submitted answers were subsequently compiled and analyzed quantitatively to inform the findings of this research.

Data Analysis

This study employed the Rasch analysis model to evaluate data obtained from the CTt. The collected responses were initially organized using Microsoft Excel and subsequently imported into Winsteps software version 5.7.1.0 for detailed Rasch analysis. To validate the CTt instrument, both content validity and internal consistency reliability were assessed. In addition, item reliability, item separation, person reliability, and person separation indices were examined to further ensure the instrument's psychometric robustness. The quality of individual test items was evaluated using three key Rasch-based fit statistics: Outfit Mean Square (MNSQ), Outfit z-standardized (ZSTD), and Point-Measure Correlation (Pt-Measure Corr), as recommended by Boone et al. (2014). To visualize the alignment between item difficulty and respondent ability, a Wright map was generated, providing a comprehensive overview of how well the items target the intended construct. Furthermore, a logit value analysis was conducted to assess the critical thinking abilities of prospective science teachers in Indonesia. Person-fit statistics were also analyzed using the same three criteria (MNSQ, ZSTD, and Pt-Measure Corr) to identify any anomalous response patterns among participants. Finally, DIF analysis was performed using Winsteps to detect any potential item bias based on gender differences, ensuring fairness and validity in the assessment outcomes.

RESULT AND DISCUSSION

This study utilized Rasch model analysis to validate the CTt by examining the suitability of the data with the model, evaluating the CT skills of pre-service science teachers, and identifying any items that may function differently across gender. The results addressing these aims are discussed in the following sections.

RQ1: Is the instrument reliable and valid according to the Rasch measurement model?

The table presents a summary of Rasch model statistics encompassing both person and item parameters derived from an assessment involving 419 respondents and 28 test items. The mean person measure was -0.48 logits, indicating that respondents, on average, performed slightly below the average item difficulty, which is standardized to 0.00. The person standard deviation (SD = 0.69) reflects moderate variation in ability across the sample, and the relatively small standard error (SE = 0.03) suggests adequate precision in the measurement estimates.

However, the person separation index (0.89) and person reliability coefficient (0.44) suggest that the instrument exhibits limited capacity to distinguish individuals based on their ability levels. According to Bond and Fox (2015), separation values below 1.0 indicate an inability to differentiate even two strata of ability within the sample, while reliability coefficients below 0.70 raise concerns regarding the consistency of the assessment in measuring a unified construct. These findings suggest a potential mismatch between item difficulty and the ability level of the sampled respondents. This limitation may be attributed to issues in item targeting, respondent heterogeneity, or instrument design, and should be explicitly addressed in future test development efforts (Elbahan et al., 2023; Syafril et al., 2022; Tsakeni, 2021). Summary statistics of people and goods are shown in Table 3.

In contrast, item-level statistics demonstrate considerably stronger psychometric qualities. The MNSQ of 1.04 with a standard deviation of 0.31 falls within the acceptable range of 0.5 to 1.5, indicating a good fit to the Rasch model (Boone et al., 2014). Item reliability was exceptionally high (0.99), and the item separation index (11.35) confirms excellent discrimination across item difficulties (Bond & Fox, 2015). These results support the conclusion that the items are functioning well to stratify respondent abilities and are structurally coherent in representing the construct being measured.

Table 3 Summary statistics of person and items

	Person	Item
N	419	28
Measure (logit)		
Mean	-0.48	0.00
SD, standard deviation	0.69	1.55
SE, standard error	0.03	0.30
Outfit mean square		
Mean	1.04	1.04
SD	0.58	0.31
Separation	0.89	11.35
Reliability	0.44	0.99
Alpha cronbach	0.46	0.46
Raw variance explained by measures	1.00	-1.00

* item and person outliers dropped from this table

These findings indicate that the items were well distributed across difficulty levels; however, the instrument showed limited effectiveness in differentiating respondent ability levels. However, the raw variance explained by measures for items was reported as -1.00, a value that is not theoretically plausible in Rasch analysis. This anomaly likely reflects a computational or data entry error and warrants further inspection before publication or use of the results.

However, the person separation index (0.89), person reliability coefficient (0.44), and Cronbach's alpha (0.46) indicate that the CTt currently has limited capacity to reliably distinguish respondents across different levels of computational thinking ability. These findings suggest that the instrument may not yet provide sufficiently stable measurement at the person level, potentially due to inadequate item targeting, restricted variability in respondent ability, or weaknesses in item design. Although item-level statistics demonstrated acceptable fit and strong item separation, the low person-level indices represent a substantial limitation of the instrument. Therefore, while the CTt shows potential for assessing computational thinking skills, further refinement is necessary to improve measurement precision, internal consistency, and discriminatory capacity across varying respondent ability levels.

Moreover, the Cronbach's alpha for both person and item responses was relatively low (0.46), indicating weak internal consistency. This suggests that the items may not be cohesively measuring a single latent construct, possibly due to multidimensionality or lack of item alignment with respondent abilities. While this presents a limitation, it also provides a valuable diagnostic insight into the challenges of assessing computational thinking among pre-service science teachers. This represents a substantial limitation of the instrument, indicating that the current version of the CTt may not yet provide sufficiently reliable discrimination of individual CT ability levels, an issue previously noted in the literature (Butler & Leahy, 2020; Fülöp et al., 2022; Zhu & Wang, 2023).

Taken together, these findings indicate that although the items display excellent psychometric functioning. Taken together, these findings suggest that the CTt demonstrates acceptable item-level functioning but limited person-level measurement performance. Although the instrument shows promise for assessing computational thinking, substantial refinement is needed to improve reliability, targeting, and discriminatory capacity. The current results are still suitable for publication, provided that the limitations are transparently discussed, and recommendations for future development such as refining the item pool, improving targeting, or adjusting sample stratification are clearly articulated (Ghani et al., 2022; Kong & Lai, 2023; Tongal et al., 2024). Future development should prioritize improving item targeting, expanding item variability, and strengthening the instrument's ability to distinguish varying respondent ability levels. This transparency not only upholds academic rigor but also contributes constructively to the ongoing discourse on computational thinking assessment in science teacher education.

Table 4 presents the Rasch model summary statistics for the CTt, comprising seven indicators—Basic Sequence and Decomposition (BSD), Repetition/Looping (RT), Recognizing and Using Patterns (RUC), If Statement (If), If/Else Statement (If/Else), While/Condition (WC), and System Functioning (SF)—each containing four items, totaling 28 items. Across all indicators, the mean item Outfit MNSQ values ranged from 0.97 to 1.08, falling within the acceptable range of 0.75 to 1.34, thus indicating a good overall fit to the Rasch model (Boone et al., 2014). Similarly, item Infit MNSQ values were stable, with a mean of 0.99 for the entire test, confirming internal consistency in how respondents engaged with items of varying difficulty. Person Outfit and Infit MNSQ values followed comparable trends, with no values indicating substantial misfit, thereby supporting the construct validity of the CTt.

Table 4 Summary of Rasch parameters for the computational thinking test and for each task.

Psychometrics attribute	BSD	RT	RUC	If	If/Else	WC	SF	CT test
Number of items	4	4	4	4	4	4	4	28
Mean								
item outfit MNSQ	1.04	1.08	0.97	1.06	1.01	0.97	1.00	1.04
item Infit MNSQ	0.99	1.00	0.98	1.01	0.98	1.00	0.96	.99
person outfit MNSQ	1.03	1.08	0.96	1.06	1.01	1.00	1.00	1.04
person Infit MNSQ	0.99	0.96	0.60	0.98	1.00	0.97	1.00	1.01
Item separation	9.70	0.00	6.93	8.48	4.09	7.76	3.96	11.33
Person separation	0.30	8.98	0.00	0.00	0.00	0.00	0.00	.89
Unidimensionality								
Raw variance by measure	2.61	3.48	9.42	2.07	0.83	1.70	0.84	36.2%
Unexplained variance 1st contrast	1.54	1.49	2.05	1.55	1.54	1.72	1.71	5.1%

In terms of reliability, item separation values were strong across most indicators, with the highest separation observed in BSD (9.70), RUC (6.93), and If (8.48), suggesting that these item sets can reliably distinguish among varying levels of difficulty (Bond & Fox, 2015). The total item separation index was 11.33, reinforcing the robustness of the CTt in establishing a hierarchical item structure. However, person separation indices were notably low for all indicators except RT (8.98), with the remaining dimensions recording values close to zero. This implies that, aside from RT, most indicators lacked sufficient spread in person abilities to effectively stratify respondents by their computational thinking skills.

Regarding dimensionality, the raw variance explained by the measures varied between indicators, with RUC (9.42%) and RT (3.48%) showing stronger alignment with the Rasch dimension, whereas If/Else and SF reported notably lower values (0.83% and 0.84%, respectively). Despite this variation, the unexplained variance in the first contrast remained below the critical threshold of 3.0 for all indicators, indicating an acceptable level of unidimensionality and the absence of major secondary dimensions.

To further evaluate the dimensional structure of the CTt, Principal Component Analysis (PCA) of Rasch residuals was conducted. The raw variance explained by the Rasch measures for the total scale was 36.2%, indicating that a substantial proportion of variance was accounted for by the primary construct. The unexplained variance in the first contrast was 5.1%, which remained within an acceptable range for supporting essential unidimensionality. Although the CTt consists of seven conceptual indicators, the residual structure did not reveal a dominant secondary dimension. In addition, Rasch fit statistics across the seven indicators generally fell within acceptable ranges, suggesting that each subdimension functioned adequately as part of the broader CT construct. However, some clustering tendencies among items within the same indicator may suggest potential local dependence, particularly among conceptually similar items. Variations in person separation and variance explained across indicators also indicate that the measurement strength of each subdimension was not entirely uniform. Therefore, while the total CTt score may still be interpreted as representing an overall CT construct, the

multidimensional nature of the subdomains should be considered in future instrument refinement and validation studies.

In summary, the CTt demonstrates acceptable psychometric quality in terms of item fit and separation. However, the instrument shows limited ability to differentiate respondent abilities across most indicators. This outcome suggests the need for further refinement of item content and distribution to ensure better alignment between item difficulty and respondent competence levels. As supported by the use of Wright Maps in similar studies (Ismail et al., 2024), aligning item difficulty with a wider range of respondent abilities is critical for developing more sensitive and effective measurement tools.

Table 5 presents the item-fit analysis of the 28 items in the CTt, based on three key Rasch model criteria: MNSQ, ZSTD, and Pt-Measure Corr. An item is only considered misfitting if all three indicators fall outside acceptable ranges: Outfit MNSQ beyond 0.5- 1.5, ZSTD beyond ±2.0, and Pt-Measure Corr. below 0.20. Item fit was evaluated holistically using Outfit MNSQ, Outfit ZSTD, and Point-Measure Correlation statistics. Particular attention was given to items with substantially elevated Outfit MNSQ values, as these may indicate unpredictable response patterns and potential threats to measurement quality.

Overall, most CTt items demonstrated good model fit. While several items exceeded acceptable limits on one or two criteria, none violated all three simultaneously, except for item RUC3, which reported an Outfit MNSQ of 2.50, a ZSTD of 6.07, and a Pt-Measure Corr. of -0.23. These values indicate that RUC3 may not be functioning consistently with the intended construct and should therefore be considered for revision or exclusion in future applications. Other items such as SF4 and SF1 showed high ZSTD values (2.99 and 2.59, respectively) and elevated Outfit MNSQ (1.29 and 1.21), yet maintained acceptable Pt-Measure Corr. (0.09 and 0.26). Hence, these items are retained but warrant further review to ensure clarity and construct alignment. Similar patterns were found in IF/ELSE4 (ZSTD = 2.68) and WC2 (ZSTD = 0.75), where item performance remained acceptable despite exceeding one threshold.

Items such as BDS3 and BDS4 also had notably high Outfit MNSQ values (2.88 and 3.41) and negative ZSTD values (-2.18 and -3.41, respectively). However, their high Pt-Measure Corr. scores (0.42 and 0.47) support their continued inclusion, suggesting they still measure aspects of the intended construct. Most items demonstrated acceptable fit to the Rasch model; however, several items, particularly RUC3, BDS3, and BDS4, require further evaluation due to elevated misfit statistics that may affect measurement precision. On the lower end of the difficulty scale (logits < -2.0), items like RT1, RT2, BDS1, and RUC1 recorded Outfit and ZSTD values within acceptable ranges, alongside positive Pt-Measure Corr. These items, although easier, remain valid for identifying students with lower ability levels. Conversely, high-difficulty items such as IF2 and WC1 (logits > 2.0) displayed acceptable fit and correlation statistics, indicating their appropriateness for measuring higher-order computational thinking. Item-fit analysis is listed in Table 5.

Table 5 Item-fit analysis

ITEM	Measure (Logit)	Infit MNSQ	Outfit MNSQ	Outfit ZSTD	Ptmeasur-Al Corr.
IF2	2.07	1.01	1.08	0.45	0.13
WC1	2.03	0.96	0.84	-0.86	0.24
RUC3	1.97	1.13	2.5	6.07	-0.23
WC2	1.76	1.01	1.11	0.75	0.16
SF2	1.76	0.99	0.93	-0.4	0.18
SF3	1.76	0.97	0.93	-0.39	0.18
IF/ELSE3	1.28	1.03	1.18	1.5	0.14
RUC2	0.89	1.03	1.02	0.21	0.2
SF4	0.87	1.07	1.29	2.99	0.09
IF/ELSE2	0.76	1.03	1.02	0.25	0.22
IF4	0.68	1.01	1	0.02	0.25
SF1	0.65	1.11	1.21	2.59	0.26
IF/ELSE4	0.52	1.07	1.2	2.68	0.14
IF3	0.41	0.95	0.97	-0.46	0.27

ITEM	Measure (Logit)	Infit MNSQ	Outfit MNSQ	Outfit ZSTD	Ptmeasur-Al Corr.
WC3	0.16	1.04	1.03	0.64	0.29
IF/ELSE1	0	1.08	1.09	1.99	0.29
WC4	-0.14	1	1	0	0.3
RT3	-0.26	0.95	1.73	-0.41	0.36
BDS3	-0.35	0.92	2.88	-2.18	0.42
BDS4	-0.39	0.88	3.41	-3.41	0.47
RUC4	-0.68	1.03	1.04	0.97	0.31
IF1	-0.69	1.04	1.14	1.14	0.34
BDS2	-1.69	0.89	-1.94	-1.98	0.45
RT2	-2.33	0.9	-1.07	-1.92	0.4
RT1	-2.49	0.88	-1.19	-1.65	0.3
RUC1	-2.65	0.89	-1.68	-1.68	0.29
BDS1	-2.83	0.92	-1.86	-1.86	0.23
RT4	-3.07	1.03	0.95	-1.23	0.21

In summary, only RUC3 was identified as a true misfit according to all three Rasch fit criteria and should be carefully examined in future use. All other items either met all thresholds or violated only one or two, and thus are deemed suitable for inclusion. These findings support the psychometric soundness of the CTt, affirming that the majority of items are valid measures of computational thinking ability. In summary, most items demonstrated acceptable fit to the Rasch model. However, RUC3 showed substantial misfit across multiple indicators, while BDS3 and BDS4 also demonstrated elevated Outfit statistics that warrant further review and refinement.

The Wright Map displayed in Figure 2 provides a visual representation of the alignment between respondent abilities and item difficulty levels within the CTt On the left side of the map, the distribution of respondents is represented, with individuals of higher ability located towards the top and those with lower ability toward the bottom. On the right side, test items are plotted based on their difficulty, with more difficult items positioned at the top and easier items at the bottom. This alignment enables a direct comparison of how well item difficulties match the range of abilities in the sample group (Bond, & Fox, 2015).

From the Wright Map, it is evident that the majority of respondents are clustered between logit values of -1 and 1, indicating a moderate level of computational thinking ability. among respondents from the participating universities. However, several items—such as IF2, RUC3, WC1, SF2, SF3, and WC2—are located at the top of the scale (logits 2 to 3), indicating that these items are relatively more difficult and may only be correctly answered by respondents with higher-than-average abilities. Conversely, items such as RT4, BDS1, and RUC1, located at the bottom of the map (logits -3 to -4), are among the easiest and likely answered correctly by most respondents. Notably, some items like IF1, RUC4, and BDS3 align closely with the mean respondent ability (logit ~0), suggesting these are well-targeted for the average test taker. In contrast, the presence of multiple items far above or below the average ability level signals potential misalignment between item difficulty and participant competence, which may affect the precision of measurement for both high- and low-ability individuals.

Overall, the Wright Map reveals a moderately balanced distribution between item difficulties and respondent abilities. However, certain items may require revision or restructuring to ensure more comprehensive coverage across the full spectrum of participant abilities. Items clustered at the higher difficulty levels could be reviewed for clarity or conceptual overload, while items at the lower end may need to be expanded or removed if they no longer contribute meaningful variance. This analysis is essential for refining the instrument to improve its diagnostic precision and pedagogical relevance.

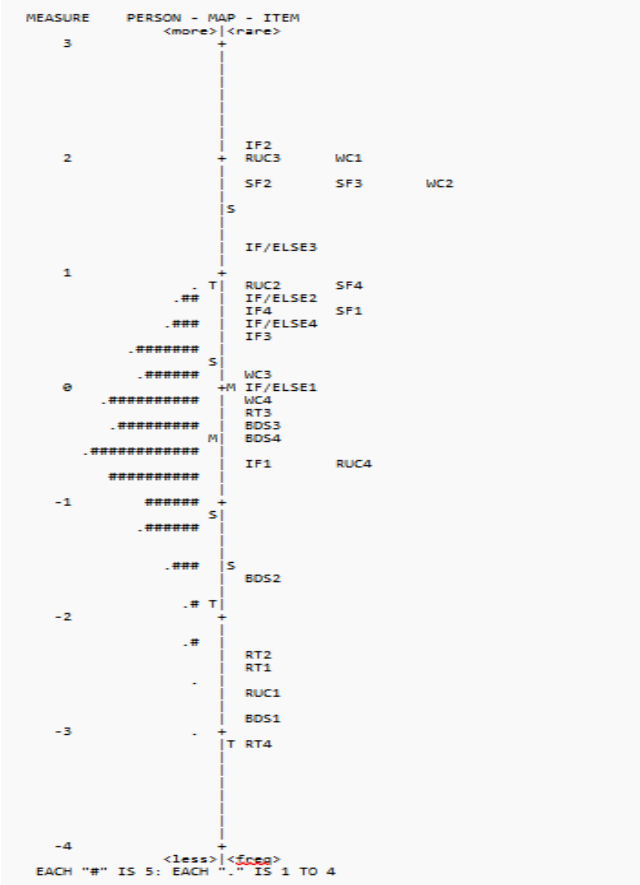

Figure 2 The wright map

RQ2: Are there any significant gender-based item biases identified through DIF analysis?
Figure 3 displays the DIF plot based on gender for a subset of items in the CTt. This analysis examines whether certain items function differently for male and female students of similar overall ability levels. Items that show significant gaps in DIF measures may suggest potential bias or differences in item interpretation across gender groups (Bond & Fox, 2015). In the figure, it is evident that several items were more favorable to one gender over the other. For example, items SF2, SF3, and WC2 show considerably higher DIF values for male students, suggesting that males performed better than females on these items, even when their overall abilities were comparable. The largest gender gap appears in item SF3, where the DIF measure for males peaks significantly, indicating that male students found this item much easier.

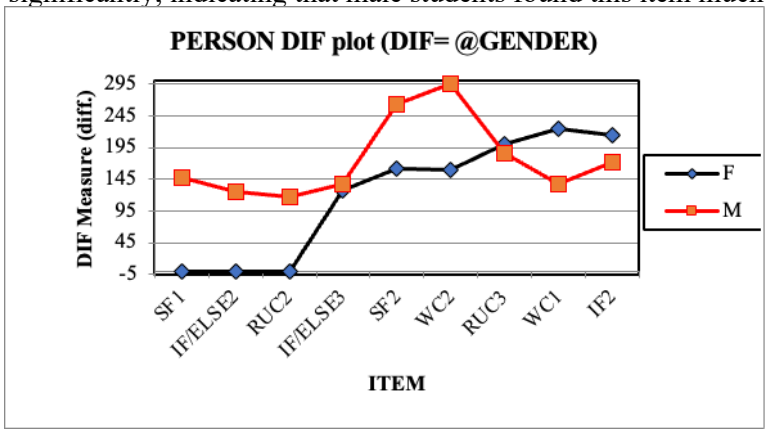

Figure 3 The DIF plot based on gender

Figure 3 presents the DIF analysis based on gender (male and female) for the CTt items. This graph aims to reveal whether certain items exhibit gender bias, potentially impacting the fairness and validity of the assessment. Following the two standard criteria for DIF interpretation—significant probability ($p < 0.05$) and DIF contrast (size). the DIF sizes are categorized into three levels (Zwick et al., 1999):

- Category A: Negligible DIF ($|DIF| < 0.43$ logits)
- Category B: Slight to moderate DIF ($|DIF| \geq 0.43$ logits)
- Category C: Moderate to large DIF ($|DIF| \geq 0.64$ logits)

From Figure 3, it can be observed that most CTt items fall within the negligible DIF range ($|DIF| < 0.43$ logits), meaning they are categorized as Category A. No item surpasses the 0.43 logit threshold. Thus, statistically, the CTt does not exhibit significant gender bias overall, reinforcing its suitability for evaluating computational thinking skills across male and female respondents. Among respondents included in this study.

However, a closer look at individual item trends reveals subtle patterns: Items such as SF1, IF/ELSE2, RUC2, and IF/ELSE3 show higher DIF values for female students. This is evident from the steeper rise in the female DIF line compared to the flatter male DIF line for these items. Although the absolute DIF values are still within the negligible range, the relative difference suggests that these items might align more naturally with the way female students process sequencing and logic-based tasks. In contrast, items like WC1, RUC3, and IF2 exhibit balanced DIF lines between genders, indicating fair functioning across male and female students without meaningful bias. Interestingly, item SF3 shows a small tendency favoring male students, as evidenced by a slightly elevated DIF value for males, although still negligible according to the classification. Thus, even though no item reaches the moderate DIF threshold, the slight variations warrant attention. SF3 may require a review for potential construct-irrelevant variance that slightly advantages male students. IF/ELSE3 may benefit from refinement to ensure it does not inadvertently favor female students.

In conclusion, the graph supports that the CTt is broadly fair and free from major gender bias. However, continuous monitoring and refinement, especially of items showing small but noticeable DIF trends, are important to uphold and enhance the fairness, particularly for future applications in diverse educational contexts. Several limitations should be acknowledged. First, the sample was drawn from only two universities, which limits the generalizability of the findings to the broader population of pre-service science teachers in Indonesia. Second, the unequal gender composition of participants may have influenced the stability of the DIF analysis and interpretation of gender-related findings. Therefore, the results should be interpreted within the context of the sampled institutions, and future studies are recommended to involve more diverse institutions and more balanced participant distributions.

CONCLUSION

This study investigated the CT skills of 419 pre-service science teachers in Indonesia using the Rasch measurement model. The results revealed that, overall, the CTt items demonstrated good psychometric quality, with high item reliability and acceptable item fit statistics. The Rasch model analysis confirmed that most items were productive for measurement and exhibited predictable response patterns. However, person-level reliability and separation indices were relatively low, indicating limited effectiveness in distinguishing varying levels of CT ability among respondents. This suggests a need for refinement in test targeting and possible enhancement of the item pool to better match the ability spectrum of pre-service teachers. DIF analysis indicated that several items functioned differently between male and female students. Specifically, items such as SF2, SF3, and WC2 were more favorable to male students, while items like IF/ELSE2, RUC2, and IF/ELSE3 favored female respondents. These findings highlight potential gender-related disparities in how CT items are interpreted or approached.

Nevertheless, the majority of items functioned equitably across gender groups, affirming the general fairness of the instrument. The use of Rasch analysis in this context offers robust psychometric validation and contributes to the limited literature on CT assessments in teacher education. These results can inform the development of more targeted instructional strategies and assessment tools, ultimately supporting the integration of CT into science teacher education and curriculum reform efforts.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

Albayrak, E., & Yilmaz Ozden, Ş. (2021). Improvement of pre-service teachers' computational thinking skills through an educational technology course. *Journal of Individual Differences in Education*, *3*(2), 97–112. https://doi.org/10.47156/jide.1027431

Aminger, W., Hough, S., Roberts, S. A., Meier, V., Spina, A. D., Pajela, H., McLean, M., & Bianchini, J. A. (2021). Preservice secondary science teachers' implementation of an ngss practice: Using mathematics and computational thinking. *Journal of Science Teacher Education*, *32*(2), 188–209. https://doi.org/10.1080/1046560X.2020.1805200

Bati, K. (2021). Integration of python into science teacher education, developing computational problem solving and using information and communication technologies competencies of pre-service science teachers. *Informatics in Education*. https://doi.org/10.15388/infedu.2022.12

Bond T. G., & Fox C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6857-4

Boulden, D. C., Rachmatullah, A., Oliver, K. M., & Wiebe, E. (2021). Measuring in-service teacher self-efficacy for teaching computational thinking: Development and validation of the T-STEM CT. *Education and Information Technologies*, *26*(4), 4663–4689. https://doi.org/10.1007/s10639-021-10487-2

Bradley, K. D., Peabody, M. R., Akers, K. S., & Knutson, N. M. (2015). Rating Scales in Survey Research: Using the Rasch model to illustrate the middle category measurement flaw. *Survey Practice*, *8*(1), 1–12. https://doi.org/10.29115/SP-2015-0001

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. *AERA*.

Butler, D., & Leahy, M. (2020). Using classroom practice as "an object to think with" to develop preservice teachers' understandings of computational thinking. In T. Brinda, D. Passey, & T. Keane (Eds.), *Empowering Teaching for Digital Equity and Agency* (Vol. 595, pp. 56–65). Springer International Publishing. https://doi.org/10.1007/978-3-030-59847-1_6

Caballero, M. D., & Merner, L. (2018). Prevalence and nature of computational instruction in undergraduate physics programs across the United States. *Physical Review Physics Education Research*, *14*(2), 020129. https://doi.org/10.1103/PhysRevPhysEducRes.14.020129

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. *Journal of Computers in Education*, *8*(2), 213–236. https://doi.org/10.1007/s40692-020-00177-2

Connolly, C., Hijón Neira, R., & Garcia-Iruela, M. (2021). Developing and assessing computational thinking in secondary education using a tpack guided scratch visual execution environment. *International Journal of Computer Science Education in Schools*, *4*(4), 3–23. https://doi.org/10.21585/ijcses.v4i4.98

Dong, W., Li, Y., Sun, L., & Liu, Y. (2023). Developing pre-service teachers' computational thinking: A systematic literature review. *International Journal of Technology and Design Education*. https://doi.org/10.1007/s10798-023-09811-3

Elbahan, H., Elbahan, M. H., & Balbağ, M. Z. (2023). Determining the level of computational thinking skills of science teacher candidates. *Osmangazi Journal of Educational Research*, *10 (Special Issue)*, 254–272. https://doi.org/10.59409/ojer.1369711

Fülöp, M. T., Udvaros, J., Gubán, Á., & Sándor, Á. (2022). Development of computational thinking using microcontrollers integrated into oop (object-oriented programming). *Sustainability*, *14*(12), 7218. https://doi.org/10.3390/su14127218

Ghani, A., Griffiths, D., Salha, S., Affouneh, S., Khalili, F., Khlaif, Z. N., & Burgos, D. (2022). Developing teaching practice in computational thinking in palestine. *Frontiers in Psychology*, *13*, 870090. https://doi.org/10.3389/fpsyg.2022.870090

González, M. R. (2015). *Computational thinking test: Design Guidelines and Content Validation*. In *EDULEARN15 proceedings* (pp. 2436-2444). IATED. https://doi.org/10.13140/RG.2.1.4203.4329

Ismail, I., Riandi, R., Kaniawati, I., Sopandi, W., Supriyadi, S., Suhendar, S., & Hidayat, F. A. (2024). Gender roles in understanding and implementing green energy technology in indonesian schools: Rasch analysis. *Qubahan Academic Journal*, *4*(3), 298–314. https://doi.org/10.48161/qaj.v4n3a752

Ismarmiaty, I., Agustin, K., Madani, M., Sriwinarti, N. K., Zainuddin, Z., & Supatmiwati, D. (2022). Penguatan kemampuan computational thinking pada pemberdayaan guru dan siswa sekolah dasar di pulau lombok. *Transformasi: Jurnal Pengabdian Masyarakat*, *18*(2), 253–267. https://doi.org/10.20414/transformasi.v18i2.5034

KaleliOğlu, F., Gülbahar, Y., & Kukul, V. (2016). A framework for computational thinking based on a systematic research review. *Baltic J. Modern Computing*, *4*(3), 583–596.

Khine, M. S. (2020). Objective measurement in psychometric analysis. In M. S. Khine (Ed.), *Rasch Measurement* (pp. 3–7). Springer Singapore. https://doi.org/10.1007/978-981-15-1800-3_1

Kong, S., & Lai, M. (2023). Effects of a teacher development program on teachers' knowledge and collaborative engagement, and students' achievement in computational thinking concepts. *British Journal of Educational Technology*, *54*(2), 489–512. https://doi.org/10.1111/bjet.13256

Kumala, F. N., Yasa, A. D., Jait, A. B. H., Wibawa, A. P., & Hidayah, L. (2023). Patterns of computational thinking skills for elementary prospectives teacher in science learning: Gender analysis studies. *International Journal of Elementary Education*, *7*(4), 646–656. https://doi.org/10.23887/ijee.v7i4.68611

Lee, I., Grover, S., Martin, F., Pillai, S., & Malyn-Smith, J. (2020). Computational thinking from a disciplinary perspective: Integrating computational thinking in k-12 science, technology, engineering, and mathematics education. *Journal of Science Education and Technology*, *29*(1), 1–8. https://doi.org/10.1007/s10956-019-09803-w

Liu, T., Gonzalez-Maldonado, D., Harlow, D. B., Edwards, E. E., & Franklin, D. (2023). Qupcakery: A puzzle game that introduces quantum gates to young learners. *SIGCSE - Proc. ACM Tech. Symp. Comput. Sci. Educ.*, *1*, 1143–1149. Scopus. https://doi.org/10.1145/3545945.3569837

Lyon, J. A., & J. Magana, A. (2020). Computational thinking in higher education: A review of the literature. *Computer Applications in Engineering Education*, *28*(5), 1174–1189. https://doi.org/10.1002/cae.22295

M Esteve-Mon, F., Adell-Segura, J., Ángeles Llopis Nebot, M., Valdeolivas Novella, G., & Pacheco Aparicio, J. (2019). The development of computational thinking in student teachers through an intervention with educational robotics. *Journal of Information Technology Education: Innovations in Practice*, *18*, 139–152. https://doi.org/10.28945/4442

Marifah, S. N. (2022). *Systematic literatur review: Integrasi computational thinking dalam kurikulum sekolah dasar di indonesia*. *COLLASE (Creative of Learning Students Elementary Education)*, *5*(5), 928-938.

Newman, C. (2014). Time to address gender discrimination and inequality in the health workforce. *Human resources for health*, *12*(1), 25.

Odden, T. O. B., & Burk, J. (2020). Computational essays in the physics classroom. *The Physics Teacher*, *58*(4), 252–255. https://doi.org/10.1119/1.5145471

Peel, A., Sadler, T. D., & Friedrichsen, P. (2022). Algorithmic explanations: An unplugged instructional approach to integrate science and computational thinking. *Journal of Science Education and Technology*, *31*(4), 428–441. https://doi.org/10.1007/s10956-022-09965-0

Peters-Burton, E., Rich, P. J., Kitsantas, A., Laclede, L., & Stehle, S. M. (2022). High school science teacher use of planning tools to integrate computational thinking. *Journal of Science Teacher Education*, *33*(6), 598–620. https://doi.org/10.1080/1046560X.2021.1970088

Prastowo, A., & Fitriyaningsih, F. (2020). Learning material changes as the impact of the 2013 curriculum policy for the primary school/madrasah ibtidaiyah. *Edukasia : Jurnal Penelitian Pendidikan Islam*, *15*(2), 251. https://doi.org/10.21043/edukasia.v15i2.7947

Subramaniam, S., Maat, S. M. M., & Mahmud, M. S. M. (2023). Designing problem-solving module based on computational thinking in mathematics education: Nominal group technique approach. *International Journal of Academic Research in Progressive Education and Development*, *12*(2), Pages 1381-1396. https://doi.org/10.6007/IJARPED/v12-i2/17262

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

Syafril, S., Rahayu, T., & Ganefri, G. (2022). Prospective science teachers' self-confidence in computational thinking skills. *Jurnal Pendidikan IPA Indonesia*, *11*(1), 119–128. https://doi.org/10.15294/jpii.v11i1.33125

Tongal, A., Yıldırım, F. S., Özkara, Y., Say, S., & Erdoğan, Ş. (2024). Examining teachers' computational thinking skills, collaborative learning, and creativity within the framework of sustainable education. *Sustainability*, *16*(22), 9839. https://doi.org/10.3390/su16229839

Tsakeni, M. (2021). Preservice teachers' use of computational thinking to facilitate inquiry-based practical work in multiple-deprived classrooms. *Eurasia Journal of Mathematics, Science and Technology Education*, *17*(1), em1933. https://doi.org/10.29333/ejmste/9574

Verawati, N. N. S. P., Rijal, K., & Grendis, N. W. B. (2023). Examining stem students' computational thinking skills through interactive practicum utilizing technology. *International Journal of Essential Competencies in Education*, *2*(1), 54–65. https://doi.org/10.36312/ijece.v2i1.1360

Wardani, I. R. W., Putri Zuani, M. I., & Kholis, N. (2023). Teori belajar perkembangan kognitiv lev vygotsky dan implikasinya dalam pembelajaran. DIMAR: Jurnal Pendidikan Islam, 4(2), 332–346. https://doi.org/10.58577/dimar.v4i2.92

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35. https://doi.org/10.1145/1118178.1118215

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1881), 3717–3725. https://doi.org/10.1098/rsta.2008.0118

Zhu, M., & Wang, C. (2023). Core competencies of K-12 computer science education from the perspectives of college faculties and K-12 teachers. *International Journal of Computer Science Education in Schools*, *6*(2). https://doi.org/10.21585/ijcses.v6i2.161

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical bayes approach to mantel-haenszel dif analysis. *Journal of Educational Measurement*, *36*(1), 1–28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x