Indonesian Journal of Educational Development (IJED)
Volume 7, Issue 2, 2026, pp. 892-906
ISSN: 2722-1059 (Online); ISSN: 2722-3671 (Print)
DOI: https://doi.org/10.59672/ijed.v7i2.7023

Rasch evaluation of matched positively and negatively worded CTS items among high school students

Ahmad Fauzi[1], Hartono*)[2], Sunyoto Eko Nugroho[3], Arif Widyatmoko[4]
[1]Universitas Negeri Semarang, Jawa Tengah, Indonesia; fauziuns@students.unnes.ac.id
[2]Universitas Negeri Semarang, Jawa Tengah, Indonesia; hartono@mail.unnes.ac.id
[3]Universitas Negeri Semarang, Jawa Tengah, Indonesia; ekonuphysed@mail.unnes.ac.id
[4]Universitas Negeri Semarang, Jawa Tengah, Indonesia; arif.widiyatmoko@mail.unnes.ac.id
*)Corresponding author: Hartono; E-mail addresses: hartono@mail.unnes.ac.id

Article Info

Article history:
Received July 24, 2026
Revised August 03, 2026
Accepted August 22, 2026
Available online August 26, 2026

Keywords: Computational Thinking Scale, Modified, Negatively worded, Positively worded, Rasch model

Copyright ©2026 by Author. Published by Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas PGRI Mahadewa Indonesia

Abstract. Instruments composed of items worded in the same direction may be vulnerable to response bias. This study evaluated a modified Computational Thinking Scale (CTS) comprising 19 matched pairs of positively and negatively worded items. An instrument-development design adapted from ADDIE was used. The study population comprised science-track senior high school students in the Surakarta Residency. The sample included 393 Grade XI students from three purposively selected schools. Data were collected using a 38-item, five-category self-report questionnaire. Five experts assessed content validity, yielding coefficients of 0.80–0.95 and a mean of 0.89. Responses were analyzed using the Rasch model in Winsteps 5.7.3.0. Cronbach's alpha and person reliability were 0.66, with person separation of 1.40, whereas item reliability was 1.00 and item separation was 14.67. Response categories showed ordered observed averages and Andrich thresholds, with outfit MNSQ values of 0.91–1.07. All items met the primary outfit MNSQ criterion, ranging from 0.82 to 1.21. Three algorithmic-thinking items showed statistically significant but small gender DIF contrasts of 0.30–0.35 logits. The hypothesis was partially supported because the instrument showed acceptable item-level functioning but limited person-level discrimination. Future studies should examine wording effects, matched-pair equivalence, dimensionality, and local dependence in broader, more diverse samples.

Introduction

Computational thinking (CT) is an essential skill in the twenty-first century. CT is closely associated with an individual's ability to think systematically while solving problems. Although it is rooted in computer science, CT is not required only by those studying computing-related disciplines. It is also needed by individuals across various disciplines that require logical reasoning, structured problem-solving, and data-driven decision-making (Li et al., 2020; Wing, 2006). CT involves formulating problems such that their solutions can be represented as computational steps and algorithms. This process involves abstraction, decomposition, the selection of appropriate representations or models, and the development of solutions that can be carried out systematically and efficiently (Aho, 2012; Denning, 2009). In educational contexts, CT is often understood as a problem-solving approach involving problem decomposition, data analysis and representation, abstraction, and algorithmic thinking (CSTA & ISTE, 2011; Voogt et al., 2015). CT is particularly relevant to science and STEM because both fields require scientific reasoning, data processing,

model development, and the evaluation of evidence-based solutions. Therefore, CT requires particular attention in science and STEM education (Cheng, 2023; Weintrop et al., 2016; Suryaningsih et al., 2025; Widana et al., 2021).

The integration of CT into science learning can support a wide range of learning activities, including observing phenomena, conducting experiments, analyzing data, developing models, using simulations, and interpreting results logically (Barr & Stephenson, 2011; Hurt et al., 2023; Weintrop et al., 2016). Research on the integration of CT into different areas of science and technology has also expanded to fields such as biology, physics, chemistry, STEM, and digital technology-based learning (Chongo et al., 2021; Matsumoto & Cao, 2017; Ogegbo & Ramnarain, 2022). Therefore, measuring CT among senior high school students in the science track is important because science learning is directly associated with scientific reasoning, problem-solving, data analysis, and the modeling of scientific concepts (Huda & Rohaeti, 2024; Weintrop et al., 2016).

CT has been integrated into curricula in many countries, from primary to higher education. However, its implementation in Indonesia faces several challenges. These include teacher readiness, limited supporting facilities, uneven levels of student understanding, and the limited availability of assessment instruments suited to the local educational context (Angeli & Giannakos, 2020; Huda & Rohaeti, 2024). Teachers generally have positive perceptions of CT integration into learning, but its classroom implementation has not yet been fully optimized. This condition is partly due to the limited availability of professional training, contextual learning resources, implementation guidelines, and school policy support (Yudi Hartawan et al., 2026). Strengthening CT, therefore, requires support not only in instructional readiness but also in assessment instruments that accurately capture students' CT tendencies. This is important because studies of CT among Indonesian senior high school students, particularly those in the science track, remain more limited than studies conducted in mathematics, preservice teacher education, or programming contexts (Huda & Rohaeti, 2024; Sondakh et al., 2022). Therefore, the quality of instruments used to measure the intensity of students' CT tendencies needs to be empirically examined before the instruments are used more widely in learning contexts (Tang et al., 2020).

CT assessment can be conducted using cognitive and noncognitive approaches (Tang et al., 2020). Cognitive assessment measures students' performance in completing CT-based tasks or test items, whereas noncognitive assessment measures their tendencies, dispositions, or the intensity of their computational thinking behaviors when dealing with problems (Korkmaz et al., 2017). Noncognitive CT assessments are less common than cognitive CT assessments (Tang et al., 2020). The Computational Thinking Scale (CTS) developed by Korkmaz et al. (2017) is a widely used noncognitive instrument for measuring CT. The CTS measures CT through five main dimensions: creativity, algorithmic thinking, cooperativity, critical thinking, and problem solving (Korkmaz et al., 2017). Through these five dimensions, the CTS measures students' tendencies to apply CT when dealing with everyday problems (Huda & Rohaeti, 2024; Korkmaz et al., 2017).

The Computational Thinking Scale has previously been used to assess computational thinking skills among Indonesian senior high school students in the science track (Huda & Rohaeti, 2024). Based on a validation study involving 526 students in Yogyakarta, the 19-item Indonesian version of the CTS was found to represent five dimensions of CT: creativity, algorithmic thinking, cooperativity, critical thinking, and problem-solving. However, the item wording directions were not evenly distributed across all dimensions. The instrument contained 15 positively worded items and 4 negatively worded items, all of which belonged to the problem-solving dimension.

The Rasch analysis showed that the CTS generally met the model-fit criteria, demonstrated adequate internal reliability, and could be used to describe students' CT levels and identify

differences in item functioning based on coding experience (Huda & Rohaeti, 2024). However, this study did not examine the psychometric functioning of an instrument in which positively and negatively worded items were distributed across all CTS dimensions. Because all negatively worded items belonged to the problem-solving dimension, the psychometric performance of a CTS containing oppositely worded items across the other dimensions remained unclear. In Likert-type scales, positively and negatively worded items are often used to reduce response bias, particularly respondents' tendency to repeatedly select high response categories without carefully considering the content of each statement (Setiawati et al., 2022; Suárez-Álvarez et al., 2018). However, negatively worded items do not always improve the measurement quality. Such items may reduce reliability, introduce method variance, increase respondents' cognitive burden, and produce response patterns that may appear to disrupt the unidimensionality of the instrument (Suárez-Álvarez et al., 2018). Therefore, a CTS containing matched positively and negatively worded items within each dimension needs to be empirically evaluated to determine its overall psychometric quality when statements with comparable content are formulated in opposite semantic directions.

The Rasch model is widely used to evaluate the quality of educational instruments because it can place respondent ability and item difficulty on the same logit scale (Boone et al., 2014; Khine, 2020). Through this model, item quality can be evaluated using outfit mean-square statistics, outfit z-standardized values, and point-measure correlations (Bond & Fox, 2013; Sumintono & Widhiarso, 2015). The Rasch model also provides information on person and item reliability and separation, response-category functioning, item difficulty, and Differential Item Functioning (DIF), which can be used to identify items that function differently across particular groups of respondents (Boone et al., 2014; Chan, 2020; Sumintono & Widhiarso, 2015). Therefore, the Rasch model is suitable for evaluating whether the modified CTS, containing matched positively and negatively worded items, continues to demonstrate adequate reliability, response-category functioning, and item fit.

Based on the preceding discussion, this study aimed to evaluate the psychometric quality of a modified Computational Thinking Scale containing matched positively and negatively worded items among senior high school students in the science track using Rasch analysis. The evaluation focused on five main dimensions of the CTS: creativity, algorithmic thinking, cooperativity, critical thinking, and problem-solving. More specifically, the study examined the instrument's content validity, reliability, response-category functioning, item difficulty based on logit estimates, item fit to the Rasch model, and possible gender-based differences in item functioning through DIF analysis.

The preceding evidence identifies an important measurement ambiguity in the Indonesian version of the CTS: because its negatively worded items were confined to the problem-solving dimension, wording effects could not be distinguished from dimension-specific content. This ambiguity may compromise the interpretation of students' CT tendencies and consequently limit the instrument's usefulness in educational assessment. Based on the literature indicating both the potential benefits and limitations of oppositely worded items, the present study addressed this problem by pairing 19 positively worded items with 19 negatively worded counterparts of comparable content across the five CTS dimensions. This study's novelty lies in systematically pairing positively and negatively worded items across all five CTS dimensions: creativity, algorithmic thinking, cooperativity, critical thinking, and problem-solving. The study addressed the following research question: To what extent does the modified CTS demonstrate adequate content validity, reliability, response-category functioning, item fit, an interpretable distribution of item difficulty, and no substantively meaningful gender DIF under the Rasch model? It was hypothesized that the modified CTS would demonstrate adequate content validity and reliability, ordered response categories, acceptable item fit, an interpretable distribution of item difficulty, and no substantively meaningful gender DIF.

Accordingly, this study aimed to evaluate the psychometric quality of the modified 38-item CTS among science-track senior high school students using Rasch analysis.

Method

Research Design, Procedures, and Setting

This study involved instrument development and psychometric evaluation using an adapted ADDIE framework (Branch, 2009). During the analysis stage, the CT construct, the five CTS dimensions, the validated Indonesian CTS, and its suitability for Grade XI science-track students were reviewed. The design stage involved mapping the retained CTS items to their respective dimensions and pairing each primary item with a counterpart formulated in the opposite semantic direction. The development stage covered expert review, instrument revision, and preparation for field administration. The implementation stage involved administering the revised instrument to the participants. In the final evaluation stage, negatively worded items were reverse-scored, and the responses were analyzed using the Rasch rating-scale model in Winsteps 5.7.3.0. The study was conducted from February to July 2025 at SMAN 4 Surakarta, SMAN Kartasura, and SMAN 7 Surakarta.

Participants and Sampling

The study population comprised grade XI science-track senior high school students in the Surakarta Residency. Three schools were purposively selected as the study sites. A total of 393 Grade XI science-track students from these schools participated voluntarily and were retained for analysis, comprising 157 male and 236 female students. Only one item response was missing, which was treated as missing data in the Rasch analysis.

Product Specifications and Instrument Blueprint

The modified CTS consisted of 38 items organized into 19 matched pairs across five dimensions: creativity, algorithmic thinking, cooperativity, critical thinking, and problem-solving. Each pair contained one primary item and a counterpart formulated in the opposite semantic direction. Responses were recorded on a five-category frequency scale: 1 = never, 2 = rarely, 3 = sometimes, 4 = usually, and 5 = always. Table 1 presents the correspondence between the original CTS item codes, their matched positively and negatively worded counterparts, and the S1–S38 item labels used in the Winsteps output.

Table 1. Blueprint of the Modified Computational Thinking Scale

CT dimension	Pair	Positively worded item	Negatively worded item
Creativity	1	C1 (S1)	C1-R (S2)
Creativity	2	C4 (S3)	C4-R (S4)
Creativity	3	C5 (S5)	C5-R (S6)
Algorithmic thinking	4	A1 (S7)	A1-R (S8)
Algorithmic thinking	5	A3 (S9)	A3-R (S10)
Algorithmic thinking	6	A4 (S11)	A4-R (S12)
Algorithmic thinking	7	A6 (S13)	A6-R (S14)
Cooperativity	8	O1 (S15)	O1-R (S16)
Cooperativity	9	O2 (S17)	O2-R (S18)
Cooperativity	10	O3 (S19)	O3-R (S20)
Cooperativity	11	O4 (S21)	O4-R (S22)
Critical thinking	12	T1 (S23)	T1-R (S24)
Critical thinking	13	T2 (S25)	T2-R (S26)

CT dimension	Pair	Positively worded item	Negatively worded item
Critical thinking	14	T3 (S27)	T3-R (S28)
Critical thinking	15	T5 (S29)	T5-R (S30)
Problem-solving	16	P1-R (S32)	P1 (S31)
Problem-solving	17	P2-R (S34)	P2 (S33)
Problem-solving	18	P3-R (S36)	P3 (S35)
Problem-solving	19	P4-R (S38)	P4 (S37)

Note. Codes in parentheses (S1–S38) indicate the item labels used in the Winsteps analysis.

Instrument Development and Validation

The 38-item draft was reviewed by five experts in physics education, educational evaluation, and instrument development. The review addressed clarity, instructions, language, CT relevance, construct alignment, and readability. Content validity was assessed using Aiken's V based on ratings from 1 to 5. The Aiken's V value was calculated using the following formula:

$$V=\frac{\sum r-l_0}{n(c-1)} \quad (1)$$

Where r is the validator score, l_0 is the lowest possible rating, c is the number of rating categories, and n is the number of experts/raters. Values above 0.80 were classified as highly valid, values from 0.60 to 0.80 as valid, and values below 0.60 as having low validity.

Data Collection

The 38 items were presented in a randomized order to reduce order effects and patterned responding. The questionnaire was administered during regular classroom sessions under the supervision of the researchers and teachers. Before completing the questionnaire, participants were informed about the purpose of the study and given instructions on how to respond. They were also told that participation was voluntary and that their responses would remain confidential.

Data Analysis and Decision Criteria

For analysis, responses to the randomly ordered items were reorganized by dimension and labeled S1–S38, as shown in Table 1. Negatively worded items were reverse-scored before analysis. Psychometric functioning was examined using the Rasch rating-scale model in Winsteps 5.7.3.0. Analyses included Cronbach's alpha, person and item reliability and separation, response-category functioning, item measures, outfit MNSQ and ZSTD, point-measure correlations, and gender-based differential item functioning (DIF).

Outfit MNSQ values of 0.50–1.50 were considered productive for measurement (Wright & Linacre, 1994). ZSTD values between −2.00 and +2.00 and positive point-measure correlations were treated as supporting evidence of item fit. Items with point-measure correlations ≤ .20 were reviewed rather than automatically removed. Category functioning was evaluated using category use, increasing observed averages, ordered Andrich thresholds, and category outfit MNSQ values below 2.00 (Linacre, 2002). DIF was considered statistically significant at $p < .05$ and substantively meaningful when the absolute DIF contrast was at least 0.50 logits. Smaller but statistically significant DIF values prompted substantive review rather than automatic item removal.

Results and Discussion

Results

Content Validity Results

The expert-validation results indicated that the modified CTS had acceptable content validity at the instrument level. The Aiken's V coefficients for the ten evaluation criteria ranged from 0.80 to 0.95, with a mean of 0.89. The criterion-level content-validation results are presented in Table 2.

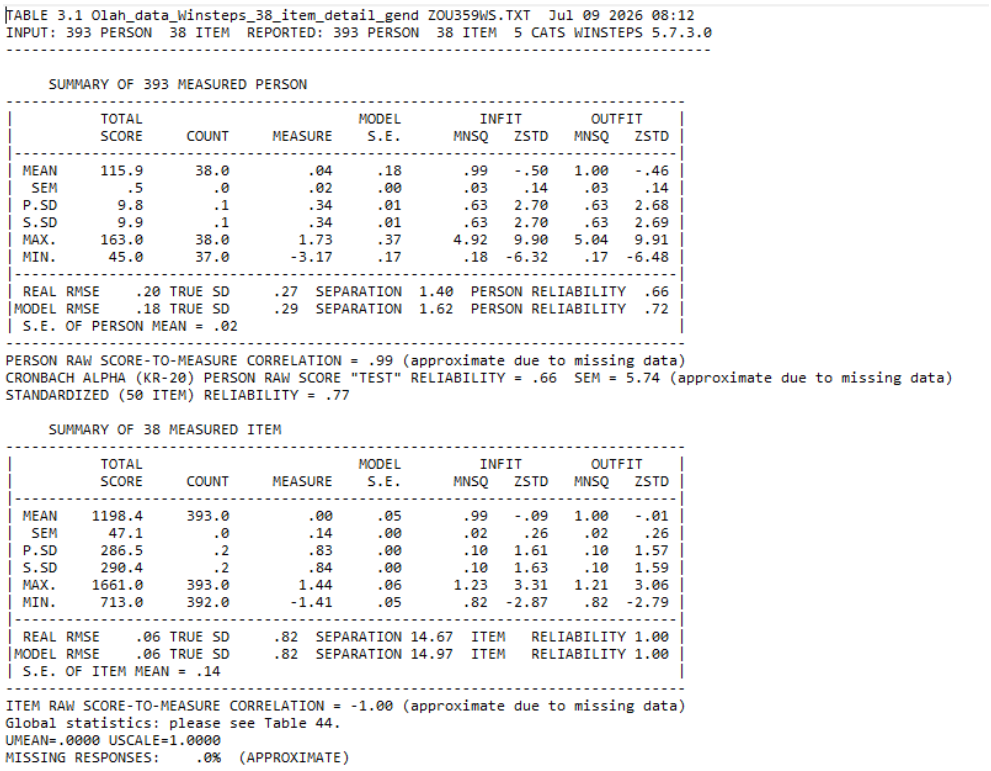
Table 2. Criterion-Level Expert Validation Results

No.	Evaluation criterion	Aiken's V	Interpretation
1.	Item clarity	0.90	Highly valid
2.	Clarity of completion instructions	0.95	Highly valid
3.	Language appropriateness for senior high school students	0.85	Highly valid
4.	Alignment of questionnaire format with the CT construct	0.80	Valid
5.	Relevance of statements to computational thinking	0.95	Highly valid
6.	Accuracy of information conveyed	0.90	Highly valid
7.	Completeness of ideas expressed	0.85	Highly valid
8.	Readability	0.85	Highly valid
9.	Language effectiveness	0.95	Highly valid
10.	Conformity with standard Indonesian language conventions	0.90	Highly valid
	Mean	0.89	Highly valid

As shown in Table 2, nine criteria obtained coefficients above 0.80, while one criterion—the alignment of the questionnaire format with the CT construct, obtained a coefficient of exactly 0.80. Thus, all ten criteria met the prespecified acceptability threshold. The evaluated criteria included the clarity of each statement, clarity of the questionnaire instructions, appropriateness of the language for senior high school students, alignment of the questionnaire format with the CT indicators, relevance of the statements to CT content, accuracy of the information expressed, completeness of the ideas presented, ease of understanding, language conciseness and effectiveness, and conformity with standard Indonesian spelling conventions. Overall, these results supported the general content appropriateness of the modified CTS. Because the coefficients were calculated for broad validation criteria rather than for each of the 38 items or each matched pair, they should not be interpreted as independent evidence that every positively and negatively worded item pair was semantically equivalent.

Reliability and Response-Category Structure

All 393 respondents and 38 items were included in the Winsteps analysis. Only one item response was missing, corresponding to approximately 0.01% of the expected responses and rounded to 0.0% in the software output. The analysis yielded a Cronbach's alpha of 0.66, a person reliability of 0.66, and a person separation of 1.40. These values indicate moderate reliability and limited person-level discrimination. In contrast, an item reliability of 1.00 and an item separation of 14.67 indicated that the hierarchy of item difficulty was highly stable. This means that the items had a clear and reliable distribution of difficulty levels when administered to samples with comparable characteristics. The direct software output is presented in Image 1, while Table 3 provides an interpretive summary.



```
TABLE 3.1 Olah_data_Winsteps_38_item_detail_gend ZOU359WS.TXT  Jul 09 2026 08:12
INPUT: 393 PERSON  38 ITEM  REPORTED: 393 PERSON  38 ITEM  5 CATS WINSTEPS 5.7.3.0
-------------------------------------------------------------------------------

     SUMMARY OF 393 MEASURED PERSON
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     115.9      38.0         .04     .18       .99   -.50   1.00   -.46 |
|  SEM        .5        .0         .02     .00       .03    .14    .03    .14 |
| P.SD       9.8        .1         .34     .01       .63   2.70    .63   2.68 |
| S.SD       9.9        .1         .34     .01       .63   2.70    .63   2.69 |
| MAX.     163.0      38.0        1.73     .37      4.92   9.90   5.04   9.91 |
| MIN.      45.0      37.0       -3.17     .17       .18  -6.32    .17  -6.48 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .20 TRUE SD     .27  SEPARATION  1.40  PERSON RELIABILITY  .66 |
|MODEL RMSE    .18 TRUE SD     .29  SEPARATION  1.62  PERSON RELIABILITY  .72 |
| S.E. OF PERSON MEAN = .02                                                   |
-------------------------------------------------------------------------------
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99 (approximate due to missing data)
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .66  SEM = 5.74 (approximate due to missing data)
STANDARDIZED (50 ITEM) RELIABILITY = .77

     SUMMARY OF 38 MEASURED ITEM
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN    1198.4     393.0         .00     .05       .99   -.09   1.00   -.01 |
|  SEM       47.1        .0         .14     .00       .02    .26    .02    .26 |
| P.SD     286.5        .2         .83     .00       .10   1.61    .10   1.57 |
| S.SD     290.4        .2         .84     .00       .10   1.63    .10   1.59 |
| MAX.    1661.0     393.0        1.44     .06      1.23   3.31   1.21   3.06 |
| MIN.     713.0     392.0       -1.41     .05       .82  -2.87    .82  -2.79 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .06 TRUE SD     .82  SEPARATION 14.67  ITEM   RELIABILITY 1.00 |
|MODEL RMSE    .06 TRUE SD     .82  SEPARATION 14.97  ITEM   RELIABILITY 1.00 |
| S.E. OF ITEM MEAN = .14                                                     |
-------------------------------------------------------------------------------
ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00 (approximate due to missing data)
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000
MISSING RESPONSES:    .0%  (APPROXIMATE)
```

Image 1. Direct Winsteps Output for Person and Item Summary Statistics

Table 3. Summary of Person and Item Reliability Based on Rasch Analysis

Indicator	Value	Interpretation
Cronbach's alpha	0.66	Moderate internal reliability
Person reliability	0.66	Moderate person reliability
Person separation	1.40	The instrument's ability to distinguish respondents is not yet strong
Item reliability	1.00	Very high item reliability
Item separation	14.67	The item-difficulty hierarchy is highly stable
Mean person measure	0.04 logits	The mean person measure is slightly above the mean item difficulty
Mean item measure	0.00 logits	Serves as the item-calibration reference point

The response category structure also showed adequate results. The observed averages increased sequentially from Categories 1 to 5, with values of −0.86, −0.56, 0.01, 0.62, and 0.88. All response categories were used by the respondents, with proportions ranging from 13% to 34%, while the category outfit MNSQ values ranged from 0.91 to 1.07. The Andrich thresholds also increased sequentially, although the distances between Categories 2 and 3 and between Categories 4 and 5 were relatively small. Therefore, the five-point response scale functioned logically and could be retained in Rasch analysis. The direct Winsteps output is presented in Image 2, while Table 4 summarizes its interpretation.

TABLE 3.2 Olah_data_Winsteps_38_item_detail_gend ZOU375WS.TXT Jul 09 2026 09:05
INPUT: 393 PERSON 38 ITEM REPORTED: 393 PERSON 38 ITEM 5 CATS WINSTEPS 5.7.3.0

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
1	1	1883	13	-.86	-.90	1.06	1.07	NONE	(-2.56)	1
2	2	2724	18	-.56	-.51	.90	.92	-1.08	-1.13	2
3	3	5143	34	.01	.02	.94	.94	-.89	-.02	3
4	4	3136	21	.62	.55	.89	.91	.79	1.12	4
5	5	2047	14	.88	.92	1.07	1.07	1.17	(2.61)	5
MISSING		1	0	.95						

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

CATEGORY LABEL	JMLE STRUCTURE MEASURE	S.E.	SCORE-TO-MEASURE AT CAT.	----ZONE----		50% CUM. PROBABLTY	COHERENCE M->C	C->M	RMSR	ESTIM DISCR	
1	NONE		(-2.56)	-INF	-1.89		71%	2%	1.3599		1
2	-1.08	.03	-1.13	-1.89	-.56	-1.53	34%	57%	.7739	.88	2
3	-.89	.02	-.02	-.56	.53	-.64	47%	42%	.6596	1.04	3
4	.79	.02	1.12	.53	1.91	.59	36%	64%	.7230	1.08	4
5	1.17	.03	(2.61)	1.91	+INF	1.58	40%	2%	1.3154	.87	5

M->C = Does Measure imply Category?
C->M = Does Category imply Measure?

Category Matrix : Confusion Matrix : Matching Matrix
Predicted Scored-Category Frequency

Obs Cat Freq	1	2	3	4	5	Total
1	540.49	553.35	587.40	156.41	45.34	1883.00
2	585.71	726.40	956.07	332.68	123.13	2724.00
3	582.24	971.05	1923.62	1077.05	589.03	5143.00
4	130.52	327.16	1061.57	914.66	702.09	3136.00
5	47.77	145.67	608.64	654.16	590.77	2047.00
Total	1886.73	2723.63	5137.30	3134.97	2050.36	14933.00

Image 2. Direct Winsteps Output for the Five-Category Rating-Scale Structure

Table 4. Summary of the Response-Category Structure

Category-Structure Indicator	Result	Interpretation
Number of response categories	5 categories	The 1–5 response scale was used in the analysis
Observed average for Category 1	−0.86	Represents the lowest overall response level
Observed average for Category 2	−0.56	Higher than the preceding category
Observed average for Category 3	0.01	Located near the midpoint of the scale
Observed average for Category 4	0.62	Represents a higher overall response level
Observed average for Category 5	0.88	Represents the highest overall response level
Response proportions	13%–34%	All categories were used by the respondents
Category outfit MNSQ	0.91–1.07	All categories were within an acceptable range
Andrich thresholds	Increased sequentially	The category structure functioned logically

Based on this summary, the response categories showed no serious dysfunction. Therefore, the five-category scale functioned adequately in the present sample.

Fit of the Matched Positively and Negatively Worded Items

Item fit was analyzed using the outfit mean square (MNSQ), outfit z-standardized (ZSTD), and point-measure correlation values. The analysis of the 38 items showed that all outfit MNSQ values ranged from 0.82 to 1.21. These values were within the acceptable range, indicating that none of the items showed serious misfit according to the outfit MNSQ criterion. Therefore, all 38 items were retained.

Nevertheless, several items required further review because their Z-standard deviation (ZSTD) values fell outside the range of −2 to +2. Items S30, S21, and S24 had high positive ZSTD values, whereas S7, S5, S12, S9, and S10 had negative ZSTD values below −2. Positive ZSTD values indicate responses that were more unexpected than predicted by the model, whereas negative ZSTD values indicate response patterns that were overly predictable or tended to overfit the model. However, because the outfit MNSQ values of all items still met the acceptable criterion, these items did not need to be removed immediately. Instead, their wording and suitability for the student context should be reviewed.

Based on the point-measure correlations, all items had positive correlations, ranging from 0.18 to 0.37. This indicates that all items functioned in the intended direction relative to the overall Rasch measure after reverse-scoring. However, several items had relatively low correlations, particularly S33, S16, S32, and S35. Overall, neither wording direction appeared to be the dominant source of problems because the items requiring review were distributed across both positively and negatively worded items. Therefore, matched oppositely worded items may be retained for further investigation, but their pairwise psychometric equivalence requires direct testing.

Distribution of Person Measures and Item Difficulty

The distribution of person measures and item difficulty was analyzed on the Rasch logit scale. The mean person measure was 0.04 logits, which was very close to the mean item measure of 0.00 logit. This finding indicates that the difficulty levels of the items were relatively well matched to the respondents' CT tendency levels. The slightly higher mean person measure suggests that the items were, on average, marginally easier to endorse than the mean item difficulty.

Based on the item measure order, the most difficult item was S30, with a value of 1.44 logits, followed by S8 at 1.31 logits, S22 at 1.10 logits, S23 at 1.04 logits, and S16 at 1.03 logits. These five items were distributed across the critical-thinking, cooperativity, and algorithmic-thinking dimensions: S30 and S23 belonged to critical thinking, S22 and S16 to cooperativity, and S8 to algorithmic thinking. In contrast, the least difficult item was S2, with a value of −1.41 logits, followed by S1 at −1.18 logits, S28 at −1.12 logits, S18 at −1.04 logits, and S15 at −0.97 logits. Thus, the relatively difficult-to-endorse CT behaviors were not concentrated in a single dimension. Items with higher logit values represented CT behaviors that were displayed less frequently, whereas those with lower logit values represented behaviors that were easier to endorse or occurred more frequently among respondents. Further examination of the item content is needed to determine whether these differences were associated with the complexity of the behaviors described, wording direction, or other item characteristics.

Differential Item Functioning by Gender

Differential Item Functioning (DIF) analysis was conducted to determine whether any items functioned differently for male and female respondents. The Winsteps output showed that the DIF grouping by gender was successfully recognized, comprising 157 male and 236 female respondents.

The analysis indicated that three items showed a statistically significant DIF: S8, S9, and S12. All three items belonged to the algorithmic thinking dimension of the questionnaire. Item S8 had a

probability value of 0.0037 and a DIF contrast of −0.34, S9 had a probability value of 0.0013 and a DIF contrast of −0.35, and S12 had a probability value of 0.0054 and a DIF contrast of 0.30. Although statistically significant, the DIF contrast values remained within the range of 0.30–0.35 logits and did not exceed the threshold of 0.50 logits. Therefore, the three items did not need to be removed immediately, but their wording and the contextual relevance of their content should be reviewed.

The direction of these differences indicates that S8 and S9 were relatively more difficult for female respondents, whereas S12 was relatively more difficult for male respondents after controlling for the overall person measure. Because all three items belonged to the algorithmic thinking dimension, items involving symbols, mathematical concepts, equations, and relationships among numbers or images may be more sensitive to gender-related response differences. Table 5 summarizes the items requiring review based on the ZSTD values, point-measure correlations, and DIF indications.

Table 5. Summary of Items Requiring Review Based on Rasch Analysis Indicators

Basis for Review	Items	Finding	Decision
High positive ZSTD	S30, S21, S24	Responses were more unexpected than predicted by the model	Reviewed but not removed
Low negative ZSTD	S7, S5, S12, S9, S10	Response patterns were overly predictable or showed overfit	Reviewed but not removed
Low point-measure correlation	S33, S16, S32, S35	Correlations were positive but relatively low	Reviewed in terms of wording and context
Gender-based DIF	S8, S9, S12	Statistically significant, but the DIF contrast was below 0.50 logits	Reviewed but not removed

Based on the content-validity assessment and Rasch measurement results, the hypothesis was partially supported. The modified CTS demonstrated adequate content validity, ordered response categories, acceptable item-level fit, an interpretable distribution of item difficulty, and no substantively meaningful gender DIF. However, the person reliability of 0.66 and person separation of 1.40 indicated only moderate reliability and limited person-level discrimination. Several items require further review based on their ZSTD values, point-measure correlations, and DIF results. There was insufficient evidence to remove any item because all items met the primary outfit MNSQ criterion and showed positive point-measure correlations.

Discussion

The findings indicate that the CTS modified by pairing positively and negatively worded items generally has adequate psychometric quality. All 38 items had outfit MNSQ values ranging from 0.82 to 1.21, indicating that none showed a serious misfit. This finding suggests that both item types produced response patterns consistent with the Rasch model. Thus, matched items across all CTS dimensions showed no serious item-level misfit. Nevertheless, fit to the Rasch model alone is not sufficient to demonstrate that positively worded items and their negatively worded counterparts function equivalently in a psychometric manner. Several items still require further review based on their ZSTD values and point-measure correlations because items that are appropriate in terms of content do not necessarily function optimally in empirical terms.

The content-validation results showed that the Aiken's V coefficients for each criterion provided evidence of adequate content validity for the modified CTS in terms of relevance, clarity,

readability, and linguistic appropriateness. Content validation used ten general criteria rather than separate item-by-item and matched-pair analyses; therefore, these coefficients did not establish semantic or psychometric equivalence within each item pair. The observed mean person measures for each response category and the Andrich threshold values showed an ordered pattern. This finding indicates that respondents used the five response categories according to their intended levels, consistent with the expectations of the Rasch rating-scale model (Boone et al., 2014; Khine, 2020). Nevertheless, the relatively narrow distances between some adjacent thresholds should be re-examined using a larger sample.

The Indonesian version of the CTS, adapted by Huda and Rohaeti (2024), consisted of 19 items, including 15 positively worded items and 4 negatively worded items. All negatively worded items in the instrument were confined to the problem-solving dimension, whereas the creativity, algorithmic thinking, cooperativity, and critical thinking dimensions were measured entirely using positively worded items. The CTS validation results showed a Cronbach's alpha of 0.79, a person reliability of 0.74, and an item reliability of 0.99. In the present study, each original item was paired with a statement written in the opposite semantic direction. The modified instrument therefore consisted of 19 matched item pairs, or 38 statements, with each pair containing one positively worded item and one negatively worded item.

The reliability analysis yielded a Cronbach's alpha of .66, a person reliability of .66, and an item reliability of 1.00. Compared with Huda and Rohaeti (2024), distributing negatively worded items across all CTS dimensions through matched item pairs was not accompanied by higher Cronbach's alpha or person reliability. This finding suggests that, in the present sample, balancing positively and negatively worded items did not necessarily improve internal consistency or person reliability. However, the differences between the two studies cannot be attributed solely to the use of matched item pairs because they involved different sample characteristics and measurement procedures. The item reliability of 1.00 indicates that the hierarchy of item difficulty was estimated with a high degree of stability, whereas the person separation value of 1.40 suggests that the instrument had limited ability to distinguish respondents across different levels of computational thinking. A direct comparison of person separation was not possible because this statistic was not reported in their study.

Several studies have shown that combining positively and negatively worded items does not necessarily improve the reliability of the instrument. Negatively worded items may produce lower internal consistency than instruments in which all items are directly written (Barnette, 2000). Some respondents provide inconsistent answers to reverse-worded items because they require more complex cognitive processing (Swain et al., 2008). Negatively worded items may also fail to prevent response bias. Although adding items increases the length of an instrument, the inclusion of negatively worded items may reduce the average inter-item correlation (van Sonderen et al., 2013). Negatively worded items may disrupt a unidimensional structure and generate response variation that is more closely related to wording direction than to the construct being measured (Suárez-Álvarez et al., 2018). Respondents must not only understand the content of a statement but also recognize its negative direction and adjust their answers to the available response categories accordingly. This increases the likelihood of response errors and inconsistencies.

Matched positively and negatively worded items within an instrument are not necessarily psychometrically equivalent. Most matched pairs of positively and negatively worded items yield different scores, even though both statements are intended to measure the same construct (Sweeney et al., 1996). Positively worded items tend to have better discriminatory power, and extreme responses to positively worded items are not always followed by equally extreme responses in the opposite direction to negatively worded items (Matlock et al., 2018). Therefore, linguistic

reversal does not necessarily produce a complete reversal in psychometric functioning. In addition, models that combine positively and negatively worded items without accounting for wording-related method factors tend to show poorer fit (Zeng et al., 2020).

The algorithmic-thinking dimension requires further attention. Although only S8 belonged to the five most difficult items, three items in this dimension (S8, S9, and S12) showed statistically significant but small gender DIF contrasts below the 0.50-logit threshold. Although these findings do not indicate substantively meaningful gender DIF, their concentration within this dimension warrants further examination of item content and interpretation across respondent groups. Huda and Rohaeti (2024) similarly reported that algorithmic thinking was the CT dimension least frequently demonstrated by science-track senior high school students, although their study found no gender-based DIF in the original Indonesian CTS. Future studies should therefore examine whether male and female respondents understand algorithmic-thinking items equivalently, while instructional development may emphasize modeling, data representation, organized solution steps, and pattern-based problem-solving.

This study provides provisional support for the modified CTS as a self-report measure of CT tendencies. Its novelty lies in extending the Indonesian CTS from an uneven wording structure to 19 content-matched, positively and negatively worded item pairs distributed across all five dimensions. This design provided a balanced representation of both wording directions within each dimension, unlike the earlier instrument's uneven distribution. Acceptable item-level fit alone does not demonstrate psychometric equivalence within each matched pair. In addition, balancing positively and negatively worded items did not improve person reliability. Therefore, the instrument demonstrated acceptable item-level functioning but limited person-level discrimination.

Theoretical and Practical Implications

From a theoretical perspective, these findings indicate that item-level Rasch fit and balanced wording direction do not provide sufficient evidence that positively and negatively worded items function equivalently. Wording direction should therefore be treated as a potential source of response variation related to method effects and examined empirically. From a practical perspective, instrument developers should not add negatively worded items solely to reduce acquiescent response tendencies without evaluating their effects on reliability, dimensionality, and item functioning. The person reliability of 0.66 and person separation of 1.40 indicate that the modified CTS still has a limited ability to distinguish CT tendencies among respondents. Items S8, S9, and S12, as well as items with relatively low point-measure correlations or high positive or low negative ZSTD values, should be reviewed before the instrument is used more widely.

Limitations and Recommendations

This study has several limitations. The participants were volunteers from three purposively selected schools in the Surakarta Residency; therefore, the findings cannot yet be broadly generalized to other student populations. The Rasch analysis evaluated individual-item fit but did not directly establish psychometric equivalence within the 19 matched pairs. Furthermore, the DIF analysis was limited to gender, and the gender-related findings require replication using an independent sample. Future studies should refine the identified items and examine wording effects, matched-pair equivalence, dimensionality, and local dependence. They should also investigate DIF across other respondent characteristics and evaluate the stability of the modified CTS in independent samples representing broader and more diverse student populations.

Conclusion

The modified CTS demonstrated adequate content validity, appropriate functioning of the five response categories, acceptable item-level fit, and a stable item-difficulty hierarchy. All 38 items met the primary outfit MNSQ criterion, and no substantively meaningful gender DIF was identified. However, the person reliability of 0.66 and person separation of 1.40 indicated moderate reliability and a limited ability to distinguish CT tendencies among respondents. Thus, the research hypothesis was partially supported. The main contribution of this study is the development of 19 content-matched positive–negative item pairs distributed across all five CT dimensions. Further studies should refine the identified items and examine wording effects, matched-pair equivalence, dimensionality, local dependence, and stability across independent samples representing broader and more diverse student populations.

Acknowledgements

The authors used OpenAI's ChatGPT to review sentence clarity, translate selected passages from Indonesian into English, and improve the manuscript's organization and readability. Grammarly was then used to check English grammar. The authors independently developed the research questions, designed the study, collected and analyzed the data, interpreted the Rasch results, verified the references, and formulated the arguments and conclusions. All AI-assisted text was reviewed and revised by the authors, who remain fully responsible for the final manuscript.

Bibliography

Aho, A. V. (2012). Computation and computational thinking. *The Computer Journal*, *55*(7), 832–835. https://doi.org/10.1093/comjnl/bxs074

Angeli, C., & Giannakos, M. (2020). Computational thinking education: Issues and challenges. *Computers in Human Behavior*, *105*, 106185. https://doi.org/10.1016/j.chb.2019.106185

Barnette, J. J. (2000). Effects of stem and likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60*(3), 361–370. https://doi.org/10.1177/00131640021970592

Barr, V., & Stephenson, C. (2011). Bringing computational thinking to K-12: what is Involved and what is the role of the computer science education community? *ACM Inroads*, *2*(1), 48–54. https://doi.org/10.1145/1929887.1929905

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model*. Psychology Press. https://doi.org/10.4324/9781410614575

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

Branch, R. M. (2009). *Instructional design: The ADDIE approach*. Springer US. https://doi.org/10.1007/978-0-387-09506-6

Chan, S. W. (2020). Computational thinking activities in number patterns: A study in a Singapore secondary school. *ICCE 2020 - 28th International Conference on Computers in Education, Proceedings*, *1*, 171–176.

Cheng, L. (2023). The effects of computational thinking integration in STEM on students' learning performance in K-12 education: A Meta-analysis. *Journal of Educational Computing Research*, *61*(2), 416–443. https://doi.org/10.1177/07356331221114183

Chongo, S., Osman, K., & Nayan, N. A. (2021). Impact of the plugged-in and unplugged chemistry computational thinking modules on achievement in chemistry. *Eurasia Journal of Mathematics, Science and Technology Education*, *17*(4), em1953. https://doi.org/10.29333/ejmste/10789

CSTA, & ISTE. (2011). *Operational definition of computational thinking for K–12 education*. http://www.iste.org/docs/pdfs/Operational-Definition-of-Computational-Thinking.pdf

Denning, P. J. (2009). The profession of IT beyond computational thinking. *Communications of the ACM*, *52*(6), 28–30. https://doi.org/10.1145/1516046.1516054

Huda, N., & Rohaeti, E. (2024). Computational thinking skill level of senior high school students majoring in natural science. *International Journal of Learning, Teaching and Educational Research*, *23*(1), 339–359. https://doi.org/10.26803/ijlter.23.1.17

Hurt, T., Greenwald, E., Allan, S., Cannady, M. A., Krakowski, A., Brodsky, L., Collins, M. A., Montgomery, R., & Dorph, R. (2023). The computational thinking for science (CT-S) framework: operationalizing CT-S for K–12 science education researchers and educators. In *International Journal of STEM Education*, *10*(1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1186/s40594-022-00391-7

Khine, M. S. (2020). *Rasch measurement: Applications in quantitative educational research*. Springer Singapore. https://doi.org/10.1007/978-981-15-1800-3

Korkmaz, Ö., Çakir, R., & Özden, M. Y. (2017). A validity and reliability study of the computational thinking scales (CTS). *Computers in Human Behavior*, *72*, 558–569. https://doi.org/10.1016/j.chb.2017.01.005

Li, Y., Schoenfeld, A. H., diSessa, A. A., Graesser, A. C., Benson, L. C., English, L. D., & Duschl, R. A. (2020). Computational thinking is more about thinking than computing. *Journal for STEM Education Research*, *3*(1), 1–18. https://doi.org/10.1007/s41979-020-00030-2

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

Matlock, K. L., Turner, R. C., & Gitchel, W. D. (2018). A study of reverse-worded matched item pairs using the generalized partial credit and nominal response models. *Educational and Psychological Measurement*, *78*(1), 103–127. https://doi.org/10.1177/0013164416670211

Matsumoto, P. S., & Cao, J. (2017). The development of computational thinking in a high school chemistry course. *Journal of Chemical Education*, *94*(9), 1217–1224. https://doi.org/10.1021/acs.jchemed.6b00973

Ogegbo, A. A., & Ramnarain, U. (2022). A systematic review of computational thinking in science classrooms. *Studies in Science Education*, *58*(2), 203–230. https://doi.org/10.1080/03057267.2021.1963580

Setiawati, F. A., Nurhayati, S. R., Amelia, R. N., & Darojat, A. A. (2022). Study on the threats of reverse-worded items to the psychometric properties of the marital quality scale. *The Open Psychology Journal*, *15*(1), e187435012208150. https://doi.org/10.2174/18743501-v15-e2208150

Sondakh, D. E., Pungus, S. R., & Putra, E. Y. (2022). Indonesian undergraduate students' perception of their computational thinking ability. *CogITo Smart Journal*, *8*(1), 68–80. https://doi.org/10.31154/cogito.v8i1.387.68-80

Suárez-Álvarez, J., Pedrosa, I., Lozano, L., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema*, *2*(30), 149–158. https://doi.org/10.7334/psicothema2018.33

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan (Application of Rasch modeling in educational assessment)*. Trim Komunikata.

Suryaningsih, N. M. A., Poerwati, C. E., Lestari, P. I., & Parwata, M. Y. (2025). Early childhood literacy skills: Implementation of the local genius-based STEM learning model. *Indonesian Journal of Educational Development (IJED)*, *6*(1), 160–172. https://doi.org/10.59672/ijed.v6i1.4627

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed likert items. *Journal of Marketing Research*, *45*(1), 116–131. https://doi.org/10.1509/jmkr.45.1.116

Sweeney, C. T., Pillitteri, J. L., & Kozlowski, L. T. (1996). Measuring drug urges by questionnaire: Do not balance scales. *Addictive Behaviors*, *21*(2), 199–204. https://doi.org/10.1016/0306-4603(95)00044-5

Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. *Computers & Education*, *148*, 103798. https://doi.org/10.1016/j.compedu.2019.103798

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLoS ONE*, *8*(7). https://doi.org/10.1371/journal.pone.0068967

Voogt, J., Fisser, P., Good, J., Mishra, P., & Yadav, A. (2015). Computational thinking in compulsory education: Towards an agenda for research and practice. *Education and Information Technologies*, *20*(4), 715–728. https://doi.org/10.1007/s10639-015-9412-6

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology*, *25*(1), 127–147. https://doi.org/10.1007/s10956-015-9581-5

Widana, I. W., Sopandi, A. T., Suwardika, I. G. (2021). Development of an authentic assessment model in mathematics learning: A science, technology, engineering, and mathematics (STEM) approach. *Indonesian Research Journal in Education,* *5*(1), 192-209. https://doi.org/10.22437/irje.v5i1.12992

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

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

Yudi Hartawan, I. G. N., Pujawan, I. G. N., & Wibawa, N. A. (2026). An exploration of teachers' perspectives on computational thinking in mathematics learning. *Indonesian Journal of Educational Development (IJED)*, *6*(4), 1173–1188. https://doi.org/10.59672/ijed.v6i4.5594

Zeng, B., Wen, H., & Zhang, J. (2020). How does the valence of wording affect features of a scale? The method effects in the undergraduate learning burnout scale. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.585179