CHAPTER III
RESEARCH METHODS

3.1 Reseach Design

This study employs the Development-Based Research (DBR) method with the research design Analysis, Design, Develop, Implement, Evaluate (ADDIE), which is also used as the stages in the development and implementation of ARSEN (Molenda, 2003). The research design specifically can be seen in Figure 3.1. The first stage in this research is Analysis. At this stage, a thorough literature review and field study are conducted to understand students' alternative conceptions and the specific needs for tools that can identify these conceptions. The literature review draws from reputable journal sources that discuss methods for identifying students' misconceptions related to the concept of heat and strategies for developing ARSEN (Augmented Reality for Special Education Needs). Simultaneously, the field study gathers data on the learning needs of students in inclusive educational settings. This dual approach ensures a comprehensive understanding of the requirements for ARSEN's development.

Figure 3. 1. ADDIE Research Design

The second stage is Design. Building on the findings from the analysis stage, ARSEN is conceptualized with key components: augmented reality simulations, teaching materials, and diagnostic instruments to identify students' conceptions. The design process is guided by the identified needs, ensuring the tool is tailored for differentiated and inclusive learning environments. The structure and content of ARSEN are meticulously planned to address diverse learning styles and support the understanding of physics concepts, particularly heat.

In the third stage, Development, the ARSEN prototype is created based on the design specifications. This involves developing the augmented reality simulations, assembling the teaching materials, and crafting the diagnostic tools. Once the prototype is complete, it undergoes rigorous testing to evaluate its usability, functionality, and educational effectiveness. Feedback from this testing phase is used to refine and enhance ARSEN, ensuring it meets the desired educational standards and goals.

The fourth stage is Implementation. At this stage, the refined version of ARSEN is introduced and tested in a real-world educational setting. The implementation takes place at an inclusive high school in Bandung, West Java, where students study physics. This phase focuses on how effectively ARSEN supports differentiated and inclusive learning, addressing the diverse needs of students and improving their understanding of the concept of heat.

Finally, the fifth stage is Evaluation. This stage assesses the impact of ARSEN on students' learning outcomes. Specifically, it evaluates the improvement in students' conceptions and the tool's effectiveness in facilitating differentiated and inclusive learning. The evaluation employs Rasch analysis, a robust statistical method for measuring learning progress and validating the reliability of the diagnostic instruments. The insights gained from this evaluation will inform future iterations of ARSEN, ensuring its continual improvement and broader applicability in inclusive education.

3.2 Population and Sampling

The population in this study consists of thirteen students enrolled in an inclusive high school in Bandung, West Java. Since these thirteen students represent the entirety of the student body at the school, they are considered the population rather than a sample. In this context, no sampling technique is required, as all students within the school are included in the study. This approach ensures that the research captures the full diversity and characteristics of the student body within this inclusive educational setting.

In addition to the students, a group of experts specializing in physics education, special needs education, and educational technology will participate in the validation of ARSEN (Augmented Reality for Special Education Needs) as an AR-based learning media. These experts serve as a secondary sample in the study, as their input is critical to the validation process. Each expert will evaluate ARSEN using a structured validation sheet designed to assess its quality and effectiveness. The evaluation process considers several dimensions, including visual design, functional capabilities, accessibility features, usability, and alignment with the physics concept of heat.

The experts will carefully examine how ARSEN's features and interface design support an interactive and engaging learning experience for diverse student populations, particularly those with special educational needs. Their assessments provide essential feedback on the appropriateness of ARSEN for enhancing students' scientific conceptions of heat. This ensures that the learning media not only meets educational standards but also offers an inclusive and adaptable environment for all learners.

While the findings from the student population offer valuable insights into the specific educational environment of this inclusive school, they may not be directly generalizable to other schools or larger populations without further contextual analysis. Similarly, the feedback from the expert validation process is specific to ARSEN's design and intended use in inclusive science learning, which may require additional adaptation for broader applications.

By involving both the entire student population and a sample of experts, this research provides a comprehensive evaluation of ARSEN's impact on learning experiences and its potential as an innovative learning media for inclusive physics education.

3.3 Research Instruments

Research instruments are tools or facilities used by researchers to collect data, making their work easier and the results more accurate, comprehensive, and

systematic, thus simplifying the data processing (Rampean & Rohaeti, 2025). The instruments used in this study consist of test and non-test instruments.

3.3.1 Heat Concept Inventory Four-Tier Test Instrument

The four-tier test instrument is a test format with four levels, designed to diagnose students' conceptions about a specific physics concept. The test used in this study is the Heat Concept Inventory Four-Tier Test (HCIF-TT), developed using the ADDIE model (Analysis, Design, Develop, Implementation and Evaluation). The four levels of the test include the student's concept answer, their confidence level in that answer, their reasoning for the answer, and their confidence level in the reasoning. This instrument enables teachers to understand the depth of students' conceptual understanding, review the severity of alternative conceptions, identify concepts requiring further emphasis, and plan instruction to enhance students' scientific conceptions (Samsudin et al., 2024). Consequently, physics instruction can be more effective in enhancing students' scientific conceptions of heat. The development of the HCIF-TT followed the structured stages of the ADDIE model:

1. Analysis

   A field study was conducted at an inclusive school in Bandung, West Java, involving 13 students. This stage aimed to explore students' conceptions of heat using a standardized three-tier diagnostic test. The third tier collected open-ended student reasoning, which was carefully analyzed to identify common patterns of misconceptions and scientifically accurate understandings.

2. Design

   Based on the results of the initial study, the design of the HCIF-TT was constructed. This involved mapping concept indicators, developing a test blueprint, and constructing question items aligned with key heat concepts—such as temperature, heat, heat transfer, thermal expansion, specific heat capacity, conduction, and radiation.

3. Develop

   The students' written justifications from the third tier of the three-tier test were adapted to develop the reasoning options for the third tier in the HCIF-TT. This process transformed open-ended responses into structured multiple-choice reasoning options, which ensured content relevance and contextual accuracy. A panel of experts (comprising physics educators and inclusive education specialists) reviewed and validated the instrument to ensure content validity. Revisions were made based on the feedback provided.

4. Implementation

   A limited trial of the developed HCIF-TT instrument was conducted with 33 students from a regular senior high school in Bandung. The trial aimed to evaluate the instrument’s construct validity and reliability.

5. Evaluation

   Data from the implementation were analyzed using the Rasch model to evaluate the instrument's psychometric properties. The evaluation included content and construct validity, reliability analysis, item difficulty, and item discrimination.

3.3.1.1 Validity Testing

Validity testing is a step to evaluate the content of an instrument to measure its accuracy in research (Leacock & Nesbit, 2007; N. Nieveen, 1999). Validity testing is divided into two types: content validity and construct validity.

3.3.1.1.1 Content Validity

Content validity is assessed by experts evaluating the developed test instrument. For the Heat Concept Inventory Four-Tier Test (HCIF-TT), the validators include three university lecturers and two physics teachers. The evaluation results from each validator are analyzed using Multifaceted Rasch Measurement (MFRM), examining criteria such as item fit, observed average, and reliability as shown by the Minifac software. If the obtained values meet the expectations of all experts, indicated by the logit ruler, the content in HCIF-TT can

be declared valid. Any indicators falling below expert expectations will undergo revision. The HCIF-TT content validity analysis, based on the Wright Map generated through the Minifac software (see Figure 3.2), confirms that HCIF-TT is a valid assesment tool for revealing students' scientific concepts. The analysis highlights the evaluation of key indicators such as aligned with misconceptions, expert concept match, and measures understanding, proper language use, student-friendly language, logical options, single correct answer, no clues, and no "All Correct/All Wrong".

The Wright Map visualization provides valuable insight into how each indicator aligns with expert expectations. Indicators such as **"**Expert Concept Match", "Measures Understanding", and "Proper Language Use" are placed above the mean logit value, signifying strong agreement among validators regarding their importance and quality within the HCIF-TT. These indicators affirm that the test is scientifically grounded, effectively measures conceptual understanding, and is articulated in a manner appropriate for students.

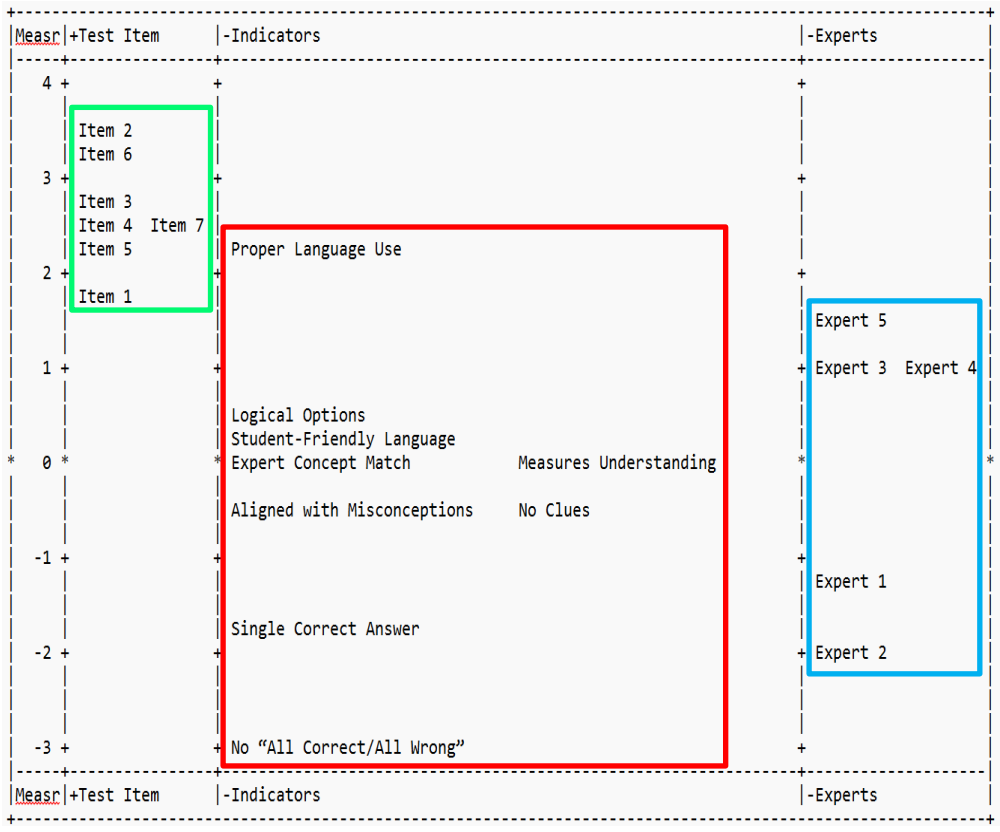
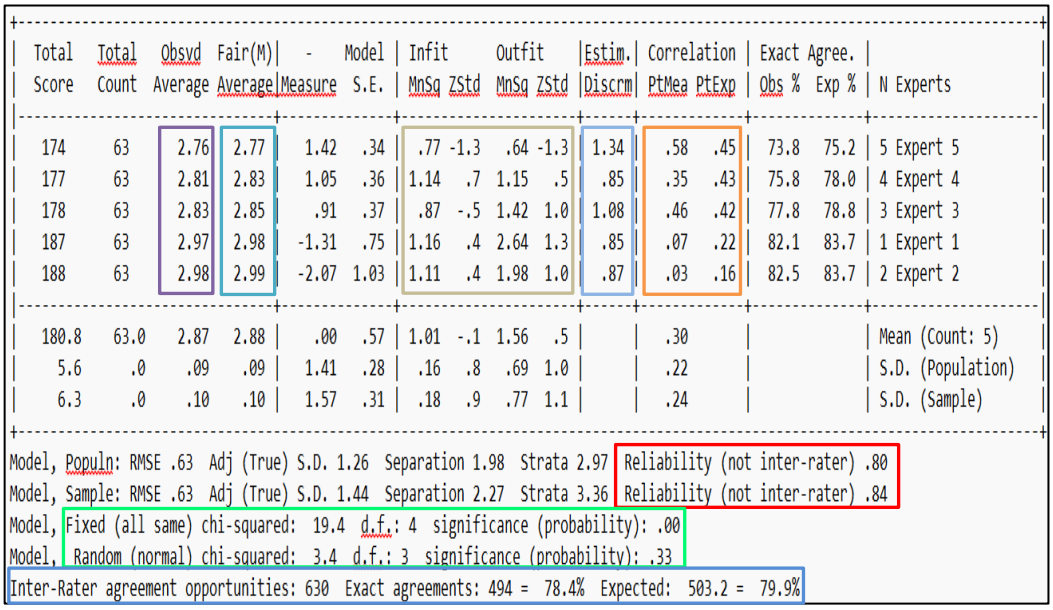


Figure 3. 2. Wright Map of HCIF-TT Content Validy

Moreover, indicators like **"**Aligned with Misconceptions" and "No Clues" are situated around the mean logit value (0 logits), indicating that while they meet expectations, slight refinements may further enhance their clarity and diagnostic power. These indicators are particularly essential in a four-tier diagnostic test like HCIF-TT, where the goal is to uncover specific student misconceptions rather than just measure correct knowledge.

Other indicators, including "Logical Options", "Student-Friendly Language", and "Single Correct Answer", are positioned just below the mean logit, suggesting moderate alignment with expert expectations. While these indicators are functionally effective, the Wright Map reveals opportunities to enhance the test's accessibility and clarity, especially for diverse student populations. Notably, the "No 'All Correct/All Wrong'" indicator appears on the lower end of the logit scale, indicating expert concern about its current implementation. This reflects a potential issue in distractor balance or scoring structure that could affect the diagnostic precision of the test. Expert distribution across the Wright Map confirms consistent scoring behavior, as no extreme inconsistencies are evident. Experts 3 and 4, for example, align closely with the central logit band, indicating calibration with the test's design principles. Experts 1 and 2, while more critical, help to highlight specific areas where the test may not fully meet the standards of robust assessment design. Such input provides a healthy spectrum of perspectives that enhances the instrument's validity through rigorous review.

To further strengthen the content validity of HCIF-TT, a detailed examination of expert rating behavior was conducted using Rasch analysis as seen in Figure 3.3. The figure shows that the observed average ratings by each expert ranged from 2.76 to 2.98, while the corresponding model fair average values ranged similarly from 2.77 to 2.99, suggesting consistent scoring behavior among the raters.



Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	N Experts
174	63	2.76	2.77	1.42	.34	.77	-1.3	.64	-1.3	1.34	.58	.45	73.8	75.2	5 Expert 5
177	63	2.81	2.83	1.05	.36	1.14	.7	1.15	.5	.85	.35	.43	75.8	78.0	4 Expert 4
178	63	2.83	2.85	.91	.37	.87	-.5	1.42	1.0	1.08	.46	.42	77.8	78.8	3 Expert 3
187	63	2.97	2.98	-1.31	.75	1.16	.4	2.64	1.3	.85	.07	.22	82.1	83.7	1 Expert 1
188	63	2.98	2.99	-2.07	1.03	1.11	.4	1.98	1.0	.87	.03	.16	82.5	83.7	2 Expert 2
180.8	63.0	2.87	2.88	.00	.57	1.01	-.1	1.56	.5		.30				Mean (Count: 5)
5.6	.0	.09	.09	1.41	.28	.16	.8	.69	1.0		.22				S.D. (Population)
6.3	.0	.10	.10	1.57	.31	.18	.9	.77	1.1		.24				S.D. (Sample)

Model, Populn: RMSE .63 Adj (True) S.D. 1.26 Separation 1.98 Strata 2.97 Reliability (not inter-rater) .80
Model, Sample: RMSE .63 Adj (True) S.D. 1.44 Separation 2.27 Strata 3.36 Reliability (not inter-rater) .84
Model, Fixed (all same) chi-squared: 19.4 d.f.: 4 significance (probability): .00
Model, Random (normal) chi-squared: 3.4 d.f.: 3 significance (probability): .33
Inter-Rater agreement opportunities: 630 Exact agreements: 494 = 78.4% Expected: 503.2 = 79.9%

Figure 3. 3. Summary Statistic of HCIF-TT Content Validy.

This is further supported by the minimal standard error (SE), which stays under 1.05 across all experts. Infit and outfit mean square (MnSq) statistics largely remain within acceptable thresholds (0.5 to 1.5), with minor deviations. For instance, Expert 5 shows a low infit (MnSq = 0.77) and outfit (MnSq = 0.64), indicating slightly overfitting behavior—meaning their ratings are more predictable than expected. Conversely, Expert 1 and Expert 2 show slightly higher outfit values (MnSq = 2.64 and 1.98 respectively), which may suggest occasional unexpected responses (Bond & Fox, 2013; Eckes, 2023). However, the standardized fit (ZStd) values across experts mostly remain within the ±2 range, indicating acceptable fit to the Rasch model.

From the discrimination index (Discrm), all experts except Expert 5 exhibit moderate discriminative ability (ranging from 0.85 to 1.08), while Expert 5 shows slightly higher discrimination (1.34), indicating their responses more sharply differentiate between high- and low-quality indicators (Linacre, 2002; Wright & Masters, 1982). However, the point-measure correlations (PtMea and PtExp) fluctuate, with Experts 1 and 2 showing lower correlations—highlighting a possible misalignment in their evaluations relative to the Rasch-predicted hierarchy of indicators. Inter-rater agreement was another critical aspect of the analysis. The

observed exact agreement among the raters was 78.4%, which exceeded the minimum acceptable threshold of 50% (Eckes, 2023). This level of agreement highlighted a moderate degree of consistency among the raters, although it also indicated room for improvement in achieving complete uniformity.

The model summary statistics further reinforce the instrument's quality. The separation values of 1.98 (population) and 2.27 (sample) imply that the instrument can distinguish approximately three strata of expert severity or strictness in rating. Reliability indices are robust, at 0.80 (population) and 0.84 (sample), indicating that the data from expert ratings are sufficiently reliable for further interpretation and refinement of the instrument (Sumintono, 2018). The fixed (all same) chi-square test for differences among experts yields a significant result ($\chi^2 = 19.4$, df = 4, $p < 0.01$), confirming that there are statistically significant differences in the way the five experts rate the indicators. This diversity, however, is valuable for uncovering inconsistencies and improving the assessment instrument. The random (normal) chi-square statistic ($\chi^2 = 3.4$, df = 3, $p = 0.33$) shows that variations are within expected norms under the assumption of a normal distribution, suggesting no extreme outliers among expert ratings.

In conclusion, the HCIF-TT demonstrates strong content validity, with most indicators scoring within or above the expected range. The test's strengths lie in its alignment with scientific concepts, ability to detect misconceptions, and overall pedagogical clarity. Areas scoring lower on the Wright Map, such as balance in answer key patterns and language accessibility, offer constructive feedback for further revision. Collectively, these findings affirm that the HCIF-TT is a well-developed instrument for diagnosing students' heat conceptions and guiding conceptual change through targeted instructional interventions.

3.3.1.1.2 Construct Validity, Fit Statistic and Difficulty Level

Construct validity of the HCIF-TT instrument was examined through empirical testing with a sample of 33 students from a regular senior high school in Bandung. This evaluation aimed to assess the instrument's construct validity, item fit statistics, item difficulty levels, and reliability using Ministep software. The

Rasch model was employed for the analysis, focusing on key indicators such as unidimensionality, Outfit Mean Square (MNSQ), Outfit Z-Standardized (ZSTD), Point-Measure Correlation (Pt-Measure Corr), and logit measures to establish the validity and measurement precision of the instrument. Reliability is discussed in the following section.

Unidimensionality is a fundamental assumption in Rasch modeling and must be confirmed before interpreting other statistical outputs such as item difficulty and fit indices. Based on the unidimensionality analysis (see Figure 3.4), the HCIF-TT instrument satisfies key assumptions of the Rasch model. The raw variance explained by the measures reached 65.0%, categorizing it as *"Excellent"* based on established criteria (>60%). The unexplained variance in the first contrast was recorded at 2.4493 eigenvalue units (below the 3.0 threshold), and the observed variance for the first contrast was 12.3%, well under the 15% limit. These indicators collectively support the conclusion that the HCIF-TT instrument is unidimensional and valid for assessing students' scientific conceptions.

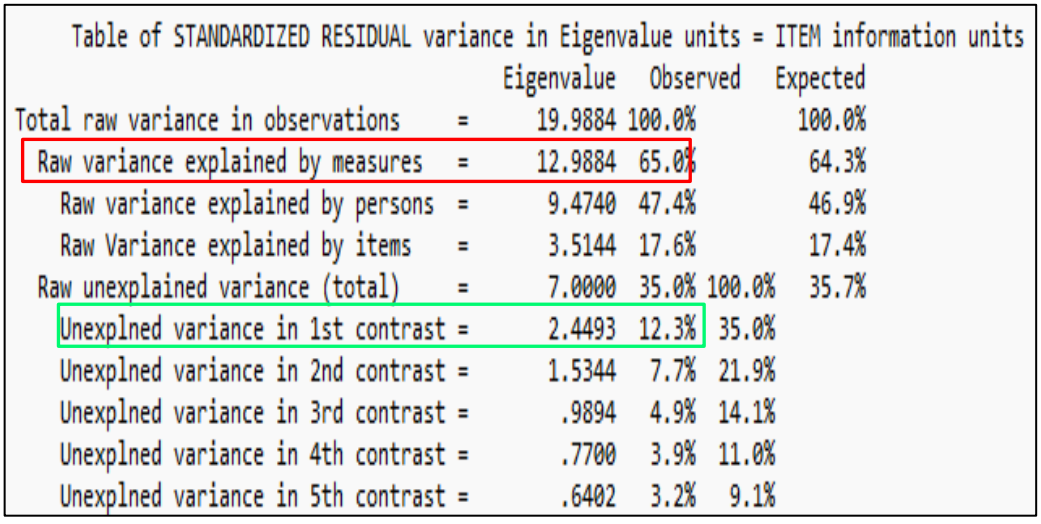


Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	19.9884	100.0%		100.0%
Raw variance explained by measures =	12.9884	65.0%		64.3%
Raw variance explained by persons =	9.4740	47.4%		46.9%
Raw Variance explained by items =	3.5144	17.6%		17.4%
Raw unexplained variance (total) =	7.0000	35.0%	100.0%	35.7%
Unexplned variance in 1st contrast =	2.4493	12.3%	35.0%	
Unexplned variance in 2nd contrast =	1.5344	7.7%	21.9%	
Unexplned variance in 3rd contrast =	.9894	4.9%	14.1%	
Unexplned variance in 4th contrast =	.7700	3.9%	11.0%	
Unexplned variance in 5th contrast =	.6402	3.2%	9.1%	

Figure 3. 4. Unidimensionality of HCF-TT.

The Wright Map generated from the Rasch analysis (see Figure 3.5) offers strong evidence for the construct validity of the HCIF-TT instrument. This map presents a joint distribution of person abilities (on the left) and item difficulties (on the right) along a common logit scale, allowing for a direct and meaningful

comparison. The distribution reveals that the items are appropriately dispersed across the measurement continuum and align well with the students' ability levels.

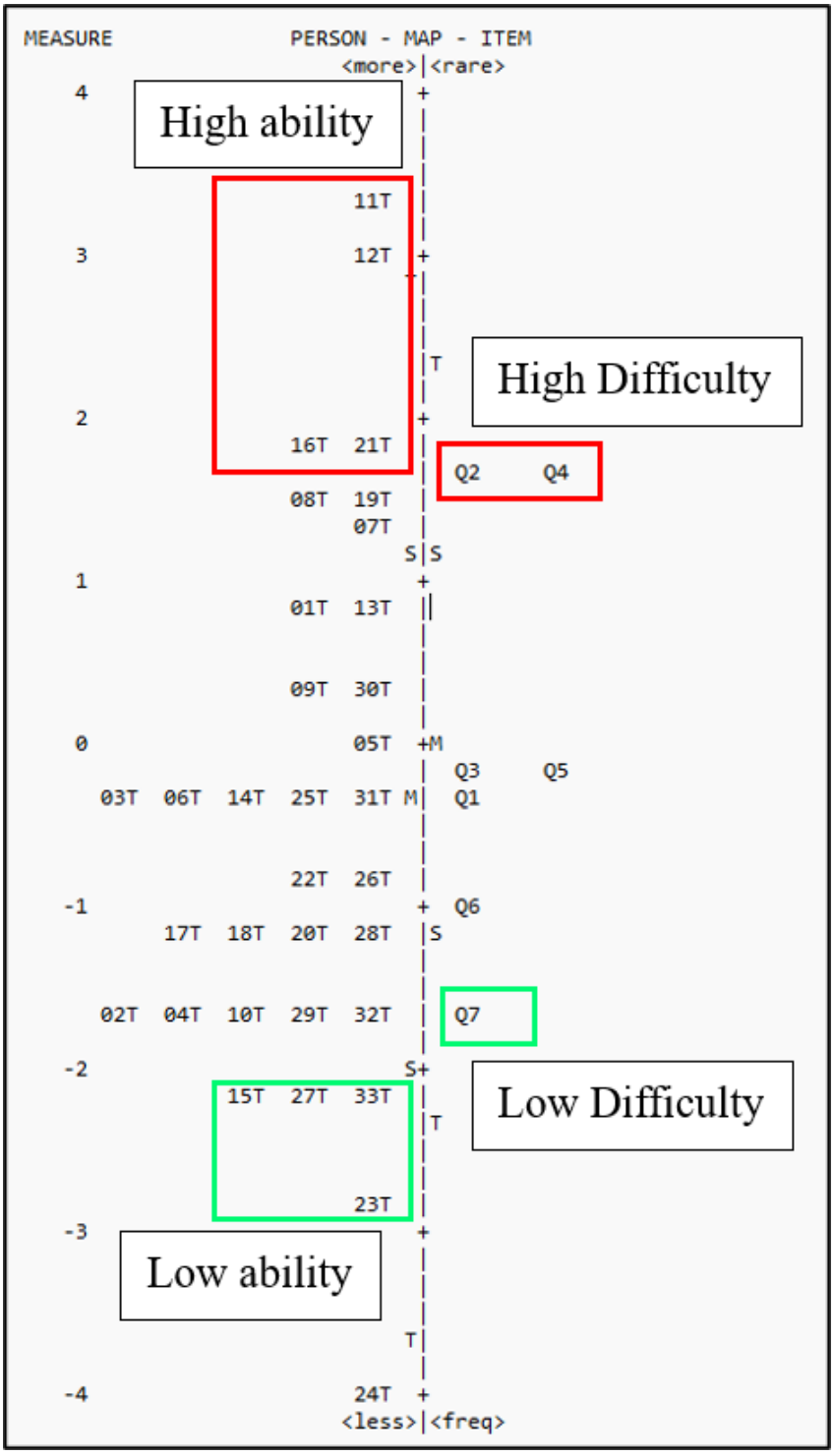


Figure 3. 5. Wright Map of HCF-TT.

Specifically, Item Q2 and Item Q4 appear at the higher end of the scale, indicating that they were more challenging and likely required a higher level of scientific conceptual understanding. Conversely, Item Q6 and Item Q7 are located at the lower end, suggesting they were relatively easier for most students. The remaining items occupy intermediate positions, reflecting a range of difficulty levels suitable for differentiating among students' conceptual proficiencies.

Based on the item statistics as presented in Figure 3.6 below, Item Q2 (measure = +1.74) and Item Q4 (measure = +1.63) were identified as the most difficult items, suggesting that these items require a higher level of conceptual understanding. On the other hand, Item Q7 (measure = -1.60) and Item Q6 (measure = -1.08) were categorized as the easiest items, indicating they are more accessible to students with lower levels of conceptual mastery. The remaining items fall within a moderate difficulty range, further supporting the instrument's ability to capture a diverse spectrum of student understanding.

ITEM STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
2	35	33	1.74	.32	.46	-2.39	.42	-2.63	.82	.66	84.8	66.4	Q2
4	36	33	1.63	.32	1.01	.15	1.02	.17	.44	.66	60.6	66.5	Q4
3	56	33	-.16	.28	1.05	.27	1.09	.40	.83	.74	63.6	59.1	Q3
5	57	33	-.23	.27	1.10	.45	1.12	.49	.64	.74	45.5	58.8	Q5
1	58	33	-.31	.27	1.30	1.15	1.24	.89	.88	.75	60.6	58.8	Q1
6	69	33	-1.08	.26	1.18	.76	1.26	.98	.63	.76	36.4	54.4	Q6
7	77	33	-1.60	.25	.64	-1.56	.60	-1.55	.82	.76	60.6	54.6	Q7
MEAN	55.4	33.0	.00	.28	.96	-.17	.96	-.18			58.9	59.8	
P.SD	14.4	.0	1.17	.03	.28	1.20	.30	1.27			14.1	4.6	

Figure 3. 6. Item Statistics of HCF-TT.

The analysis of item difficulty using the Rasch model further confirms the appropriateness of the HCIF-TT instrument. The item logit values ranged from -1.60 to +1.74, indicating a good spread across varying levels of conceptual difficulty. Items with higher positive logit values are considered more challenging,

as they are positioned above the mean ability level (0 logits), whereas items with negative logit values are relatively easier and lie below the mean student ability.

Item fit statistics were evaluated using three indicators: Outfit MNSQ, Outfit ZSTD, and Point-Measure Correlation. According to Rasch modeling guidelines, an item is considered to fit well when: (1) outfit MNSQ is within the range of 0.5 to 1.5, (2) ZSTD falls between -2.0 and +2.0, and (3) point-measure correlation ranges from 0.4 to 0.85. Based on these criteria, items are categorized in Table 3.1.

Table 3. 1. Fit Statistics of HCF-TT

Item	Outfit		PT-Measure	Fit Category
	MNSQ	ZSTD		
Q1	1.24	0.89	0.88	very well-fitting
Q2	0.42	-2.63	0.82	very well-fitting
Q3	1.09	0.40	0.83	very well-fitting
Q4	1.02	0.17	0.44	very well-fitting
Q5	1.12	0.49	0.64	very well-fitting
Q6	1.26	0.98	0.63	very well-fitting
Q7	0.60	-1.55	0.82	very well-fitting

All seven items demonstrated an excellent fit with the Rasch model, meeting all three fit criteria. This outcome signifies that each item functions effectively in measuring the intended construct of scientific conception and contributes meaningfully to the overall instrument quality without introducing statistical noise or distortion.

3.3.1.2 Reliability Testing

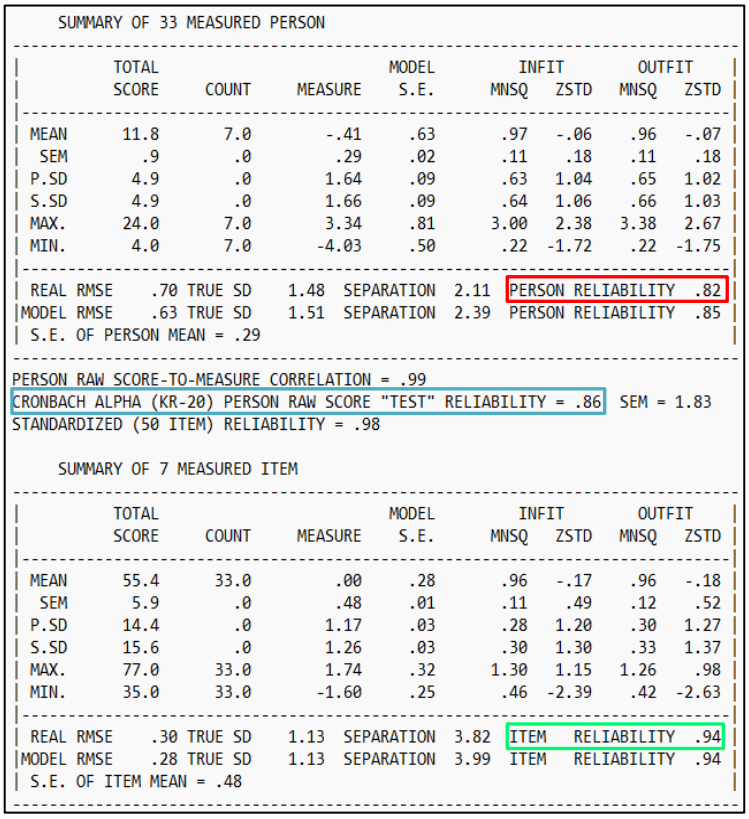
Instrument reliability refers to the consistency of an instrument in measuring during research or the consistency of respondents in answering the test questions (Bond & Fox, 2013; Eckes, 2023; Linacre, 2002; Wright & Masters, 1982). Repeated measurements should yield consistent or identical results. Consistent reliability indicates that an instrument administered to the same individuals at

different times will produce similar outcomes. This equivalence demonstrates that the instrument is reliable (Sumintono, 2018). The reliability level is empirically indicated by a coefficient reliability value. In this study, reliability testing is conducted using Rasch modeling analysis with the Ministep software. According to Sumintono (2018) value of reliability can be categorized as seen in Table 3.2.

Table 3. 2. Interpretation of Item, Person Reliability and Cronbach's Alpha

Summary statistic	Value	Interpretation
Item and person reliability	$\mathbf{r > 0,94}$	Excellent
	$\mathbf{0,90 < r \leq 0,94}$	Very Good
	$\mathbf{0,80 < r \leq 0,90}$	Good
	$\mathbf{0,67 < r \leq 0,80}$	Sufficient
	$\mathbf{r \leq 0,67}$	Low
Cronbach's Alpha	$\boldsymbol{\alpha \geq 0,80}$	Very High
	$\mathbf{0,70 \leq \alpha < 0,80}$	High
	$\mathbf{0,60 \leq \alpha < 0,70}$	Good
	$\mathbf{0,50 \leq \alpha < 0,60}$	Moderate
	$\boldsymbol{\alpha < 0,50}$	Low

The summary statistics presented in Figure 3.7 provide robust psychometric evidence supporting the quality and reliability of the HCIF-TT instrument. These statistics include key indicators such as item reliability, person reliability, and internal consistency measures, which are essential for establishing the soundness of the assessment tool within the Rasch measurement framework.



SUMMARY OF 33 MEASURED PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	11.8	7.0	-.41	.63	.97	-.06	.96	-.07
SEM	.9	.0	.29	.02	.11	.18	.11	.18
P.SD	4.9	.0	1.64	.09	.63	1.04	.65	1.02
S.SD	4.9	.0	1.66	.09	.64	1.06	.66	1.03
MAX.	24.0	7.0	3.34	.81	3.00	2.38	3.38	2.67
MIN.	4.0	7.0	-4.03	.50	.22	-1.72	.22	-1.75

REAL RMSE .70 TRUE SD 1.48 SEPARATION 2.11 PERSON RELIABILITY .82
MODEL RMSE .63 TRUE SD 1.51 SEPARATION 2.39 PERSON RELIABILITY .85
S.E. OF PERSON MEAN = .29

PERSON RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .86 SEM = 1.83
STANDARDIZED (50 ITEM) RELIABILITY = .98

SUMMARY OF 7 MEASURED ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	55.4	33.0	.00	.28	.96	-.17	.96	-.18
SEM	5.9	.0	.48	.01	.11	.49	.12	.52
P.SD	14.4	.0	1.17	.03	.28	1.20	.30	1.27
S.SD	15.6	.0	1.26	.03	.30	1.30	.33	1.37
MAX.	77.0	33.0	1.74	.32	1.30	1.15	1.26	.98
MIN.	35.0	33.0	-1.60	.25	.46	-2.39	.42	-2.63

REAL RMSE .30 TRUE SD 1.13 SEPARATION 3.82 ITEM RELIABILITY .94
MODEL RMSE .28 TRUE SD 1.13 SEPARATION 3.99 ITEM RELIABILITY .94
S.E. OF ITEM MEAN = .48

Figure 3. 7. Summary Statistics of HCF-TT.

The item reliability of the HCIF-TT instrument was recorded at 0.94, which falls under the classification of *Very Good*. This high level of reliability suggests that the item difficulty hierarchy is well-established and stable, indicating that the sample size is sufficient to confirm the consistency of item calibration across different administrations. The item reliability of the HCIF-TT instrument was recorded at 0.94, which falls under the classification of *Very Good*. This high level of reliability suggests that the item difficulty hierarchy is well-established and stable, indicating that the sample size is sufficient to confirm the consistency of item calibration across different administrations. In terms of person reliability, the instrument achieved a value of 0.82, which is classified as *Good*. This suggests that the HCIF-TT is effective in reliably distinguishing among students with different levels of conceptual understanding in heat and temperature. Moreover, the instrument demonstrated strong internal consistency, as evidenced by the Cronbach's alpha (KR-20) value of 0.86. This value exceeds the commonly

accepted threshold of 0.70 for educational instruments and indicates that the set of items measures a coherent construct with minimal internal error variance.

3.3.2 Interview

The interview guideline is a supporting instrument used by the researcher to guide interviews with student representatives from each class (control and experimental classes). Each class is represented by two students (one male and one female) who are considered representative of the entire class. The purpose of the interviews is to gather supporting data that provides information about students' responses to ARSEN, including its strengths and weaknesses during the learning process. Interview data offers depth and detail in describing students' experiences in the classroom (Creswell & Creswell, 2017), providing additional insights into ARSEN's characteristics.

3.4 Research Procedure

The research procedure follows the ADDIE model, which consists of five stages: Analysis, Design, Development, Implementation, and Evaluation. Each stage is detailed in the table 3.3 below.

Table 3. 3. Detailed Stages of Research Procedure

Stage	Sub-Stage	Details
Analysis	Literature Review	Conducted to understand students' alternative conceptions and the need for tools to identify them.
	Field Study	Collects data on the learning needs of students in inclusive settings.
Design	Research Instruments	Validation instruments will be used to evaluate the validity and reliability of ARSEN. This also applies to the design of HCIF-TT.
	AR Media	Involves designing AR simulations and the application interface.
	Students Worksheets	Activities for students will be designed.
Develop	Expert Validation	ARSEN and HCIF-TT will be validated by experts. If revisions are necessary,

Stage	Sub-Stage	Details
		ARSEN and HCIF-TT will be revised and undergo limited testing.
	Limited Testing	ARSEN and HCIF-TT will be tested in a small group.
	Refinement	ARSEN and HCIF-TT will be refined based on feedback from limited testing.
Implement	The ARSEN Intervention	The ARSEN intervention will be conducted at an inclusive high school in Bandung, using a one-group pretest-posttest design.
Evaluation	Enhancement of Students' Scientific Conceptions	The results from the pretest and posttest will be analyzed to determine the enhancement of students' scientific conceptions.
	The ARSEN in Inclusive Learning	The practicality of ARSEN will be analyzed, with feedback gathered from students in an inclusive classroom.

For better interpretation of the research procedure, see Figure 3.8. The image presents an overview of the development process for the ARSEN prototype, highlighting each stage from analysis to evaluation. The journey begins with the Initial Stage, where a thorough literature review is conducted. This review delves into key topics such as conceptual understanding, conceptual change, and the unique needs of inclusive students. It also explores differential learning, inclusive education practices, and student characteristics. Complementing this, a field study is carried out to gather real-world data on the learning needs of students in inclusive settings. Together, these efforts help identify the core research problems that the ARSEN prototype aims to address.

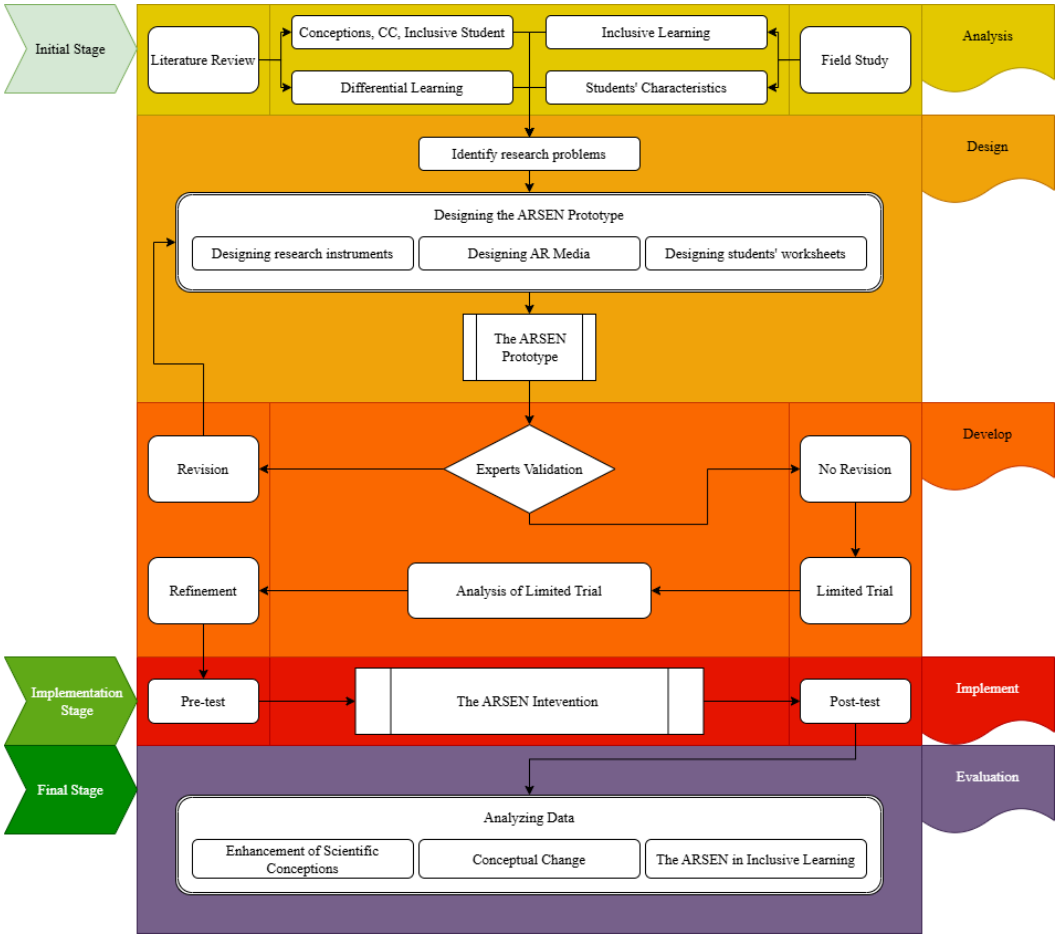


Figure 3. 8. Research Procedure

Moving into the Design Stage, the focus shifts to creating the ARSEN prototype. This involves several critical tasks, such as designing research instruments that will later be used to assess the prototype's validity and reliability. Concurrently, the development of augmented reality media and the application interface takes place, ensuring the AR experience is both engaging and educational. Additionally, worksheets tailored to the students' activities are designed to enhance their learning experience and complement the AR media.

The Development Stage marks the creation of the initial ARSEN prototype. Once developed, it undergoes expert validation, where specialists in the field evaluate its effectiveness and relevance. Depending on their feedback, the prototype may be revised to address any identified shortcomings or proceed without revisions if deemed adequate. Subsequently, a limited trial is conducted, introducing the

prototype to a small group of participants. The results from this trial are carefully analyzed, and further refinements are made based on the insights gathered.

In the Implementation Stage, the ARSEN intervention is prepared for a broader audience. Before its implementation in an inclusive high school setting, a pre-test is administered to establish a baseline understanding of the students' knowledge. Following this, the intervention takes place, and after its completion, a post-test is conducted. This allows for a direct comparison of students' knowledge before and after the intervention.

Finally, the process concludes with the Evaluation Stage, where the collected data is thoroughly analyzed. This analysis focuses on several key outcomes, including the enhancement of students' scientific conceptions and any noticeable shifts in their conceptual understanding. Additionally, the practicality of using ARSEN in inclusive learning environments is assessed, ensuring that the tool not only enhances learning but is also feasible for regular classroom use. This narrative encapsulates the meticulous process of developing and refining the ARSEN prototype, ensuring it effectively meets the educational needs of inclusive classrooms.

3.5 Data Analysis

The purpose of data analysis is to draw conclusions based on the research conducted or to establish the foundation of arguments for answering each research question outlined in Chapter 1. The data analysis in this study will cover the analysis of the characteristics (feasibility) of ARSEN, the analysis of the enhancement of students' scientific conceptions in the concept of heat, and the analysis of students' perceptions regarding the application of ARSEN in inclusive learning.

3.5.1 Characteristics Analysis of ARSEN

The characteristics of ARSEN are assessed through three criteria: validity, practicality, and effectiveness. Validity includes both content validity and construct validity, which are validated by experts through validity testing. Practicality will be assessed by students and teachers, while effectiveness is closely related to the next

section, which discusses the enhancement of students' scientific conceptions. For validity testing, Table 3.4. provide guidelines on the ARSEN validation sheet.

Table 3. 4. Scoring Guidelines for ARSEN Validity Test

Code	Explanation	Description	Score
VWR	Valid Without Revision	Can be used without revision	3
VR	Valid with Revision	Can be used with revision	2
TV	Not Valid	Cannot be used	1

The validity test will be conducted by several experts, consisting of physics lecturers and high school physics teachers specializing in physics education, inclusive learning, and educational technology. Based on Table , scores from each expert will be collected and processed using MFRM analysis with the aid of Minifac (Facets) Rasch software. The data obtained include both quantitative and qualitative data. The qualitative data consist of detailed descriptions of the revisions suggested by the experts.

The practicality of ARSEN for use is evaluated through trial implementation. The ARSEN trial sheet is a questionnaire comprising 15 positive statements on a 4-point Likert scale. The scoring guidelines used are shown in Table 3.5.

Table 3. 5 Scoring Guidelines for ARSEN Trial

Code	Explanation	Score
SA	Strongly Agree	4
A	Agree	3
D	Disagree	2
SD	Strongly Disagree	1

Based on Table 3.5, the scores from the questionnaire are processed using MFRM analysis with the aid of Minifac (Facets) Rasch software. The interpretation of practicality is determined based on the values of observed average, as shown in Table 3.6 (adapted from Astuti et al. (2022) and Hermita et al. (2020)).

Table 3. 6. Practicality Criteria for ARSEN

Percentage Requirement (%)	Level of Practicality
81-100	Very Practical
61-80	Practical
41-60	Moderate
21-40	Not Practical
0-20	Very Not Practical

In addition to quantitative data, qualitative data in the form of suggested improvements are also accommodated in this sheet, enabling improvements in practicality.

3.5.2 Analysis of the Enhancement of Students' Scientific Conceptions

The quantity of students' conceptual changes can be determined by their state before and after the treatment. The students' states in each condition are revealed using the HCIF-TT instrument through quantitative analysis during the pre-test and post-test stages. A multi-tier instrument HCIF-TT, referencing categories such as R-SEN (Robust Scientific Equitable Notion), Pa-SEN (Partial Scientific Equitable Notion), Ne-SEN (Negative Scientific Equitable Notion), Mi-SEN (Misaligned Scientific Equitable Notion), Ab-SEN (Absence of Scientific Equitable Notion), and No-SEN (None Scientific Equitable Notion). Rasch analysis and Nvivo 14 will be utilized to assess the improvement in students' conceptions. Additionally, the analysis of conceptual change will be conducted, categorized into GC (Great Change), NC (No Change), and U-GC (Un-Great Change) (Samsudin et al., 2024), as illustrated in Table 3.7.

Table 3. 7. Modified Conception Levels (see also, Aminudin et al., 2019)

Conception Levels	**1st Tier**	**2nd Tier**	**3rd Tier**	**4th Tier**	**Score**
Robust Scientific Notion (R-SEN)	Exact	Certain	Exact	Certain	4

Conception Levels	1st Tier	2nd Tier	3rd Tier	4th Tier	Score
Partial Scientific Notion (Pa-SEN)	Exact	Certain	Exact	Insecure	3
	Exact	Insecure	Exact	Certain	
	Exact	Insecure	Exact	Insecure	
Negative Scientific Notion (Ne-SEN)	Exact	Certain	Inexact	Certain	2
	Exact	Certain	Inexact	Insecure	
	Exact	Insecure	Inexact	Certain	
	Exact	Insecure	Inexact	Insecure	
	Inexact	Certain	Exact	Certain	
	Inexact	Certain	Exact	Insecure	
	Inexact	Insecure	Exact	Certain	
	Inexact	Insecure	Exact	Insecure	
Misconception (Mi-SEN)	Inexact	Certain	Inexact	Certain	1
Absence of Scientific Notion (Ab-SEN)	Inexact	Certain	Inexact	Insecure	0
	Inexact	Insecure	Inexact	Certain	
	Inexact	Insecure	Inexact	Insecure	
None of Scientific Notion (No-SEN)	None				

Based on the table above, conceptual levels are divided into six levels. Each level has its own scoring criteria. To observe the quantity of students' conceptual changes, their scores from the pre-test and post-test are collected and processed using the N-change calculation. N-change is derived from Table 3.8 (Marx & Cummings, 2007).

Table 3. 8. N-change calculation for ARSEN

Condition	Equation
Post-test > Pre-test	$c = \frac{(\text{post} - \text{test score}) - (\text{pre} - \text{test score})}{(\text{maximum score}) - (\text{pre} - \text{test score})}$
Post-test = Pre-test = 100 or 0	$C = \text{Drop}$
Post-test = Pre-test	$c = 0$

Condition	Equation
Post-test < Pre test	$c = \frac{(\text{post} - \text{test score}) - (\text{pre} - \text{test score})}{(\text{pre} - \text{test score})}$

Based on Table C.5, the N-change value for each student can be determined. The N-change values are then used to categorize the changes in students' conceptions. The determination of categories is based on Table 3.9 (adapted from (Marx & Cummings, 2007)).

Table 3. 9. N-change Category

N-change Value Requirement	Category
$0{,}7 < c \leq 1$	High
$0{,}3 < c \leq 0{,}7$	Moderate
$0 < c \leq 0{,}3$	Low
$c = 0$	No Change
$-1 < c < 0$	Negative

In addition to examining changes from the students' perspective, conceptual changes are also reviewed from the change in the level of conception for each item. To determine this, the percentage of each conceptual level for each item must first be calculated. The Enhanced Conceptions (EC) is calculated using the equation,

$$\text{EC}(\%) = \frac{\Sigma(\text{students}'\text{enhanced conceptions})}{\Sigma(\text{total students})} \times 100\%$$

To obtain the quantity of enhanced conceptions (QEC), the above equation is used to calculate the percentage of conceptual levels during the pre-test and post-test. The calculation of the percentage of quantity enhanced conceptions (QEC) for each item is derived using the equation:

$$\text{QEC}(\%) = \pm\left[\text{EC}_{\text{post}}(\%) - \text{EC}_{\text{pre}}(\%)\right]$$

The ± symbol serves as a marker to differentiate calculations between groups of conceptual levels. Groups of conceptual levels expected to show positive changes (R-SEN, Pa-SEN, and Ne-SEN) are given a positive (+) sign, while groups

of conceptual levels expected to show negative changes (Mi-SEN, Ab-SEN, No-SEN, and EC) are given a negative (-) sign. The QEC results are then interpreted into conceptual change types as shown in Table 3.10.

Table 3. 10. Categorization of Conceptual Change

QEC	Type of Conceptual Change
+	*Great Change* (GC)
–	*Un-Great Change* (U-GC)
0	*No Change* (NC)

3.5.3 Students' perceptions of ARSEN in Inclusive Learning

To evaluate the impact of ARSEN in an inclusive learning environment, semi-structured interviews were conducted to collect qualitative data on students' perceptions. This approach enabled the exploration of their experiences, thoughts, and feelings regarding the use of ARSEN in learning heat concepts.

The interview data were analyzed using thematic analysis, a method commonly employed in educational research to identify recurring patterns and themes within qualitative data. The analysis focused on key areas such as usability, engagement, and perceived effectiveness of ARSEN in supporting their conceptual understanding.

The findings from this thematic analysis provided rich, in-depth insights that complemented the quantitative results from the pre- and post-tests. Together, these data sources offered a more holistic understanding of ARSEN's effectiveness and practical application in inclusive classroom settings.