25 (1), 2024, 142-154

Jurnal Pendidikan MIPA

e-ISSN: 2685-5488 | p-ISSN: 1411-2531

http://jurnal.fkip.unila.ac.id/index.php/jpmipa/

Development and Validation of the Relativity Concept Inventory Test Using Item Response Theory Generalized Partial Credit Model 3PL

Innal Mafudi* & Edi Istiyono

Department of Physics Education, Universitas Negeri Yogyakarta, Indonesia

Abstract: This research aims to develop a valid and reliable Relativity Concept Inventory Test instrument. This instrument is based on relativity material, which includes Einstein's first and second postulates, time dilation, velocity addition, and length contraction. Methods for preparing instruments include 1) test design, 2) test validation, 3) test trials, and 4) test data analysis. The design of the test grid is based on Bloom's taxonomy C2 to C5 and produces 13 questions. The instrument is made in the form of multiple-choice questions and is equipped with a level of confidence. Instrument validation was carried out by 6 physics education lecturers and 1 high school teacher, with analysis using the V Aiken formula. The validated instrument was then tested on 130 students from 2 high schools in Madiun. Trial data was analyzed using the Generalized Partial Credit Model 3PL (GPCM-3PL). The development results show that: 1) 13 multiple choice Relativity Concept Inventory Test questions with a level of confidence were successfully developed, 2) expert validation showed that all question items got a score of 0.93, which is included in the valid, with instrument reliability of 0, 42 (very low category), 3) the results of the trial test showed that the relativity concept inventory was proven to be fit with the GPCM, with different power of the items of 0.407 and the level of difficulty showed that 11 items were valid with a range of -1.05 to 1.64, while 2 questions (numbers 8 and 9) are invalid with a value of more than -2. Apart from that, the question items have no potential to be guessed, as evidenced by the guessing value of 0 (zero). This Relativity Concept Inventory Test instrument meets the requirements for use in measuring students' conceptual understanding

Keywords: assessment, inventory concept, relativity.

▪ INTRODUCTION

Physics education at the Senior High School (SMA) level is crucial in building students' scientific knowledge framework up to the university level (Hazari et al., 2007). Albert Einstein's theory of relativity is a critical topic in the physics curriculum. This theory is the backbone of modern understanding of the universe and technological development (Hartle, 2005). However, understanding the theory of relativity is often challenging for students and teachers (Farmer, 2021). The cause is the relative effects that contradict their daily experience and intuition. This condition often causes difficulties in understanding concepts and can lead to misconceptions about relativity material (Cormier & Steinberg, 2010; Gero et al., 2019; Kulgemeyer & Wittwer, 2021; Listianingrum et al., 2022; Vicovaro, 2023). This condition must be diagnosed, and appropriate action must be taken to overcome it.

One way to find out misconceptions in students is with a concept inventory test, which is applied at the beginning and at the end of learning (Aslanides & Savage, 2013; Piacsek, 2018; Sachan et al., 2019; Siong et al., 2023). The concept inventory test is a test that is commonly used to assess learning in the field of physics (Ene & Ackerson, 2018). However, the results of literature searches show that the development of this test on relativity material in Indonesia is still very limited.

Innal Mafudi* & Edi Istiyono
*Email: Innalmafudi.2023@student.uny.ac.id

DOI: http://dx.doi.org/10.23960/jpmipa/v25i1.pp142-154
Received: 24 June 2024
Accepted: 23 July 2024
Published: 27 July 2024

Previous research related to the development of an inventory test of the concept of relativity includes several essential studies. For example, research (Scherr et al., 2001) This test successfully identified some common misconceptions but needed to improve its construct validity. Likewise, research (Aslanides & Savage, 2013) developed a relativity concept inventory test focusing on the qualitative aspects of students' understanding. Although these tests provide deep insight into student understanding, they place less emphasis on quantitative aspects that can be measured more objectively. Additionally, the test does not cover the entire range of common misconceptions and is limited in scope, which may reduce its effectiveness in diverse classroom contexts.

Research (Thacker et al., 1994) pointed out that inventory tests of the relativity concept need to consider cultural differences and educational contexts, which are often overlooked in instrument development. Specifically in Indonesia, research (Listianingrum et al., 2022) shows that misconceptions about velocity addition, time dilation, and length contraction are still often found among physics students. Although this research provides valuable insight into the types of misconceptions that occur, it has methodological weaknesses. This research is more descriptive in nature and places less emphasis on developing and validating instruments that can be used widely. In addition, this research has yet to use an in-depth quantitative approach to measure the level of misconception with high precision and has not utilized a more complex analytical model such as Item Response Theory (IRT) to evaluate the instrument.

Based on these conditions, it is essential to develop a valid and reliable inventory test of the concept of relativity. The advantage of this research lies in the empirical validation test stage, which uses IRT with the Generalized Partial Credit Model (GPCM) and Three-Parameter Logistic (3PL) with two analysis software QUEST and Parscale. The use of IRT GPCM 3PL with these two software allows for a more in-depth and comprehensive analysis of the test items. This model is able to capture information about the level of difficulty, discriminating power, and guessing level of each item, thus providing a more accurate picture of the overall quality of the test (Muraki, 1992; Samejima, 1997). The instrument developed is not only valid and reliable but also has a high level of precision in measuring students' conceptual understanding and detecting misconceptions that may exist (Baker, 2021; Embretson et al., 2006). It is hoped that this will provide an alternative for teachers to measure students' understanding of concepts accurately so that teachers obtain precise data regarding students' understanding and can determine appropriate strategies in learning activities.

▪ METHOD

Research Design and Procedures

The research is into the development of an inventory of the concept of relativity, which is based on the RCI instrument that was developed previously (Aslanides & Savage, 2013). The test instrument development model was modified from the model (D Mardapi, 2008; Istyono, 2020). The steps for developing an instrument in the form of a test are 1) test design, 2) test validation, 3) test trial, and 4) test data analysis. The stages of test development are presented in Figure 1.

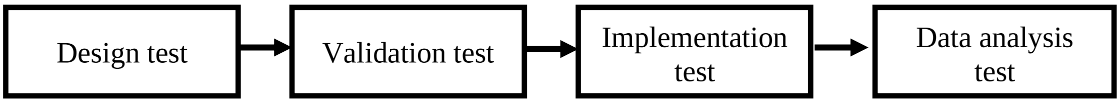


Figure 1. Instrument development steps

Participants

The research sample consisted of 130 students from two high schools in Madiun who worked online using Google Forms with Exambro security within 30 minutes. Sample selection was done using a purposive sampling technique to represent the target population. The sample selection criteria were determined based on recommendations from the physics teacher at the school concerned, with the aim that the selected students were studying relativity material.

Instruments

The instrument used in this research is the Relativity Concept Inventory Test, which consists of 13 multiple-choice questions. The questions are based on material from relativity, which includes Einstein's first and second postulates, time dilation, Velocity addition, and length contraction. Each question has a confidence level to measure students' understanding and self-confidence with scoring guidelines, as seen in Table 1. The instrument format was adapted from research (Aslanides & Savage, 2013). The instrument's validity was measured based on assessing 6 physics education lecturers as material experts and 1 physics teacher as a practitioner using formulas (Aiken, 1985). The instrument's reliability was tested through trial data analysis using the Generalized Partial Credit Model 3PL (GPCM-3PL).

Table 1. Guidelines for scoring the relativity concept inventory test instrument

Answer	Confidence options	Score	Information
Correct	Certain	3	Understanding
	Confident	2	Partial Understanding
	Unconfident	1	Partial Understanding
	Guessing	0	Not Understanding
Wrong	Certain		Misconception
	Confident		Misconception
	Unconfident		Not Understanding
	Guessing		Not Understanding

Data analysis

Data analysis was carried out using several statistical techniques with the help of software to ensure the validity and reliability of the instrument. The instrument's validity was analyzed using Excel to calculate the expert validity score based on Aiken's V formula (Aiken, 1985). Next, the trial result data was analyzed using QUEST software to determine Reliability and Instrument Item Suitability (Goodness of Fit). The reliability of the instrument is calculated to determine the consistency of the measurement results, using the categories stated (Guilford, 1956), as shown in Table 2

Table 2. Coefficient reliability criteria

Interval Coefficient (r)	Interpretation
>0.80	Very high
0.70 - 0.79	High
0.60 - 0.69	Moderate
0.50 - 0.59	Low
<0.50	Very low

The goodness of Fit is calculated to measure the suitability of the instrument items with the model used. The criteria are that an item is said to be fit if the INFIT MNSQ value is between 0.77 to 1.30 and uses INFIT t with a limit of -2.0 to 2.0 (Adams & Khoo, 1993). IRT GPCM 3PL analysis is carried out using Parscale software to determine the slope (distinguishing power), location (level of difficulty), and guessing (possibility of guessing) by referring to the value criteria proposed (Baker, 2021) as shown in Table 3. Finally, the Item Characteristic Curve (ICC) is used to describe the probability of a correct answer that changes with the student's ability

Table 3. Value criteria for slope, location and guessing analysis

Discrimination (Slope)		**Difficulty (Location)**		**Guessing**	
Interval Score	**Interpretation**	**Interval Score**	**Interpretation**	**Interval Score**	**Interpretation**
>0.70	Very High	< -2.0	Very Easy	0.00	None
1.35 – 1.70	High	-2.0 to -0.5	Easy	> 0.00 - 0.25	Low
0.65 – 1.34	Moderate	-0.5 to 0.5	Moderate	> 0.25 - 0.50	Moderate
0.00 – 0.64	Low	0.5 to 2.0	Difficult	> 0.50 - 0.75	High
		> 0.75 - 1.00	Very High		

▪ RESULT AND DISSCUSSION

Instrument Design

The design of the relativity concept inventory test instrument begins with determining the relativity material that will be used as the basis for preparing the instrument. Focus Group Discussion (FGD) with initial-level students, teachers, and lecturers agreed on several materials, including Einstein's first and second postulates, time dilation, velocity addition, and length of contraction. From these five materials, a matrix and grid were created, which will be used as a basis for writing the 13 questions. The relativity concept inventory test questions are adjusted to the cognitive taxonomy levels C2 to C5 with the distribution in Table 4.

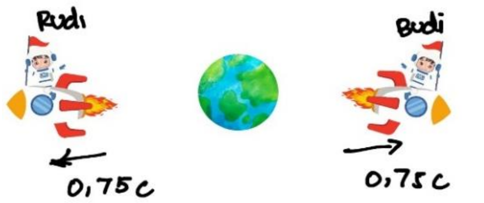
Table 4. Matrik of inventory test instruments for the concept of relativity

Aspect	**Concept**	**Indicators Concept**	**Question**
Implementing (C3)	First Postulate	The laws of physics are the same in all inertial reference frames.	1.2
	Time Dilation	The time interval between two separate events in a reference frame.	8

	Velocity addition	Velocities transform between frames such that no object can be observed traveling faster than the speed of light in a vacuum.	9.10
	Length contraction	The length of an object is the longest in the frame in which the ends of the object are at rest and is shorter in all other frames.	11
Analyzing (C4)	Second Postulate	The speed of light in a vacuum is the same in all reference frame.	4.5
	Length contraction	The length of an object is the longest in the frame in which the ends of the object are at rest and is shorter in all other frames.	12
Understanding (C2)	Second Postulate	The speed of light in a vacuum is the same in all reference frame.	6
	Length contraction	The length of an object is the longest in the frame in which the ends of the object are at rest and is shorter in all other frames.	13
Evaluate (C5)	First Postulate	The laws of physics are the same in all inertial reference frames.	3
	Time Dilation	The time interval between two separate events in a reference frame.	7

Based on the results of Matrik distribution in Table 5, the inventory test items for the concept of relativity were written in multiple choice form and equipped with levels of confidence from guessing level to definite level. For example, the relativity concept inventory question number 9 is shown in Figure 2. Multiple-choice questions equipped with a level of confidence have several advantages compared to conventional multiple-choice questions. The advantage is that there is an increase in measurement validity because confidence in the answers provides additional information about student understanding (Farrell & Leung, 2008). In addition, the level of self-confidence helps identify misconceptions because students who answer confidently incorrectly may have deep misconceptions. The level of self-confidence will also reduce the effect of guessing on test results, providing a more accurate picture of the student's abilities(Allen et al., 2006; Aslanides & Savage, 2013; Elisa et al., 2009; Goncher et al., 2015).

The following question in the scenario is: Rudi and his friend Budi decide to take separate trips on the same spaceship. They each accelerate away from Earth in opposite directions, namely rudi at v = 0.75c to the left and Budi at v = 0.75c to the right, relative to the observer on Earth.



If Rudi measures the rate of increase in distance to Budi, he will get a value, namely:
A. Equal to 1.5 c
B. Greater than 0.75c but less than c
C. Same as c

Rate how confident you are with your answer:

a	B	c	d
Guessing	Unconfident	Confident	Certain

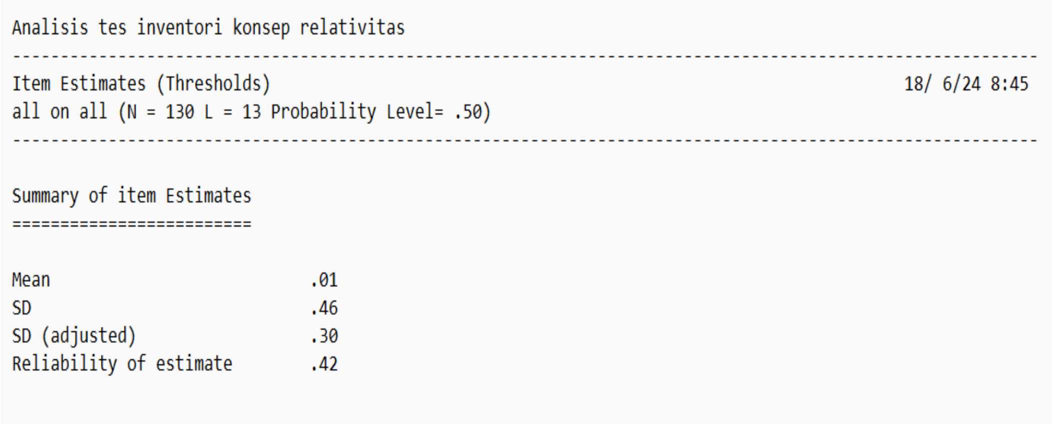
Figure 2. Question number 9 with material on velocity addition

Validity

The question items that have been successfully developed are validated through expert judgment by 7 experts consisting of 6 physics education lecturers and 1 teacher. The validation results show that a total of 13 questions received a V Aiken score of 0.93 with a valid category. The high validity value is because when preparing the instrument, input from experts who have a deep understanding of the material and measurement objectives is taken into account so that the instrument is valid for measuring students' understanding of the concept of relativity.

Reliability

Reliability aims to measure the consistency of the results obtained from the instrument. In this research, reliability was analyzed using QUEST software to produce a value of 0.42, which, according (Guilford, 1956)category, is very low, see Figure 3.



```
Analisis tes inventori konsep relativitas
--------------------------------------------------------------------------------------------------
Item Estimates (Thresholds)                                                          18/ 6/24 8:45
all on all (N = 130 L = 13 Probability Level= .50)
--------------------------------------------------------------------------------------------------

Summary of item Estimates
=========================

Mean                          .01
SD                            .46
SD (adjusted)                 .30
Reliability of estimate       .42
```

Figure 3. Output of the reliability test estimate for the inventory test of the concept of relativity

The very low reliability value indicates that the measurement results with this instrument are inconsistent. This condition occurred because it was influenced by several factors, including the limited number of participants, namely only 130 students. Small sample sizes can cause high variability in reliability estimates because small samples are less representative of the broader population (Dai et al., 2021). In addition, the timing of tests close to school summative assessments may affect student motivation and concentration. Students may experience fatigue or burnout, so they do not perform at their best during tests (Sievertsen et al., 2016). Other factors that may also have an influence are the absence of rewards or incentives for students and the low risk of the test, which decreases students' motivation to answer questions (Agnew et al., 2021). By considering these factors, it can be concluded that several aspects of the implementation and design of the instrument need to be improved to increase the reliability of the instrument in future research.

Instrument Testing

Instrument Item Suitability (Goodness Fit)

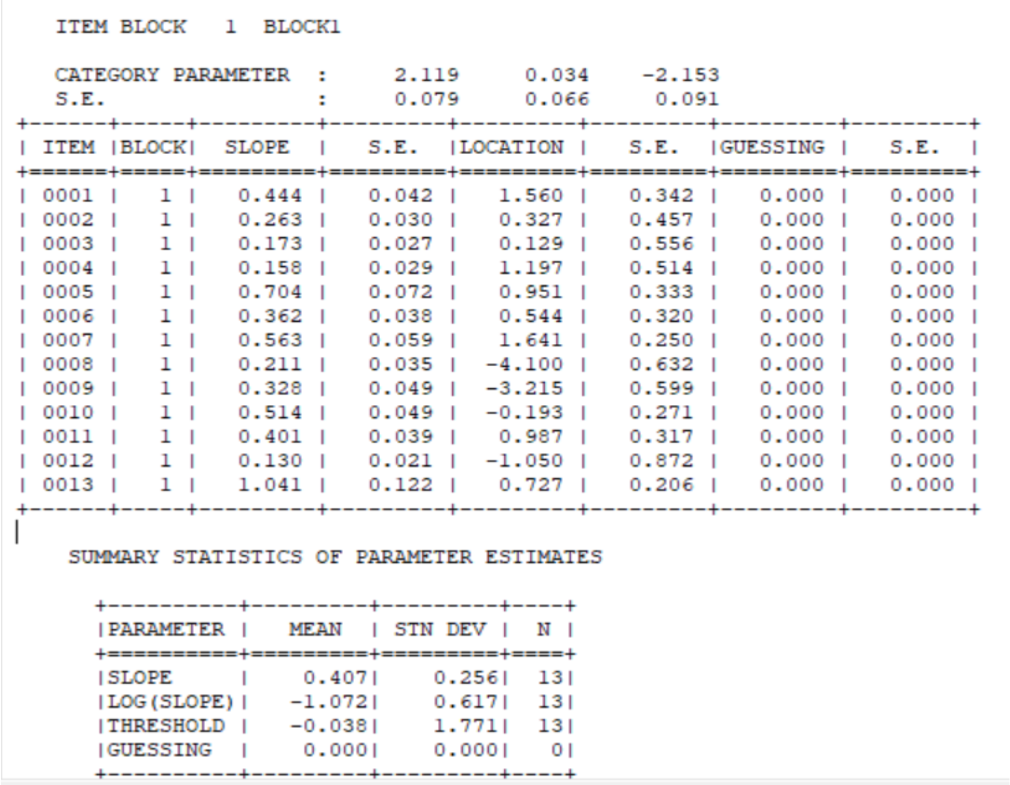
Testing to determine the goodness fit of each item follows the rules (Adams & Khoo, 1993). An item is said to be fit if the INFIT MNSQ value is between 0.77 to 1.30, and using INFIT t with a limit of -2.0 to 2.0, then suitable items are obtained that meet goodness of fit. The INFIT MNSQ inventory value of the concept of relativity from the results of analysis using QUEST software is between 0.93 to 1.07 and INFIT t -0.6 to 0.7 in Table 6. With the item acceptance limits using INFIT MNSQ and INFIT t, 13 items were declared fit All.

Table 5. INFIT MNSQ and INFIT t inventory values for the concept of relativity

Question	INFIT MNSQ	INFIT t	Criteria
1	1.00	.0	Fit
2	1.07	.9	Fit
3	1.01	.2	Fit
4	.96	-.5	Fit
5	1.00	.1	Fit
6	1.06	.7	Fit
7	.96	-.2	Fit
8	.96	-.2	Fit
9	.93	-.6	Fit
10	1.00	.1	Fit
11	1.03	.3	Fit
12	1.02	.3	Fit
13	.94	-.6	Fit

Analysis of Slope, Location and Guessing

In item response theory, IRT GPCM 3PL analysis uses three main parameters, namely parameter a (different power or slope), parameter b (level of difficulty or location), and parameter c (guessing). These parameters describe the characteristics of the items used to measure student abilities. The following are the results of the analysis using Parscale software, which can be seen in Figure 4



```
   ITEM BLOCK   1  BLOCK1

   CATEGORY PARAMETER  :      2.119     0.034    -2.153
   S.E.                :      0.079     0.066     0.091
+------+-----+---------+---------+---------+---------+---------+---------+
| ITEM |BLOCK|  SLOPE  |   S.E.  |LOCATION |   S.E.  |GUESSING |   S.E.  |
+======+=====+=========+=========+=========+=========+=========+=========+
| 0001 |   1 |   0.444 |   0.042 |   1.560 |   0.342 |   0.000 |   0.000 |
| 0002 |   1 |   0.263 |   0.030 |   0.327 |   0.457 |   0.000 |   0.000 |
| 0003 |   1 |   0.173 |   0.027 |   0.129 |   0.556 |   0.000 |   0.000 |
| 0004 |   1 |   0.158 |   0.029 |   1.197 |   0.514 |   0.000 |   0.000 |
| 0005 |   1 |   0.704 |   0.072 |   0.951 |   0.333 |   0.000 |   0.000 |
| 0006 |   1 |   0.362 |   0.038 |   0.544 |   0.320 |   0.000 |   0.000 |
| 0007 |   1 |   0.563 |   0.059 |   1.641 |   0.250 |   0.000 |   0.000 |
| 0008 |   1 |   0.211 |   0.035 |  -4.100 |   0.632 |   0.000 |   0.000 |
| 0009 |   1 |   0.328 |   0.049 |  -3.215 |   0.599 |   0.000 |   0.000 |
| 0010 |   1 |   0.514 |   0.049 |  -0.193 |   0.271 |   0.000 |   0.000 |
| 0011 |   1 |   0.401 |   0.039 |   0.987 |   0.317 |   0.000 |   0.000 |
| 0012 |   1 |   0.130 |   0.021 |  -1.050 |   0.872 |   0.000 |   0.000 |
| 0013 |   1 |   1.041 |   0.122 |   0.727 |   0.206 |   0.000 |   0.000 |
+------+-----+---------+---------+---------+---------+---------+---------+

    SUMMARY STATISTICS OF PARAMETER ESTIMATES

      +----------+---------+---------+----+
      |PARAMETER |   MEAN  | STN DEV |  N |
      +==========+=========+=========+====+
      |SLOPE     |    0.407|    0.256|  13|
      |LOG(SLOPE)|   -1.072|    0.617|  13|
      |THRESHOLD |   -0.038|    1.771|  13|
      |GUESSING  |    0.000|    0.000|   0|
      +----------+---------+---------+----+
```

Figure 4. Result analysis Slope (a), Location (b), and Guessing (c)

The results of this analysis showed that the average difference (slope) value obtained was 0.407, which was low and below the ideal value of> 0.7, as suggested by (Baker, 2021). Low discriminating indicates that the questions are not effective enough in differentiating between students with high and low abilities. Research by (Singh Rana, 2014) shows that low discrimination is often caused by a lack of variation in the level of difficulty of the questions. This result is in line with the difficulty level score (location) of the questions, which are in the difficulty range (-1,050 to +1,641); namely the majority of questions are in the easy and medium range. This is also reinforced by two questions that are outside the range (-2 to +2), namely item number 8 (b = -4,100) and number 9 (b = -3,215), which means that these two questions are invalid because they are too easy. A guessing value of 0 (zero) indicates that there is no possibility that students can answer correctly just by guessing, which is a positive indicator for the validity of the questions because good questions should not be easy to guess (Baker, 2021). Based on these results, the developed Relativity Concept Inventory Test instrument has advantages compared to previous research; namely, the guessing value shows that all the questions are of very good quality because there is no chance for students to answer correctly just by guessing. This advantage is significant compared to several other instruments, such as the Relativity Concept Inventory (RCI) by (Aslanides & Savage, 2013) which does not completely eliminate guessing. Second, the use of two different software, namely QUEST for reliability and goodness of fit analysis, Parscale for analysis of Slope, Location and Guessing, allows a more holistic and comprehensive approach. However, this research also still needs to improve, namely the limited number of trial respondents. The limited number of respondents can affect the accuracy and generalization of the analysis results.

Item Characteristic Curve (ICC)

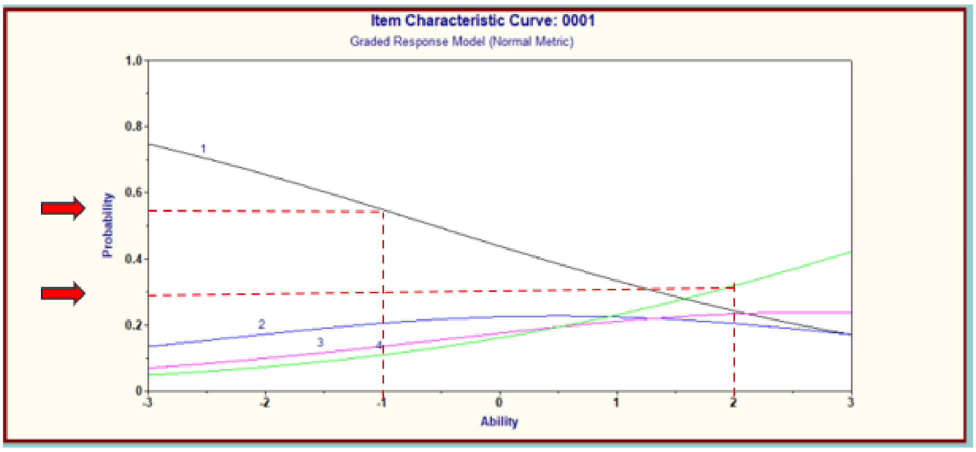
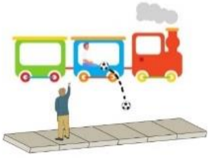
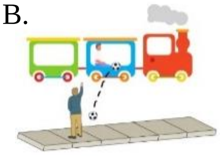
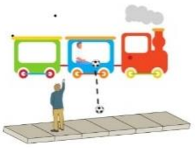


Figure 5. Item Characteristic Curve (ICC) for question number 1

Item Characteristic Curve (ICC) is used as a parameter in IRT analysis to evaluate item performance in the test (Adedoyin & Mokobi, 2013; McGrath, 2019; Pasquali & Primi, 2003). The results of the analysis obtained an ICC of 13, which is important for increasing the validity and reliability of the test. For example, shown in Figure 5 are the ICC results for question number 1. The curve provides information that in category 1, with a probability value of 0.7, most of the respondents had an ability of -1 and in category 4, with a probability value of 0.3, most of the respondents had an ability of 2. These results have provided information that item number 1 has been able to identify abilities sequentially. The form of question number 1 can be seen in Figure 6.

Rudi sits on a train that moves at a speed v from left to right relative to Budi, standing on the platform. As Rudi passes Budi, he drops the ball out of the train window.
Note: Air resistance is ignored*
Which picture will Budi observe while standing on the platform?

A. 	B. 	C. 

Rate how confident you are with your answer:

a	b	c	d
Guessing	Unconfident	Confident	Certain

Figure 6. Example question number 1

The ICC curve shows that item number 1 can identify students' abilities sequentially. Students with lower abilities answer with lower categories, while students with higher abilities tend to choose higher categories. This shows that this item has good distinguishing power. Even though the average slope value for all items in the test is 0.407, this item is still effective in differentiating students with various levels of ability. Overall, the ICC analysis shows that item number 1 is able to identify students with various levels of ability. The different probabilities of correct answers for each category indicate that these items effectively differentiate students based on their abilities. Although there is room for improvement in terms of discrimination, these results indicate that the items are valid and reliable in measuring students' relativity concept abilities. The complete ICC graph for questions 1 to 13 can be seen in Figure 7.

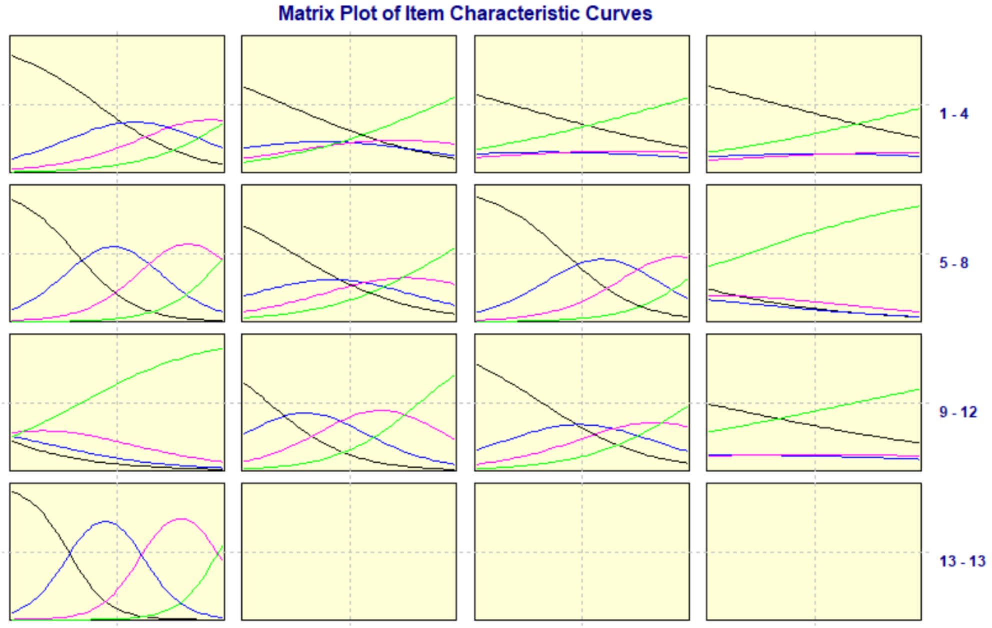


Figure 7. Item Characteristic Curve (ICC) questions number 1 to 13

Standard Error of Measurement (SEM)

The Test Information Function curve and Standard Error of the test instrument analyzed using the 3PL Generalized Partial Credit Model (GPCM) provide important information about the quality and reliability of items in the test at various levels of student ability.

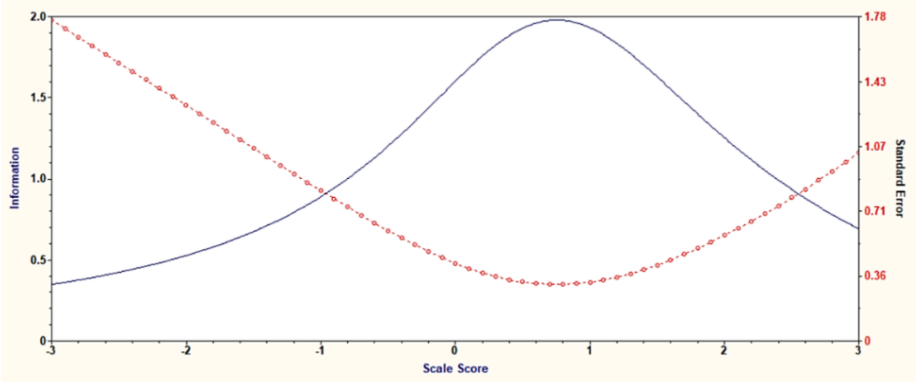


Figure 8. Information function curves and SEM

Based on Figure 8 above, it can be seen that the intersection of the Test Information Function curve and Standard Error is on a capability scale of -1.0 to +2.3. These results indicate that this test instrument is most suitable for students who have abilities in this range. The Test Information function (blue curve) shows that the instrument is most informative over this capability range, providing the most accurate and reliable information. In contrast, information decreases at both ends of the ability spectrum (very low and very high ability), indicating that the test is less informative for students with very low (< -1.0) or very high (> 2.3) ability. The Standard Error curve (dotted red line) shows the lowest standard error of measurement at a capability of 0.7, reaching a low point of around 0.36. This means that student ability estimates are accurate in the -1.0 to +2.3 range.

▪ CONCLUSION

The results of the analysis can be concluded that we have succeeded in developing a test instrument for an inventory of the concept of relativity in the form of multiple choices with a confidence level of 13 questions. All instruments have met validity through expert judgment with a validity value of 0.93 in the valid category. The reliability of the instrument is in the very low category, with a reliability coefficient of 0.42. Empirical testing with the 3PL Generalized Partial Credit Model (GPCM) stated that 11 questions were valid out of the 13 questions developed. In addition, this instrument is most suitable for students who have abilities in the -1.0 to +2.3 range or in the middle ability range. These results show that the instrument developed has gone through a holistic and comprehensive validation process so that it can be used to measure students' understanding of the concept of relativity.

▪ REFERENCES

Adams, R. J., & Khoo, S. T. (1993). Quest: the interactive test analysis system. australian council for educational research.

Adedoyin, O. O., & Mokobi, T. (2013). Using irt psychometric analysis in examining the quality of junior certificate mathematics multiple. International Journal of Asian Social Science, 3(4), 992–1011.

Agnew, S., Kerr, J., & Watt, R. (2021). The effect on student behaviour and achievement of removing incentives to complete online formative assessments. Australasian Journal of Educational Technology, 2021(4), 173–185.

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psychological Measurement, 45(1), 131–142.

Allen, K., Reed, T., & Terry, R. (2006). Work in progress: assessing student confidence of introductory statistics concepts. Proceedings - Frontiers in Education Conference, FIE, 13–14.

Aslanides, J., & Savage, C. (2013). Relativity concept inventory: Development, analysis, and results. Physical Review Special Topics-Physics Education Research, 9(1), 10118–10128.

Baker, F. (2021). The basics of item response theory. (C. Boston & L. Rudner, Eds.). ERIC Clearinghouse on Assessment and Evaluation.

Cormier, S., & Steinberg, R. (2010). The twin twin paradox: exploring student approaches to understanding relativistic concepts. The Physics Teacher, 48(9), 598–601.

D Mardapi. (2008). *Teknik penyusunan instrumen tes dan nontes* [Techniques for preparing test and non-test instruments]. Mitra Cendikia Press.

Dai, S., Vo, T., Kehinde, O., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of polytomous irt models with rating scale data: an investigation over sample size, instrument length, and missing data. Frontiers in Education, 6, 1–18.

Elisa, A., Sotos, C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? Journal of Statistics Education, 17(2), 1–21.

Embretson, S., Yang, X., Green, J. L., Camilli, G., & Elmore, P. B. (2006). Item response theory. In Handbook of complementary methods in education research.

Ene, E., & Ackerson, B. J. (2018). Assessing learning in small sized physics courses. Physical Review Physics Education Research, 14(1).

Farmer, S. (2021). Introduction of einsteinian physics to the upper secondary school physics curriculum in scotland. In Teaching Einsteinian Physics in Schools (1st ed.). Routledge.

Farrell, G., & Leung, Y. (2008). Convergence of validity for the results of a summative assessment with confidence measurement and traditional assessment.

Gero, A., Tsybulsky, D., & Levin, I. (2019). Research and design triads in the digital epoch: Implications for science and technology education. Global Journal of Engineering Education, 21(1), 80–83.

Goncher, A., Boles, W., & Jayalath, D. (2015). Using textual analysis with concept inventories to identify root causes of misconceptions. Proceedings - Frontiers in Education Conference, FIE, 2015, 1–4.

Guilford, J. P. (1956). Fundamental statistics in psychology and education, 3rd ed. In Fundamental statistics in psychology and education, 3rd ed. McGraw-Hill.

Hartle, J. B. (2005). General relativity in the undergraduate physics curriculum. American Journal of Physics, 74(1), 14–21.

Hazari, Z., Tai, R. H., & Sadler, P. M. (2007). Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors. Science Education, 91(6), 847–876.

Istyono, E. (2020). *Pengembangan instrumen penilaia dan analisis belajar fisika dengan teori klasik dan modern* [development of instruments for assessment and analysis of physics learning with classical and modern theories]. UNY Press.

Kulgemeyer, C., & Wittwer, J. (2021). When learners prefer the wrong explanation: misconceptions in physics explainer videos and the illusion of understanding. 1-29.

Listianingrum, S. A., Jumadi, J., & Zakwandi, R. (2022). Physics student misconception: relative velocity, time dilatation, and length contraction. Jurnal Ilmiah Pendidikan Fisika, 6(2), 1–7.

McGrath, K. (2019). Investigating the impact of parameter instability on item response theory proficiency estimation. Proceedings of the 2019 AERA Annual Meeting.

Muraki, E. (1992). A Generalized partial credit model: application of an em algorithm. Applied Psychological Measurement, 16(2), 159–176.

Pasquali, L., & Primi, R. (2003). Fundamentos da teoria da resposta ao item: TRI. Avaliaçao Psicologica: Interamerican Journal of Psychological Assessment, 2(2), 99–110.

Piacsek, A. A. (2018). A new pre/post test to assess student mastery of introductory level acoustics and wave mechanics. The Journal of the Acoustical Society of America, 144(3), 1785–1786.

Sachan, A., Bhadri, G. N., & Kittur, J. (2019). Design and development of concept assessment tool (cat):a concept inventory. Journal of Electrical Engineering & Technology, 33(1), 16–21. https://api.semanticscholar.org/CorpusID:214242537

Samejima, F. (1997). Graded Response Model. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of Modern Item Response Theory (pp. 85–100). Springer New York.

Scherr, R. E., Shaffer, P. S., & Vokos, S. (2001). Student understanding of time in special relativity: Simultaneity and reference frames. American Journal of Physics, 69(S1), 24–35.

Sievertsen, H., Gino, F., & Piovesan, M. (2016). Cognitive fatigue influences students' performance on standardized tests. Proceedings of the National Academy of Sciences of the United States of America, 113.

Singh Rana, S. (2014). Test item analysis and relationship between difficulty level and discrimination index of test items in an achievement test in biology. Indian Journal Of Research, 3(6), 56–58.

Siong, L. chin, Tyug, O. Y., Phang, F. A., & Pusppanathan, J. (2023). The use of concept cartoons in overcoming the misconception in electricity concepts. Participatory Educational Research, 10(1), 310–329.

Thacker, B., Kim, E., Trefz, K., & Lea, S. M. (1994). Comparing problem solving performance of physics students in inquiry-based and traditional introductory physics courses. American Journal of Physics, 62(7), 627–633.

Vicovaro, M. (2023). Grounding intuitive physics in perceptual experience. Journal of Intelligence, 11(10), 1–20.