PILLAR OF PHYSICS EDUCATION
Vol. 19 (4), 2026, page.346-357
ISSN (Print) : 2337-9618
ISSN (Online) : 2684-8880

ANALYSIS OF THE VALIDITY AND RELIABILITY OF A NEWTON'S LAWS CONCEPT-UNDERSTANDING TEST INSTRUMENT: A CLASSICAL TEST THEORY AND RASCH MODELING APPROACH

Hilda Mazlina[1]*, Mainisa[1], Ria Rahayu[2]

[1] *Department of Physics Education, Universitas Islam Negeri Ar-Raniry, Jl. Syeikh Abdul Rauf Kopelma Darussalam, Kota Banda Aceh, 23111, Indonesia*
[2]*Department of Sains Education, Universitas Samudera, Jl. Prof. Dr. Syarief Thayeb, Meurandeh, Langsa, 24416, Indonesia*
Corresponding author. Email:*hilda.mazlina@ar-raniry.ac.id*

ABSTRACT

This study aims to evaluate the psychometric quality of a Newton's Laws concept-understanding test by integrating Classical Test Theory (CTT) with the Rasch Measurement Model. A quantitative descriptive-evaluative design was applied to N = 100 tenth-grade students from two senior high schools (SMA) in Aceh, Indonesia. The instrument comprised 20 single-tier multiple-choice items covering Newton's First, Second, and Third Laws. CTT analysis (Pearson item–total correlation and Cronbach's Alpha) was performed in SPSS, and Rasch analysis (item fit, reliability, and Wright Map) was performed in WINSTEPS. Validity and reliability: 16 of the 20 items were valid under CTT (rxy > 0.195) and Cronbach's Alpha was 0.78 (acceptable). The Rasch analysis supported unidimensionality (explained variance > 40%), with a Person Reliability of 0.75 and an Item Reliability of 0.92. Four items (Q3, Q6, Q14, and Q17) showed misfit (MNSQ > 1.50) and were also invalid under CTT. Interpretation of misconceptions: the response patterns on these four items suggested possible impetus-based reasoning about inertia, neglect of friction and angled-force components, and the assumption that the normal force equals weight in an accelerating elevator; these interpretations are tentative because the test is single-tier. Together, the two approaches provide complementary evidence on instrument quality and on students' conceptual difficulties in mechanics.

Keywords : Classical Test Theory, Concept Understanding, Newton's Laws, Physics Education Research,Rasch Modeling, Reliability,Validity.

Pillar of Physics Education is licensed under a Creative Commons Attribution ShareAlike 4.0 International License.

I. INTRODUCTION

The measurement of students' conceptual understanding of fundamental mechanics, particularly Newton's Laws of Motion, constitutes one of the crucial domains within Physics Education Research (PER) [1][10]. Understanding of Newton's Laws serves as a foundational prerequisite for students to engage with more advanced physics topics such as rotational dynamics, fluid mechanics, and the mechanics of rigid bodies [29]. Nevertheless, numerous international studies report that Newton's Laws remain among the topics most frequently associated with persistent alternative frameworks, or enduring misconceptions [5][26]. Students often bring intuitive, or naive, physical reasoning that contradicts Newtonian principles—for instance, the belief that a force acting in the direction of motion is continuously required to sustain that motion, or that objects of greater mass invariably exert a greater force during an interaction [8][22].

The accuracy with which students' conceptual understanding and misconceptions can be identified depends heavily on the psychometric quality of the test instrument employed [25]. Evaluation instruments whose validity and reliability have not been established are liable to generate biased and inconsistent measurement data, which in turn can lead to erroneous pedagogical interpretations [6]. In classroom evaluation practice, item-quality analysis has thus far been dominated by Classical Test Theory (CTT) implemented through SPSS software. CTT analysis focuses on computing the Pearson item–total correlation (rxy) to test item validity and Cronbach's Alpha coefficient (α) to measure internal consistency.

Although CTT analysis provides practical information that is useful at the classroom level, this approach carries fundamental methodological limitations [3]. CTT parameters such as item difficulty (p) and discrimination index (D) are sample-dependent—that is, strongly contingent on the characteristics of the student sample tested—whereas ability estimates are test-dependent, being strongly contingent on the difficulty of the specific test administered. These limitations render CTT validation results difficult to generalize beyond the sample population concerned. Furthermore, single-tier multiple-choice tests analyzed using CTT frequently fail to distinguish genuine conceptual understanding from lucky guessing and from structured misconceptions [12]. To address limitations, modern psychometric studies in PER recommend integrating CTT with the Rasch Measurement Model (a one-parameter logistic IRT model) [18][21]. Rasch modeling transforms dichotomous raw scores into a linear interval logit scale possessing the property of parameter invariance, whereby item difficulty estimates are independent of the student sample, and estimates of students' latent ability (βn) are independent of the particular set of items administered [30][27]. Moreover, Rasch modeling provides item-fit indicators (Infit/Outfit Mean Square) capable of detecting problematic items or anomalous response patterns, as well as a Wright Map that plots the distribution of student ability alongside the hierarchy of item difficulty [14].

To address this research gap, modern psychometric studies in PER recommend integrating CTT with the Rasch Measurement Model (a one-parameter logistic IRT model) [18][21]. Rasch modeling transforms dichotomous raw scores into a linear interval logit scale possessing the property of parameter invariance, whereby item difficulty estimates are independent of the student sample, and estimates of students' latent ability (βn) are independent of the particular set of items administered [30]. Moreover, Rasch modeling provides item-fit indicators (Infit/Outfit Mean Square) capable of detecting problematic items or anomalous response patterns, as well as a Wright Map that plots the distribution of student ability alongside the hierarchy of item difficulty [14].

Previous studies have employed a range of assessment formats and psychometric approaches to examine students' understanding and misconceptions related to Newton's Laws. Established instruments such as the Force and Motion Conceptual Evaluation (FMCE) use research-based multiple-choice items to assess conceptual understanding of Newtonian mechanics [26]. More recent studies have increasingly adopted multi-tier diagnostic formats to distinguish students' conceptual difficulties from other sources of incorrect responses. For example, Kaniawati et al.[13] used a Four-Tier Newtonian Test (FTNT) with 30 high-school students and reported that the four-tier format provided more specific information about misconceptions than a two-tier test. Fratiwi et al., [7] also developed a four-tier version of a Newton's Laws diagnostic test from an earlier two-tier format using a 4D development model. Other studies have incorporated modern psychometric approaches. Samsudin et al., [22], for example, developed a multi-representation tier instrument for Newton's Laws and evaluated it using Rasch analysis with 92 high-school students. Mufit and Syamsidar [16] developed a four-tier Newton's Laws instrument and evaluated its validity and reliability using Aiken's V, point-biserial correlation, item difficulty, and KR-20 with a field-test sample of 21 students. More recently, Riau et al., [20] developed a two-tier Newton's Laws diagnostic instrument and applied Rasch analysis to data from 46 high-school students.

These studies demonstrate that Newton's Laws assessment research has progressed from conventional multiple-choice instruments toward multi-tier diagnostic formats and Rasch-based psychometric analysis. However, the existing studies have generally emphasized either the diagnosis of misconceptions through multi-tier formats or the evaluation of instrument properties using a single psychometric framework. The present study addresses a complementary need by combining Classical Test Theory (CTT) and the Rasch Measurement Model within the evaluation of a single-tier Newton's Laws concept-understanding test. CTT provides conventional evidence of item validity and internal consistency, whereas Rasch analysis provides information on item fit, item and person reliability, and the relative distribution of student ability and item difficulty. In addition, this study uses item-level misfit as a starting point for a cautious cognitive interpretation of students' responses. Thus, rather than claiming that previous studies have not combined these approaches, the contribution of the present study lies in integrating complementary CTT and Rasch evidence and linking statistical item behavior to the interpretation of possible conceptual difficulties in Newton's Laws among Indonesian senior high-school students.

Combining CTT and Rasch is necessary for this instrument for three reasons. First, it is a 20-item single-tier multiple-choice test in which a correct answer may reflect guessing rather than understanding. The item–total correlation in CTT shows which items discriminate weakly, but it cannot show where along the ability range the responses depart from expectation; Rasch Infit/Outfit statistics and the Wright Map provide this information. Second, the sample is of moderate size and was drawn from two schools, so CTT difficulty and discrimination indices are strongly sample-dependent, whereas the Rasch model places students and items on a common interval scale and allows item difficulty to be examined against student ability. Third, the two approaches can reveal different problems: an item may satisfy the CTT validity criterion yet approach the Rasch misfit threshold, as occurred for Q8 and Q18 in this study (Section 3.3). Using CTT alone would therefore overlook items with anomalous response patterns, while using Rasch alone would omit the conventional validity and reliability evidence still expected of classroom instruments.

Accordingly, this study aims to (1) examine item validity and internal-consistency reliability using CTT; (2) examine item fit, person and item reliability, and person–item targeting using the Rasch model; and (3) interpret the misfitting items cognitively to identify students' misconceptions about Newton's Laws. The novelty of the study lies in linking these psychometric results to an item-level cognitive diagnosis for a Newton's Laws test administered to N = 100 tenth-grade students from two schools in Aceh.

II. METHOD

2.1 Research Design and Subjects

This study employed a quantitative, descriptive-evaluative design aimed at analyzing and evaluating the psychometric quality of a Newton's Laws concept-understanding test instrument. The research subjects comprised 100 eleventh-grade students from two public senior high schools in Aceh, namely SMAN 3 Banda Aceh and SMAN 1 Darussalam. Fifty students were selected from SMAN 3 Banda Aceh and 50 students from SMAN 1 Darussalam. The subjects were selected using a multistage sampling technique.

In the first stage, the two schools were purposively selected based on the following criteria: (1) school accreditation status; (2) curriculum implemented; and (3) accessibility or other relevant criterion. In the second stage, students were selected from the eleventh-grade classes in each school. Student selection was based on the following criteria: (1) students had completed the Newton's Laws unit in the physics curriculum; and (2) the selected students represented heterogeneous academic achievement, as determined from the most recent physics report-card scores / another documented academic achievement measure. These criteria were applied to ensure that the sample included students with varying levels of academic achievement relevant to the assessment of the test instrument.

2.2 Test Instrument

The instrument under evaluation consisted of 20 single-tier multiple-choice items (Q1 through Q20), each with five answer options (A, B, C, D, E). The instrument was designed to measure three principal domains of Newton's Laws:

1) ***Newton's First Law (Inertia):*** Q1, Q2, Q3, Q9, Q13, Q16, Q20 (7 items). Indicators: frames of reference, states of rest, uniform rectilinear motion with zero net external force, and the concept of inertia within moving systems.

2) ***Newton's Second Law (Dynamics of Motion):*** Q4, Q5, Q6, Q10, Q11, Q14, Q17, Q19 (8 items). Indicators: the linear relationship Fnet = ma, F–a graph analysis, the effect of friction force, force components at an angle θ, and effective force in an accelerating (elevator) reference frame.

3) ***Newton's Third Law (Action–Reaction):*** Q7, Q8, Q12, Q15, Q18 (5 items). Indicators: interaction force pairs, the property of equal magnitude and opposite direction acting on two separate bodies, and the interaction of bodies with differing mass.

2.2.1 Item Development and Test Blueprint

The 20 items were newly developed by the researchers based on a test blueprint (specification grid) derived from the conceptual understanding indicators of Newton's First, Second, and Third Laws. The items were not directly adapted from existing instruments such as the Force Concept Inventory (FCI) or the Force and Motion Conceptual Evaluation (FMCE). However, the development of the items and distractors was informed by previous research on students' alternative conceptions related to Newton's Laws [15][26][23].

The blueprint mapped each item to one of the three Newton's Laws domains and to the corresponding indicators, consisting of seven items for Newton's First Law, eight items for Newton's Second Law, and five items for Newton's Third Law. All items were constructed as single-tier multiple-choice questions with one correct answer and four distractors. The distractors were designed to represent plausible alternative conceptions reported in the literature. The items incorporated conceptual reasoning, graphical representation, and numerical calculations where appropriate to assess students' understanding of Newton's Laws from different perspectives.

2.2.2. Content Validation

Before empirical testing, the draft instrument was reviewed by experts in physics education and physics instruction. The experts evaluated the relevance of each item to the specified conceptual understanding indicators, the accuracy of the physics content, the clarity of the wording, and the plausibility of the distractors. Based on the experts' feedback, several items were revised to improve the accuracy of the content, clarity of the questions, and appropriateness of the answer options and distractors. The revised instrument was then used for empirical testing.

2.3 Data Collection Procedure

Data were collected through a one-shot administration of the concept-understanding test, with a duration of 60 minutes. Student answer sheets were scored dichotomously, with correct responses assigned a score of 1 and incorrect responses a score of 0.

2.4 Data Analysis Technique

2.4.1 Classical Test Theory (CTT)

CTT analysis was performed using SPSS software. Item validity was tested using the Pearson Product-Moment correlation formula between the item score (X) and the total score (Y):

$$rxy = [N\cdot\Sigma XY - (\Sigma X)(\Sigma Y)] / \sqrt{\{[N\cdot\Sigma X^2 - (\Sigma X)^2] \cdot [N\cdot\Sigma Y^2 - (\Sigma Y)^2]\}}$$

An item was deemed valid if its coefficient rxy > rtable. For a sample of N = 100 at a significance level of $\alpha = 0.05$, the critical value is $r_{table} = 0.195$.

Internal-consistency reliability of the instrument was calculated using Cronbach's Alpha (α):

$$\alpha = [K/(K-1)] \cdot [1 - \Sigma\sigma i^2/\sigma t^2]$$

where K is the number of items (K = 20), σi^2 is the variance of item i, and σt^2 is the total test-score variance. Following Nunnally and Bernstein (1994), $\alpha \geq 0.70$ was interpreted as acceptable and $\alpha \geq 0.80$ as good internal consistency; because α is also influenced by test length and dimensionality, it was interpreted together with the Rasch evidence [25][12].

2.4.2 Rasch Modeling (1-PL IRT)

Latent psychometric analysis was conducted using WINSTEPS software based on the one-parameter logistic Rasch model. The probability Pni that student n, with ability βn, answers item i, with difficulty δi, correctly is given by:

$$P_{ni}(X_{ni} = 1 \mid \beta_n, \delta_i) = \exp(\beta_n - \delta_i) / [\, 1 + \exp(\beta_n - \delta_i) \,]$$

Item fit was evaluated on the basis of the Infit and Outfit Mean Square (MNSQ) statistics, with an ideal acceptance range of $0.50 \leq \text{MNSQ} \leq 1.50$ and a Z-Standardized (ZSTD) value of $-2.0 \leq \text{ZSTD} \leq +2.0$. [14]. Unidimensionality was examined using principal component analysis of residuals, with raw variance explained by the measures of at least 40% regarded as good [17][19][9]. Reliability was reported separately as Person Reliability and Item Reliability, and interpreted using the categories : < 0.67 weak, 0.67–0.80 acceptable, 0.81–0.90 good, 0.91–0.94

very good, and > 0.94 excellent. A Wright Map was constructed to depict the distribution of students' latent ability relative to item difficulty (Figure 1). Targeting was judged from the difference between the person and item means and from floor and ceiling effects, defined as more than 15% of respondents located beyond the easiest or hardest item.

III. RESULTS AND DISCUSSION

3.1 Results of the Combined Psychometric Analysis (CTT & Rasch)

Testing of N = 100 students from two schools yielded stable parameter estimates. Table 1 presents a synthesis of the CTT and Rasch modeling results for the 20 items of the Newton's Laws concept-understanding test.

Table 1. Synthesis of CTT and Rasch Model Psychometric Evaluation (N = 100)

Item Code	Sub-topic / Concept Indicator	p (CTT)	rxy (Pearson)	CTT Status (rtable = 0.195)	Rasch Measure (Logit)	Infit MNSQ	Outfit MNSQ	Psychometric Quality Category
Q1	Law I: Condition of inertia & frame of reference	0.78	0.521	Valid	−1.25	0.95	0.92	Very Good / Easy
Q2	Law I: Object moving with no net external force	0.65	0.482	Valid	−0.58	0.98	0.96	Very Good / Moderate
Q3	Law I: Conceptual force on a sliding object	0.32	0.182	Invalid	+1.35	1.58	1.65	Misfit / High Misconception
Q4	Law II: Linear relationship between Fnet and a	0.70	0.545	Valid	−0.82	0.91	0.89	Very Good / Easy
Q5	Law II: F–a graph analysis	0.45	0.420	Valid	+0.28	1.02	1.05	Good / Fairly Difficult
Q6	Law II: Effect of friction force on acceleration	0.28	0.150	Invalid	+1.72	1.62	1.70	Misfit / Too Difficult
Q7	Law III: Force pair on stacked objects	0.58	0.495	Valid	−0.15	0.94	0.93	Very Good / Moderate
Q8	Law III: Interaction force between two objects of different mass	0.35	0.215	Valid	+1.12	1.38	1.42	Acceptable Range / Effective Distractor
Q9	Law I: Object on a frictionless inclined plane	0.52	0.410	Valid	+0.08	1.05	1.08	Good / Moderate

Item Code	Sub-topic / Concept Indicator	p (CTT)	rxy (Pearson)	CTT Status (rtable = 0.195)	Rasch Measure (Logit)	Infit MNSQ	Outfit MNSQ	Psychometric Quality Category
Q10	Law II: Two-mass system connected by a string	0.40	0.385	Valid	+0.52	1.10	1.12	Good / Fairly Difficult
Q11	Law II: Mass change under a constant force	0.62	0.460	Valid	−0.42	0.96	0.94	Good / Moderate
Q12	Law III: Gravitational force and normal force	0.38	0.280	Valid	+0.85	1.25	1.28	Good / Fairly Difficult
Q13	Law I: Elevator moving at constant velocity	0.55	0.430	Valid	−0.02	0.99	0.97	Good / Moderate
Q14	Law II: Elevator accelerating upward (Feffective)	0.30	0.175	Invalid	+1.60	1.52	1.55	Misfit / Difficult
Q15	Law III: Action–reaction on a launching rocket	0.60	0.475	Valid	−0.31	0.93	0.91	Good / Moderate
Q16	Law I: A stopping bus and passenger inertia	0.82	0.560	Valid	−1.55	0.88	0.85	Very Good / Very Easy
Q17	Law II: Object pulled at an angle θ	0.34	0.160	Invalid	+1.48	1.56	1.60	Misfit / Complex
Q18	Law III: Collision between a truck and a small car	0.36	0.240	Valid	+1.05	1.32	1.35	Accepted / Pure Misconception
Q19	Law II: Deceleration due to braking force	0.50	0.405	Valid	+0.18	1.01	1.03	Good / Moderate
Q20	Law I: Uniform circular motion and net force	0.42	0.365	Valid	+0.45	1.08	1.10	Good / Moderate

Note: Shaded rows in Table 1 mark the four misfitting items (Q3, Q6, Q14, Q17), which are discussed further in Section 3.3. Items were classified as valid when rxy > 0.195 and as misfitting when Infit MNSQ exceeded 1.50 (acceptable range 0.50–1.50).

Why items classified as invalid are still discussed in the Rasch analysis. In CTT, an item is called invalid only when its correlation with the total score falls below the critical value, that is, when it discriminates poorly between students who score high and low on the whole test. This is a screening decision about the item's contribution to the total score, not a judgement that the item carries no information. The four flagged items were therefore retained in the Rasch analysis. Rasch fit statistics address a different question, namely whether the observed responses match the pattern expected from each student's ability and each item's difficulty, and removing the items beforehand

would have discarded exactly the responses that show why these items behave poorly. CTT was thus used to flag items and Rasch to locate and characterize the departure from expectation.

Relationship between CTT validity and Rasch fit. The two sets of evidence agree closely. All four items that were invalid under CTT also exceeded the Infit MNSQ upper limit of 1.50, and all 16 valid items fell within the acceptable range (Table 1). A low item–total correlation and Rasch misfit appear to be two views of the same problem: responses to these items are less predictable from overall ability than the model expects, for example when able students choose an incorrect option. These four items were also the most difficult items of the test and had the lowest proportions correct, so the problem is concentrated at the upper end of the difficulty scale. Two Third Law items (Q8 and Q18) passed the CTT criterion yet had Infit values close to the limit, which shows why both criteria were examined and not only one.

Educational meaning. Even after the four misfitting items are set aside, the remaining 16 items still cover all three Newton's Laws (six First Law, five Second Law, and five Third Law items), so the shortened test can serve as a classroom-level diagnostic of conceptual understanding. The easiest items (e.g., Q1 and Q16) concern the basic idea of inertia in familiar situations, whereas the items that required applying Newton's Second Law in less familiar contexts (friction, angled forces, and accelerating motion) were the most difficult and fitted the model worst. This suggests that instruction secures the qualitative statement of the laws more readily than their application, and it points to free-body-diagram construction, vector resolution, and accelerating situations as topics that deserve more teaching time.

3.2 Wright Map (Person–Item Map)

Figure 1 presents the Wright Map generated directly from WINSTEPS, showing the distribution of the 100 students' latent ability (βn) and the difficulty of the 20 test items (δi) on the same logit scale. The person distribution is displayed on the left side, while the item distribution is displayed on the right side. This map provides a visual representation of the relationship between students' ability levels and the difficulty hierarchy of the test items.

Figure 1 shows that item difficulties span almost the same range as student abilities (students: −1.70 to +1.70 logits; items: −1.55 to +1.72 logits), indicating that the test includes items covering a relatively broad range of student ability levels. Eleven of the 20 items lie between −0.6 and +0.9 logits, providing greater measurement coverage around the central range of student ability. Targeting was evaluated based on the difference between the person mean and the item mean, where a difference substantially below 1 logit indicates reasonable alignment between student ability and item difficulty [17]. The person mean was −0.12 logits, while the item mean should be interpreted based on the WINSTEPS item-measure output. Therefore, the targeting interpretation is based primarily on the observed overlap between the person and item distributions in the Wright Map [3].

The Wright Map indicates that the test provides reasonable coverage across the central ability range, although the extreme ends of the continuum are less densely represented. At the lower end, only Q1 and Q16 are located below −1.0 logit, while at the upper end, the items above +1.3 logits are primarily the four items identified as misfitting. Consequently, measurement at the extreme ends of the ability continuum should be interpreted cautiously. In particular, estimates for highly able students should be treated with caution because the items located at the upper end did not demonstrate satisfactory fit to the Rasch model [8][27].

Overall, the distribution of persons and items suggests that the test is reasonably targeted to the sampled students, particularly within the central ability range. However, the limited number of items at the extreme ends indicates that additional items with appropriate difficulty levels may be needed to improve measurement precision for students with very low or very high ability.

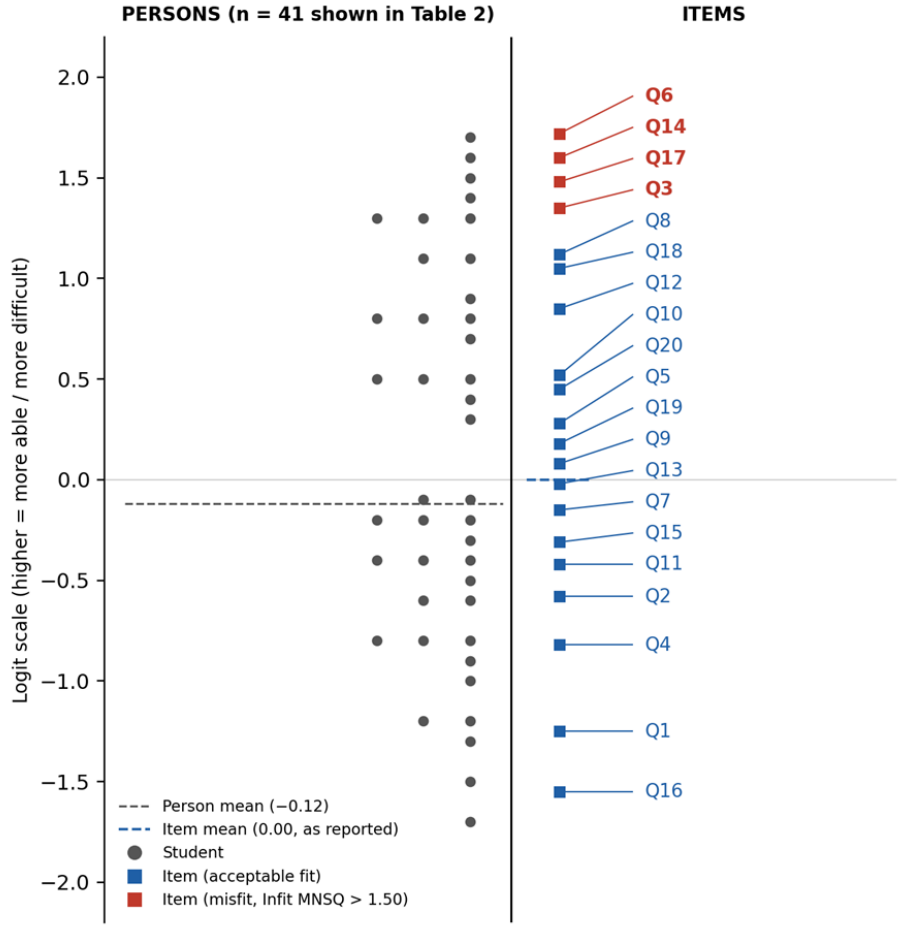


Figure 1. Wright Map (Person–Item Map) of the Newton's Laws concept-understanding test generated directly from WINSTEPS (N = 100). Students are represented by their estimated ability levels on the left, and test items are represented by their difficulty measures on the right on the same logit scale.

3.3 Cognitive Discussion and Diagnosis of Item Misfit

An in-depth analysis was conducted on the items exhibiting statistical misfit or a low item–total correlation ($r_{xy} < r_{table} = 0.195$). Two levels of evidence are distinguished in this section. The statistical evidence (p, rxy, Infit/Outfit MNSQ, and the ability of the students who chose particular options) shows that an item behaved unexpectedly. The cognitive interpretation, which explains why students may have answered as they did, is a hypothesis drawn from the physics content of the item and from earlier research [23]. Because the instrument is single-tier, only the chosen option is observed, not the reasoning or the confidence behind it, so the interpretations below indicate likely misconceptions and do not confirm them.

3.3.1 Cognitive Diagnosis of the Newton's Second Law Sub-topic

The Newton's Second Law sub-topic accounted for the largest number of misfitting items, namely Q6, Q14, and Q17.

a) **Analysis of Q6 (Kinetic Friction and Acceleration):** Q6 presents a case in which an object is pulled by a constant force F across a rough surface with kinetic friction force (fk), and students are asked to determine the object's acceleration (a). The analysis showed that Q6 had p = 0.28, rxy = 0.150, and Infit MNSQ = 1.62. Statistically, the misfit was associated with unexpectedly high-ability students ($\beta n > +1.00$ logits) selecting a distractor that disregards the friction force (a = F/m). Because the single-tier format does not provide students' reasoning, this response pattern can only be interpreted cautiously. A plausible interpretation is that these students applied (F = ma) directly without first determining the net force from the forces acting on the object, such as through a free-body diagram, where (Fnet = F - fk).

b) **Analysis of Q17 (Force Applied at an Angle θ):** Q17 tests understanding of force vectors when an object is pulled by a force F acting at an angle θ to the horizontal plane. With rxy = 0.160 and Infit MNSQ = 1.56, this item was classified as misfitting. The response pattern suggests that some students may have had difficulty resolving the horizontal component of the applied force (F cos θ). Students may have used the total force F directly when applying Newton's Second Law rather than accounting for the horizontal component. If this interpretation is correct, the response pattern would indicate a possible gap between conceptual understanding of force and its mathematical representation. However, because the instrument uses a single-tier format, the data cannot determine whether the difficulty originated from the physics concept, vector decomposition, or the associated trigonometric reasoning [28][24][17].

c) **Analysis of Q14 (An Elevator Accelerating Upward):** Q14 measures understanding of the normal force N exerted by the floor on a person inside an elevator accelerating upward. This item p = 0.30, rxy = 0.175, and Infit MNSQ = 1.52. The response pattern suggests that some students may have relied on the static assumption that the normal force is always equal to the person's weight (N = w), rather than applying Newton's Second Law to the accelerating system. For an upward-accelerating elevator, the relevant relationship is (N - w = ma), giving (N = m(g+a)). The pattern was observed among students across moderate- and high-ability levels, which may indicate that the difficulty is not attributable solely to lower ability. Nevertheless, because option-level response frequencies and students' reasoning were not available, this interpretation should be regarded as a plausible explanation rather than a confirmed misconception.

3.3.2 Cognitive Diagnosis of the Newton's First Law Sub-topic (Inertia)

Analysis of Q3 (Conceptual Force on a Sliding Object): Q3 tests understanding of the force acting on an object sliding across a frictionless surface after the initial pushing force has been released. This item exhibited misfit, with Infit MNSQ = 1.58 and rxy = 0.182. From a statistical perspective, the misfit indicates that the response pattern for Q3 was not fully consistent with the Rasch model expectation. From a cognitive perspective, students who selected an incorrect option appear to have assumed that a force in the direction of motion must continue to act for the object to keep moving. This response pattern is consistent with conceptions reported in earlier research, in which motion is assumed to require a continuing force [2][4][5]. It may also reflect difficulty in distinguishing velocity (v) from acceleration (a), particularly the assumption that uniform rectilinear motion requires a continuing net force. However, because the instrument recorded only the selected answer and did not include a confidence level or reasoning tier, this interpretation cannot be regarded as definitive evidence of a specific misconception. Guessing or other reasoning processes cannot be excluded for individual responses.

3.3.3 Cognitive Diagnosis of the Newton's Third Law Sub-topic (Action–Reaction)

Analysis of Q8 and Q18 (Interaction Force Between Two Objects of Different Mass): Q8 and Q18 measure understanding of the action–reaction force pair in interactions between two objects of markedly different mass, such as a large vehicle and a smaller vehicle. Both items were classified as valid under CTT (rxy = 0.215 and rxy = 0.240, respectively), although their Infit MNSQ values were relatively high (1.38 and 1.35). From a response-pattern perspective, the selection of an option stating that the larger or more massive object exerts a greater force is consistent with the commonly reported tendency to associate interaction-force magnitude with object size or mass rather than applying Newton's Third Law, (Faction = −Freaction) [22]. This interpretation is also consistent with the dominance principle described in previous research, in which the larger or more massive object is perceived as exerting the greater force.

Nevertheless, because Q8 and Q18 did not exceed the established Rasch fit criteria, the evidence for problematic response behavior is weaker than that observed for the four misfitting items. Moreover, the single-tier format does not provide information about students' reasoning or confidence. Therefore, the observed response pattern should be regarded as indicative of a possible conceptual difficulty rather than confirmation of a specific misconception.

3.3.4. Link to Previous Studies and Limits of Single-Tier Diagnosis

Link to previous studies.The interpretations of Q3, Q8, and Q18 correspond to well-documented alternative conceptions in mechanics: the belief that motion requires a continuing force and the belief that the heavier or larger object exerts a greater force during an interaction [2][4][5]. These patterns are consistent with findings reported in previous studies of students' understanding of Newton's Laws.

The Second Law items (Q6, Q14, and Q17) address the application of (Fnet = ma) in contexts involving kinetic friction, an accelerating elevator, and an applied force at an angle, respectively. Difficulties in these contexts are consistent with findings from previous research showing that students may have difficulty identifying the net force when friction is present, distinguishing normal force from weight in accelerating systems, and resolving an applied force into its relevant components. These difficulties suggest that applying Newton's Second Law requires students not only to recall (Fnet = ma), but also to identify and represent the forces acting in a particular physical situation.

The contribution of the present analysis is that these difficulties were identified through a combination of CTT item statistics and Rasch item-fit analysis rather than through incorrect-response frequencies alone. In particular, the misfitting items provide an indication that some response patterns were inconsistent with the expected measurement model. However, item misfit should not by itself be interpreted as direct evidence of a specific misconception. The cognitive explanations presented above therefore remain tentative and require further diagnostic investigation.

Why a four-tier test would give stronger evidence. The interpretations above should be treated as hypotheses. In a single-tier item, only the selected option is observed. An incorrect answer may reflect a misconception, lack of knowledge, a careless error, or guessing, while a correct answer may also result from guessing [6][11]. A four-tier diagnostic test provides stronger evidence by combining an answer tier, a reasoning tier, and confidence ratings for the answer and reasoning. This structure allows researchers to distinguish more carefully between a confident incorrect response accompanied by scientifically inconsistent reasoning and a low-confidence incorrect response that may indicate insufficient knowledge.

Accordingly, the four misfitting items identified in this study should not be interpreted as definitive evidence of specific misconceptions. Instead, they provide potential targets for subsequent diagnostic assessment. Developing these items into a four-tier format, as recommended in the Conclusion, would allow the proposed cognitive interpretations to be examined more directly by incorporating students' reasoning and confidence levels.

IV. CONCLUSION

This study evaluated the psychometric quality of a 20-item Newton's Laws concept-understanding test using Classical Test Theory and the Rasch model with N = 100 tenth-grade students from two schools in Aceh. The instrument showed acceptable psychometric properties: 16 of the 20 items met both the CTT validity criterion and the Rasch fit criteria, internal consistency and item reliability were acceptable to very good, and person–item targeting was adequate. The four misfitting items (Q3, Q6, Q14, and Q17) were also the most difficult items; their response patterns suggest possible misconceptions about impetus, friction, angled-force components, and the normal force in an accelerating elevator, which remain hypotheses because the test is single-tier. Educational implications: the 16 retained items can be used as a validated instrument for classroom-level screening of conceptual understanding of Newton's Laws, and the misfit results point instruction toward free-body diagrams, vector resolution, and accelerating situations. Limitations and recommendations: the sample was limited to two schools, so further validation with larger and more diverse samples (including dimensionality and differential item functioning analyses) is needed before wider use or any claim of standardization. The four misfitting items should be revised and developed into a four-tier diagnostic format to separate misconceptions from lack of knowledge and guessing.

ACKNOWLEDGMENT

The authors wish to thank the principals, physics teachers, and tenth-grade students of the two partner schools, SMAN 3 Banda Aceh and SMAN 1 Darussalam, for granting permission and actively participating in the data collection for this study.

REFERENCES

[1] Alonso, M., & Finn, E. J. (2020). Physics. Addison-Wesley.

[2] Bao, L., & Fritchofer, K. (2019). Assessment of conceptual understanding in physics: A review of instrument development. Physical Review Physics Education Research, 15(2), 020101. https://doi.org/10.1103/PhysRevPhysEducRes.15.020101

[3] Bond, T. G., Yan, Z., & Heene, M. (2020). Applying the Rasch model: Fundamental measurement in the human sciences (4th ed.). *Routledge*. https://doi.org/10.4324/9780429030499

[4] Byman, R. (2005). Curiosity and sensation seeking: A conceptual and empirical examination. *Personality and Individual Differences*, 38(6), 1365–1379. https://doi.org/10.1016/j.paid.2004.09.004

[5] Clement, J. (2008). Creative learning in students: Student misconceptions in mechanics. Springer.

[6] Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. https://psycnet.apa.org/doi/10.1037/h0040957

[7] Fratiwi, N. J., Kaniawati, I., Suhendi, E., Suyana, I., & Samsudin, A. (2017). The Transformation of Two-Tier Test into Four-Tier Test on Newton's Laws Consepts. *Mathematics, Science, and Computer Science Education (MSCEIS 2016)*, *AIP Confer*. https://doi.org/10.1063/1.4983967

[8] Geddis, A. N. (1993). Transforming subject-matter knowledge: The role of pedagogical content knowledge in learning to reflect on teaching. *International Journal of Science Education*, 15(6), 673–683. https://doi.org/10.1080/0950069930150605

[9] Herráez, A. (2006). Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education*, 34(4), 255–261. https://doi.org/10.1002/bmb.2006.494034042644

[10] Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, 30(3), 141–158. https://doi.org/10.1119/1.2343497

[11] Johnson, J. A. (1997). Units of analysis for the description and explanation of personality. In R. Hogan, J. Johnson, & S. Briggs (Eds.), Handbook of personality psychology (pp. 73–93). *Academic Press.*

[12] Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Research in Science & Technological Education*, 35(2), 238–260. https://doi.org/10.1080/02635143.2017.131009

[13] Kaniawati, I., Fratiwi, N. J., Danawan, A., & Suyana, I. (2019). Analyzing Students ' Misconceptions about Newton ' s Laws through Four-Tier Newtonian Test (FTNT). 16(1), 110–122. https://doi.org/10.12973/tused.10269a

[14] Linacre, J. M. (2021). Winsteps® Rasch measurement computer program User's Guide. Winsteps.com.

[15] Mazur, E. (2014). Peer instruction: A user's manual. Pearson.

[16] Mufit, F., & Syamsidar. (2022). Development of Four-Tier Multiple Choice Test Instrument to Identify Students ' Concept Understanding of Newton ' s Law Material. *JIPF (JURNAL ILMU PENDIDIKAN FISIKA),*7(2), 134–144.

[17] Nasledov, A. (2005). SPSS: Computer analysis of data in psychology and social sciences. Piter.

[18] Planinic, M., Ivanjek, L., & Susac, A. (2010). Rasch model-based analysis of the Force Concept Inventory. *Physical Review Special Topics – Physics Education Research*, 6(1), 010103. https://doi.org/10.1103/PhysRevSTPER.6.010103

[19] Posner, M. (2004). Neural systems and individual differences. TC Record, 106(9), 1720–1730.

[20] Riau, F., Majid, A., & Suryana, T. G. S. (2025). Pengembangan instrumen diagnostik miskonsepsi hukum newton dengan model rasch. Nebula, 2(1), 15–21.

[21] Saepuzaman, D., Feranie, S., & Karim, S. (2021). Psychometric evaluation of physics conceptual understanding tests: Transitioning from classical test theory to Rasch measurement model. *Journal of Physics: Conference Series*, 1806(1), 012015. https://doi.org/10.1088/1742-6596/1806/1/012015

[22] Samsudin, A., Fratiwi, N. J., & Suhandi, A. (2020). Diagnosing students' misconceptions on Newton's laws using a four-tier diagnostic test. *Jurnal Pendidikan IPA Indonesia*, 9(3), 332–341. https://doi.org/10.15294/jpii.v9i3.24589

[23] Sidorenko, E. V. (2002). Methods of mathematical processing in psychology. Rech.

[24] Suhandi, A., Samsudin, A., & Fratiwi, N. J. (2025). Altering misconceptions: How e-rebuttal texts on Newton's laws reconstruct students' mental models. *Journal of Science Education and Technology*, 34(2), 112–125.
[25] Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2
[26] Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation. *American Journal of Physics*, 66(4), 338–352. https://doi.org/10.1119/1.18863
[27] Thurstone, L. L. (1959). The measurement of attitude: A psycho-social method and some experiments. University of Chicago Press.
[28] Viiri, J. (2003). Research-based teaching unit on the force concept. *International Journal of Science Education*, 25(8), 993–1008. https://doi.org/10.1080/0950069032000070323
[29] Walker, J., Halliday, D., & Resnick, R. (2014). Fundamentals of physics (10th ed.). Wiley.
[30] Wright, B. D., & Stone, M. H. (1979). Best test design. MESA Press.