FORCE CONCEPT INVENTORY ANALYSIS BY USING INDEXES AND RASCH MODEL

LINDITA HAMOLLI, DODË PRENGA

Department of Physics, Faculty of Natural Sciences,

University of Tirana, Albania

e-mail: lindita.hamolli@fshn.edu.al

Abstract

Understanding basic physics concepts is crucial in high school education. The Concept Inventory (CI) test, introduced by Hestenes in 1992, is widely used to assess conceptual knowledge. The Rasch calibration technique ensures this test's accuracy and provides valuable outcomes. A structured and standardized CI test is an effective measurement tool when certain prerequisites are met. Index analysis can help address potential drawbacks. In this paper, we evaluate the conceptual knowledge inventory of students after high school studies, addressing some problems related to the use of the FCI measurement. First, we used indices to validate the calibration of the test itself with the Rasch method. Then, the result of the index analysis was used to compare the students' overall perception of the test. Further, we compare the results of the FCI calibrated tests with those collected at the end of the physics course in the Computer Science department of the Faculty of Natural Sciences and discuss the persistence of some knowledge gaps and the degree of progress after the university course. For didactic purposes and to facilitate a qualitative generalization of the results, the measurements were performed on a group of students who were assumed to have an average background in physics.

Key words: *Concept Inventory, physics knowledge, indexes, rasch model.*

Përmbledhje

Kuptimi i koncepteve bazë të fizikës është vendimtar në arsimin e mesëm. Testi i Inventarit të Konceptit (CI), i prezantuar nga Hestenes në 1992, përdoret gjerësisht për të vlerësuar njohuritë konceptuale. Teknika e kalibrimit Rasch siguron saktësinë e këtij testi dhe jep rezultate të vlefshme. Një test CI i strukturuar dhe i standardizuar është një mjet matës efektiv kur plotësohen disa parakushte. Analiza e indeksit mund të ndihmojë në adresimin e të metave *të mundshme. Në këtë punim, ne vlerësojmë inventarin konceptual të njohurive të studentëve pas studimeve të shkollës së mesme, duke trajtuar disa probleme që lidhen me përdorimin e matjes FCI. Së pari, ne përdorëm indekse për të vërtetuar kalibrimin e vetë testit me metodën Rasch. Më pas, rezultati i analizës së indeksit u përdor për të krahasuar perceptimin e përgjithshëm të studentëve për testin. Më tej, krahasojmë rezultatet e testeve të kalibruar FCI me ato të mbledhura në fund të kursit të fizikës në departamentin e Shkencave Kompjuterike të Fakultetit të Shkencave të Natyrës dhe diskutojmë për vazhdimësinë e disa boshllëqeve të njohurive dhe shkallën e përparimit pas kursit universitar. Për qëllime didaktike dhe për të lehtësuar një përgjithësim cilësor të rezultateve, matjet u kryen në një grup studentësh që supozohej se kishin një formim mesatar në fizikë.*

Fjalë kyçe: *Inventari i konceptit, njohuritë fizike, indekset, modeli rasch.*

1. Introduction

Conceptual and procedural knowledge tests are two distinct instruments used to assess students' understanding in science and to analyse educational issues and features. From a simplified point of view, the procedural test is constructed as to evaluate the ability of students to resolve problems step by step and their fluency in employing instructed methods, whereas conceptual tests aim measuring students' knowledge of fundamental relationships between variables and features of physical systems. Conceptual knowledge tests are based on the pioneering work of the Force Concept Inventory (FCI) introduced by Hestenes (1992) and have been further developed across various dimensions and disciplines.

A Concept Inventory (CI) test consists of multiple-choice items belonging to the Item Response Theory (IRT) (Embretson & Reise, 2013). The calibration of this measurement instrument is performed by the Rasch technique. Additionally, the theory of indexes provides valuable auxiliary tools regarding usefulness and representativeness (Prenga et al., 2023). We have employed both techniques in the following analysis, with a description of these methods provided in the next section. To the best of our knowledge, the use of CI analyses in didactical studies in Albania is not very common. This circumstance has limited our discussion due to the lack of historical data related to these issues.

However, the impressive advancements of our physicists abroad and the successful careers of students educated in physics in Albania, strongly suggest that the recent unsatisfactory levels in physics could a localized and temporary event. This issue can be addressed by correctly identifying the causes, which can be achieved through direct measurement and interdisciplinary analysis.

In this context, the use of the CI test to assess student knowledge in physics was initiated several years ago as part of a master's degree Thesis in AML. Aside from the practical novelty of using this instrument, initial observations indicated that conceptual knowledge in physics was not satisfactory. Initially, the analysis of conceptual knowledge in physics for our students has been considered within the context of the negative effects of online learning during COVID-19 restrictions, (Kushta et al., 2022, Prenga et al., 2023), followed by several other investigative and analytical works of a amore general view, (Prenga, 2024)

Furthermore, several problems related to physics teaching have been analysed, including infrastructural limitations that hinder the support of newly implemented teaching methods, and the physics literature used for high school education (Hafizi et al., 2023; Boçi & Prenga, 2022). In this framework, we observe that during the first two years of gymnasium, students have two 45-minute physics lessons per week. In the third year, students choose physics based on their intended university studies, having four 45-minute physics lessons per week. Mechanics is taught throughout the first year (ages 14–15) and revisited in the third year for students who select this subject. The impact of insufficient laboratory support on students' advancement in physics has been specifically highlighted (Boçi & Prenga, 2022).

Based on these findings and our everyday observations, as well as feedback gathered from physics teachers through a general questionnaire, we hypothesize that shortcomings in conceptual knowledge inherited from high school significantly affect the acquisition of knowledge after university physics courses and influence the effectiveness of the CI test itself.

For this study, we conducted direct measurements on students of Informatics, considering that they represent an interesting target group with a sufficient background in physics, though not necessarily as strong as that of physics students. From a statistical point of view, recognising the difficulty of an adequate random sampling, we considered and treated this group on the quality of the convince sampling, that permit us to perform the analysis with the price of quantitative limitations.

2. Data collection and analysis by test indexes

The school system in Albania consists of nine years of elementary school followed by three years of high school. After high school, students can continue their education at various universities and colleges. There are several types of high schools in Albania, but most students who continue their education at the Faculty of Natural Sciences come from high schools. Physics is taught as a separate and compulsory subject from the sixth grade of elementary school (ages 13–14) until the third year of gymnasium (ages 17–18). In the sixth-grade students have one 45-minute physics lessons per week, in the seventh, eighth, and ninth grades of elementary school, they have two 45-minute physics lessons per week. In high school, the number of physics lessons per week depends on the type of school.

To estimate the average level of conceptual understanding of mechanics for Albanian students at the end of gymnasium (a typical high school in the Albanian education system that prepares students for universities), a representative sample of students was pretested with the FCI at the beginning of the 2023-2024 academic year. Specifically, the FCI test was administered to first-year students in the Informatics Branch of the Faculty of Natural Sciences. Almost all these students had completed high school. This provided a dataset (N=84) that was later analysed using the Indexes and the Rasch model.

The FCI is a multiple-choice test consisting of 30 questions designed to assess students' conceptual understanding of Newtonian force with minimal reliance on mathematics. Most questions in the FCI were slightly changed over the years. The latest version of the test which we also used in this study can be found on the web (http). This test evaluates how well students grasp the concepts of force and motion after studying Newtonian mechanics. One of the advantages of the FCI is, its ability to be easily administered to large groups of students, making its results both impactful and significant. According to the FCI authors (Hestenes, 1992), a score of 60% is considered the threshold for developing Newtonian thinking. Students scoring below this threshold typically have insufficient understanding of Newtonian concepts for effective problem-solving and may struggle with university-level physics courses (Hestenes, 1995).

The test was carefully translated into Albanian. The testing was conducted anonymously. The allocated time for taking the test was 30 minutes. Participating students were informed in advance about the FCI test and the rules to be followed. No incentives, such as grades, were offered to students for taking the test. Also, the purpose of the research and the importance of the test were explained to the students in advance, they showed interest in the test and wanted to know their results.

All students who were tested had studied mechanics during their first year of gymnasium, but only 88% continued with it in their third year (refer to Figure 4). The testing took place at the start of their bachelor's studies, with a gap of three and one years between their mechanics regular course and the current FCI testing. Despite the significant gap between learning mechanics and the testing, students had been engaging with other physics topics that incorporated Newtonian concepts. Consequently, these concepts should have been utilized by the students over the years, enhancing their understanding through application in various contexts.

2.1. Compatibility of the FCI test by using Indexes

CI testing, like every measurement procedure, might suffer from inaccuracies. Often, a CI test aims to discover shortcomings and common-sense mistakes in addition to assessing knowledge. Normally, a CI test can be calibrated through the Rasch technique, but when discussing complex measurements like scientific knowledge, deviations from expected outcomes can be seen as additional information. Bearing these arguments in mind, we have performed a test compatibility and reliability analysis, as well as a factorial diagnosis, by analysing the test's indexes. The test's indexes consist of statistical estimators of testing integrity, significance, validity, discriminatory power and difficulty measure (Ding et al., 2006; Aubrecht et al., 1983). Here are their definitions and calculation formulas.

- **Item difficulty index**

The item difficulty index $\mathrm{P_i}$ measures the difficulty of a test question (i). It is calculated as the ratio of the number of correct responses $\mathrm{N_i}$ to the total number of students N who attempted the question: $\mathrm{P_i} = \mathrm{N_i}/N$. The difficulty index $\mathrm{P_i}$

might be more accurately called the "easiness index," as it represents the proportion of correct responses to a particular question. A higher P_i value indicates a higher percentage of correct answers, making the item easier for the population. The difficulty index ranges from 0 to 1, where 0 means no one answered correctly, and 1 means everyone answered correctly. While these extreme values are possible, they are generally not useful for measurement purposes and should be avoided.

There are various criteria for acceptable difficulty index values (Doran, 1980). For the FCI, we use a widely adopted criterion that requires the difficulty index to be between 0.3 and 0.9, with an optimum value of 0.5. However, controlling every item in a test can be challenging, especially as the number of items increases. Therefore, an average difficulty index value $\overline{P}$ of all items P_i in a test is often used to indicate the overall test difficulty:

$$\overline{P} = \frac{1}{K}\sum_{i=1}^{K} P_i \tag{1}$$

Figure 1 shows the difficulty index P_i values for each item in the FCI test performed before the teaching process, based on a combined sample of 84 students. The FCI item difficulty index values range from just below 0.2 to slightly above 0.6, with most items falling between 0.2 and 0.4. The average difficulty index $\overline{P}$ is 0.29, which is lower than the lower bound of the acceptable range [0.3, 0.9]. According to Ding (2006), a difficulty index below

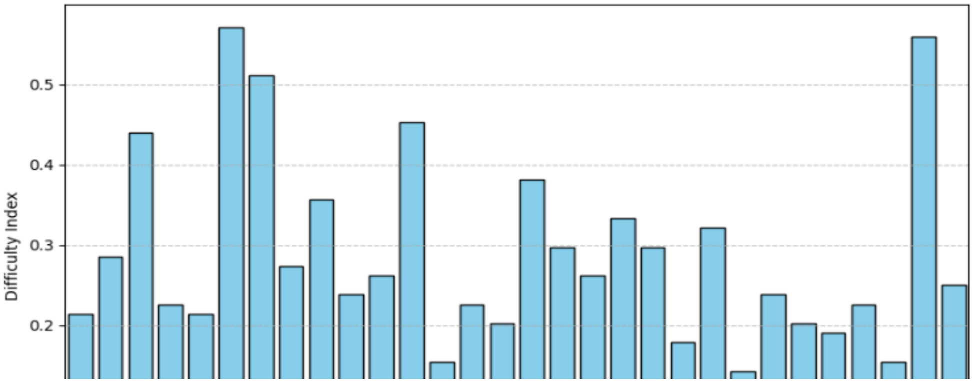


Figure 1. FCI item difficulty indices for each question, based on a sample of 84 students.

0.3 indicates serious testing issues. For the FCI, a low difficulty index suggests higher-than-expected difficulty. In the pre-course test, items 1, 2, 4, 5, 6, 10, 11, 13, 14, 15, 16, 22, 23-28, and 30 were perceived as very difficult by the 84 students, covering all mechanics subjects in the FCI. This suggests that physics is challenging for most students, despite their high school results and choice of a physics and mathematics-based branch.

- **Item discrimination index**

The item discrimination index, D measures the ability of a test item to distinguish between students who know the material well and those who do not. A high discrimination index indicates that students with robust knowledge usually answer correctly, while those with weaker understanding do not. Conversely, a flawed question might mislead thoughtful students to incorrect answers, while less thoughtful students might answer correctly. Tests with many high-discrimination items effectively separate strong students from weak ones. To calculate the discrimination index D, divide the sample into two equal groups: a high group H and a low group L, based on whether their total scores are above or below the median.

Count the number of correct responses in both groups N_H and N_L. If the total number of students is N, the discrimination index D can be calculated as: $D = (N_H - N_L)/(N/2)$. In educational and psychological studies, there are several different calculations of discrimination index often employed by researchers. The calculation described above (50%–50%) is the one which we adopted to calculate discrimination indices for FCI items. Other researchers may use the top 25% as the high group and the bottom 25% as the low group (25%–25%). For (50%–50%) calculation the discrimination index D can be expressed as:

$$D = \frac{N_H(\text{top } 50\%) - N_L(\text{bottom } 50\%)}{N/2}. \qquad (2)$$

It evaluates the discriminative power of test items considering all students, while the calculation (25%-25%) uses only the most stable individuals by discarding half of the available data. The item discrimination index D ranges from -1 to +1, with +1 being the best value and -1 the worst. Ideally, all high group students answer correctly, and all low group students answer incorrectly, resulting in D = +1. Conversely, if all low group students answer correctly and all high group students answer incorrectly, D = -1. While these extremes are rare, items with negative discrimination indices should be eliminated.

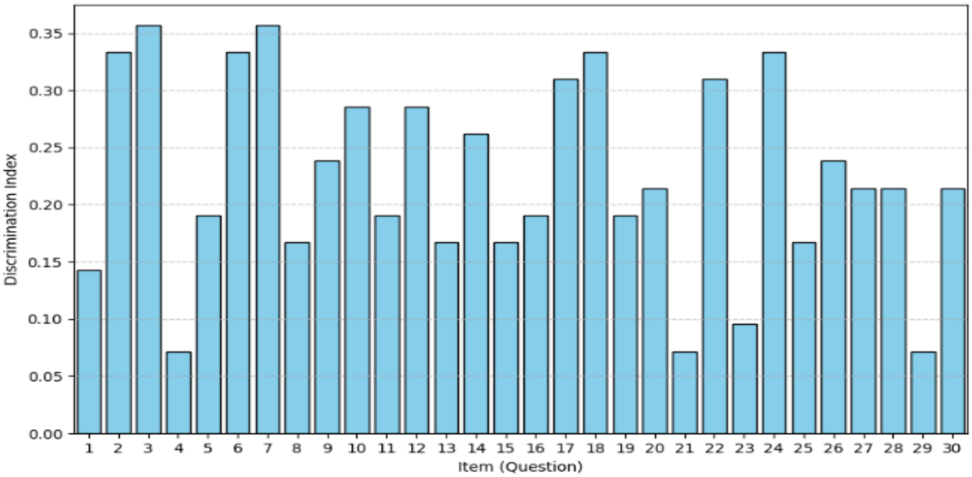


Figure 2. FCI item discrimination indices for each question, based on a combined sample of 84 students. The average discrimination index is 0.224 (50% method).

An item is considered to provide good discrimination if $D \geq 0.3$ (Doran, 1980). Items with D between 0 and 0.3 are not necessarily bad, but most items in a test should have high discrimination indexes to effectively distinguish between strong and weak mastery of the material.

Figure 2 shows the discrimination index for each FCI item. Most of the discrimination index D values for FCI items range from 0.1 to 0.5, with the majority (18 items) around 0.1–0.3. This indicates that most FCI items do not meet the required discriminatory power. We also calculated the average discrimination index $\overline{\mathrm{D}}$ for all FCI items and found it to be 0.224, which does not meet the criterion of $\overline{\mathrm{D}} \geq 0.3$. To illustrate the underestimation of the 50%-50% calculation, we also computed FCI item discrimination indices using the 25%-25% method. The index values for all 30 items increased, resulting in an average discrimination index $\overline{\mathrm{D}}$ of 0.37 with the 25%-25% calculation.

- **Point biserial coefficient**

The point biserial coefficient, sometimes referred to as the reliability index for each item, measures the consistency of a single test item with the entire test. It reflects the correlation between students' scores on an individual item and their scores on the whole test, essentially serving as a form of the correlation coefficient. The point biserial coefficient ranges from -1 to +1. A highly positive correlation indicates that students with high total scores are more likely to answer the item correctly, while a negative value suggests that students with low overall scores are more likely to get the item correct, indicating that the test item may be defective.

To calculate the point biserial coefficient for an item, one needs to determine the correlation coefficient between the item scores and the total scores. A student's score on an item is a dichotomous variable, which can only have two values: 1 (correct) or 0 (incorrect). Scores for the entire test are usually viewed as continuous, especially if the test has a relatively large number of items (e.g., 20 or more). The correlation coefficient between a set of dichotomous variables (item scores) and a set of continuous variables (total test scores) is used to calculate the point biserial coefficient (Ghiselli, 1981),

$$\mathrm{r_{pbs}} = \frac{\overline{\mathrm{X_1}} - \overline{\mathrm{X}}}{\sigma_{\mathrm{X}}} \sqrt{\frac{\mathrm{P}}{1-\mathrm{P}}}. \qquad (3)$$

Here, $\overline{\mathrm{X_1}}$ is the average total score for students who score 1 on the test item (i.e., correctly answer the item), $\overline{\mathrm{X}}$ is the average total score for the entire sample, σ_{X} is the standard deviation of the total score for the entire sample, and P is the difficulty index for this item. For instance, in item 1 of the FCI pretest, 18 out of 84 students answered correctly, resulting in P = 0.214. For those 18 students, the average total score $\overline{\mathrm{X_1}}$ is 13.17. For all 84 students in the sample, the average total score $\overline{\mathrm{X}}$ is 8.67. With the standard deviation $\sigma_{\mathrm{X}} =$ 4.88 of the total score for the entire sample, we can calculate the point biserial coefficient for FCI item 1 to be approximately 0.482. Ideally, all items in a test should be highly correlated with the total score. However, this is somewhat unrealistic for a test with many items. The widely adopted criterion for measuring the "consistency" or "reliability" of a test item is $r_{pbs} \geq 0.2$ (Kline, 1986). Items with point biserial coefficient lower than 0.2 can still remain in a test, but there should be few such items. One way to check whether there are a majority number of items satisfying $r_{pbs} \geq 0.2$ is to calculate the average point biserial coefficient $\bar{r}_{pbs}$ of all items K in a test:

$$\bar{r}_{pbs} = \frac{1}{K}\sum_{i=1}^{K}(r_{pbs})_i \qquad (4)$$

where K is the number of items and $(r_{pbs})_i$ is the point biserial coefficient for the i-th item.

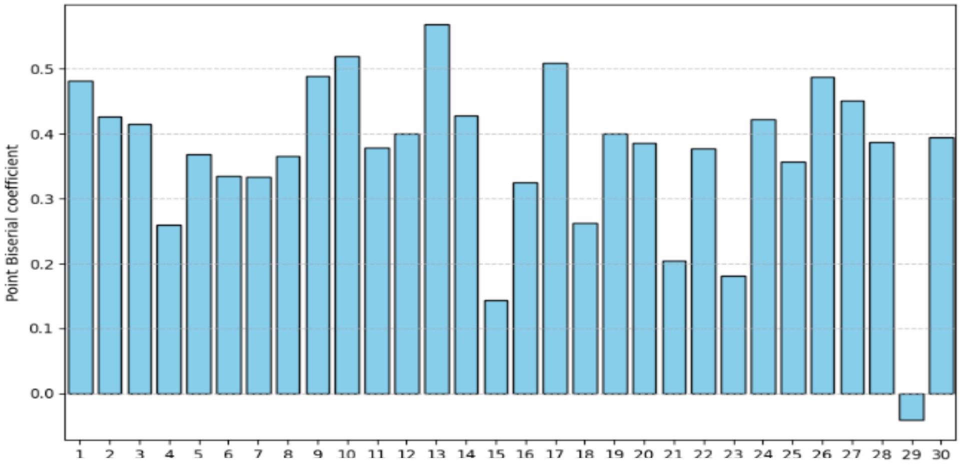


Figure 3. FCI item point biserial coefficient from a sample of 84 students.

The average point biserial coefficient for FCI is 0.37, which is greater than the criterion value 0.2, so FCI items overall have fairly high correlations with the whole test. Figure 3 provides the point biserial coefficient values for each FCI item. As one can see, almost all items have satisfactory r_{pbs} values, indicating that almost all FCI items are reliable and consistent. The discrimination index and point biserial coefficient are two different statistics measure of an item.

The first measures how effectively an item separates strong and weak students, while the second measures whether an item is consistent with the entire test. An item could have a fairly high discrimination index value but show little consistency with the test as a whole. In such a case, the item might be testing a topic different from the main subject matter of the rest of the test. Conversely, an item could be consistent with the test as a whole (high point biserial correlation coefficient) but offer little discriminatory information.

- **Kuder-Richardson reliability index**

The Kuder-Richardson reliability index measures the self-consistency of an entire test. If a test is administered twice (at different times) to the same sample of students, we would expect a highly significant correlation between the two test scores, assuming the students' performance is stable, and the test environmental conditions are the same on each occasion. The correlation coefficient between the two sets of scores is defined as the reliability index of the test. However, this approach does not provide a practical way of determining the reliability index of a test, as students may remember the test questions and study for the test, and test conditions at different times may not be identical. Kuder and Richardson further developed this idea and proposed to divide a test into its smallest components – items and defined the reliability index of a test as:

$$r_{\text{test}} = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\delta_i^2}{\delta^2}\right) = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}P_i(1-P_i)}{\delta^2}\right). \qquad (5)$$

K is the number of the test items, δ_i is the standard deviation of the i-th item score, δ is the standard deviation of the total score and P is the difficulty index of an item. Possible values for the r_{test} fall into the range [0,1]. Different tests for various purposes have different criteria. A widely accepted criterion is that tests with reliability index higher than 0.7 are reliable for group measurement and tests with reliability index higher than 0.8 are reliable for individual measurement.

Under most circumstances in physics education, evaluation instruments are designed to be used to measure a large group of students, so if a certain physics test has a reliability index greater than 0.7, one can safely claim it is a reliable test. In the FCI analysis, we adopted Kuder-Richardson formula (5) to calculate the reliability index. We find the reliability index for FCI pretest to be 0.73, which is satisfactorily high for group measurement.

- Ferguson's delta

Ferguson's delta is another whole-test statistic. It measures the discriminatory power of an entire test by investigating how broadly the total scores of a sample are distributed over the possible range. If a test is designed and employed to discriminate among students, one would like to see a broad distribution of total scores. The calculation of Ferguson's delta is based on the relationship between total scores of any two subjects (students). These scores may either be different or equal (Ding, 2006). The discriminatory power is given by the relationship

$$\delta = \frac{N^2 - \sum_{i=1}^{K} f_i^2}{N^2 - N^2/(K+1)}, \tag{6}$$

where N is the number of students in a sample, K is the number of test items, and f_i is the frequency (number of occurrence) of cases at each score. The possible range of Ferguson's delta values is [0,1]. If a test has Ferguson's delta greater than 0.9, the test is considered to offer good discrimination. Ferguson's delta for our FCI pretest is 0.95, which is greater than 0.9.

2.2. Improvement of the test indexes after the course

The reliability and discriminatory power of the FCI pretest were evaluated using five statistical tests: three for individual items and two for the overall test. After the teaching process, the same group of students took the FCI test again (post-test).

This time, each student was assigned a unique code to identify their test scores later and compare them with their pre-test results. The results before and after the teaching process are summarized in Table 1. As shown, the values of the five statistical tests have slightly improved compared to the pre-test. In the FCI post-test, all indexes fall within the validity zone, ensuring trustworthy and reliable results. Consequently, all statements regarding the CI-scores and other statistical quantities measured for the sample can be extended and considered as characteristics of the population.

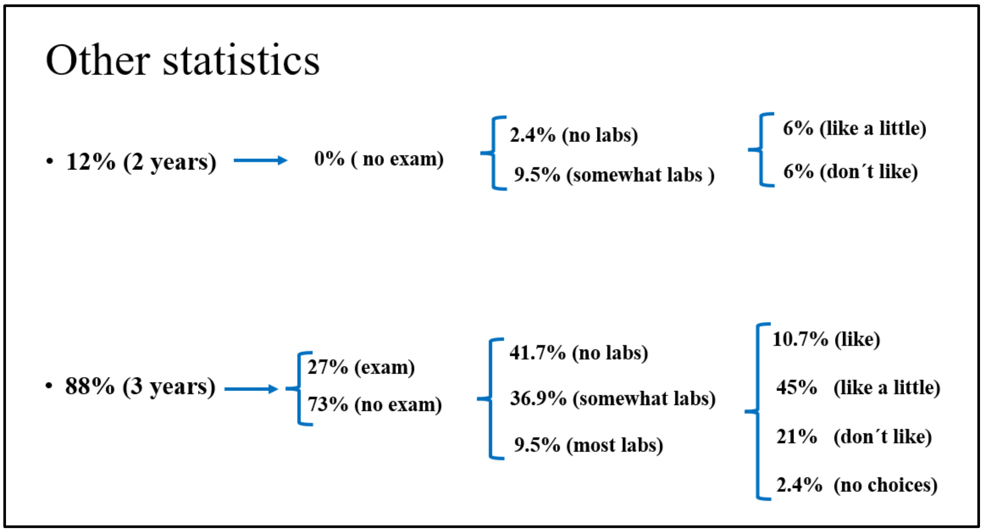
Table 1. The indexes for FCI-Before and FCI-After tests for a sample of 84 Informatics Branch Students

FCI	**Difficulty index**	**50% to 50% Discrimination index**	**Reliability Index (Point Biserial)**	**Self-consistency index**	**Discrimination power (Ferguson delta)**
Before Course	0.29	0.224	0.37	0.73	0.94
After Course	0.44	0.315	0.43	0.85	0.96
Reference values	≥ 0.3	≥ 0.3	≥ 0.2	≥ 0.7	≥ 0.9

In Figure 4, we present statistics for the group of 84 students who participated in the FCI test. Twelve percent of these students studied physics for only two years, while the remaining students studied it for three years. Among the latter group, only 27% chose physics as a subject for the matriculation exam. Approximately 10% of all students completed a laboratory course in high school, and only 10.7% expressed an interest in the subject of physics.

3. Indicatory findings of the FCI by Rash analysis

After confirming that the test indexes fall almost within the desired range, the test results are considered reliable and trustworthy. However, the results of the confidence interval should be considered indicative, since the testing procedure did not fully comply with the statistical requirements for randomness due to some objective limitations (we have tested only the Informatics students, who were required to choose physics to pursue this branch of study). The final results were obtained using Rasch analysis, which we briefly describe here.



Other statistics

- 12% (2 years) → 0% (no exam)
  - 2.4% (no labs)
  - 9.5% (somewhat labs)
    - 6% (like a little)
    - 6% (don´t like)
- 88% (3 years) →
  - 27% (exam)
  - 73% (no exam)
    - 41.7% (no labs)
    - 36.9% (somewhat labs)
    - 9.5% (most labs)
      - 10.7% (like)
      - 45% (like a little)
      - 21% (don´t like)
      - 2.4% (no choices)

Figure 4. Statistics for the group of 84 students who participated in the FCI test, detailing the duration of the high school physics course, the physics selection for the matriculation exam, the completion of their laboratory work, and their interest to the physics subject.

Rasch analysis is a psychometric technique developed to increase the accuracy of instrument construction, monitor instrument quality, and evaluate respondent performance (Boone, 2016). Notably, it provides calibrated assessment of the concept inventory outcomes: the student ability to solve the test, the items' difficulties, the estimated probability for a student to solve an item, pathological behaviours as guessing (Planinic, 2010). Here is the description of its core calculation procedure. Initially, the answers of the FCI test are recorded in a matrix $(i\ ,j\ )$ = (0,1) by assigning (1) for correct answer and (0) for incorrect one (Zaiontz, 2023). Unanswered questions are left blank. One calculates student's average scores obtained for the test $(i\ )$ and the average scores that all students realized for the item $(j\ )$:

$$P(i) = \frac{1}{NumberItems}\sum_{j=1}^{NumberItems} T(i,j);\quad P(j) = \frac{1}{NumberStudents}\sum_{i=1}^{NumberStudents} T(i,j)$$

Next, the student's ability to solve the test, β_i and the item's difficulty perceived by all students, δ_j are calculated by equations:

$$\beta_i = ln\frac{P(i)}{1-P(i)}, \quad \delta_j = ln\frac{1-P(j)}{P(j)} \qquad (7)$$

Quantities in equations (7), are measured in logit units, which are linear and homogeneous (Dode et al., 2023). Using them, the probability $P_e(i,j)$ that student (i) having the ability β_i could solve the item (j), whose difficulty is perceived δ_j is

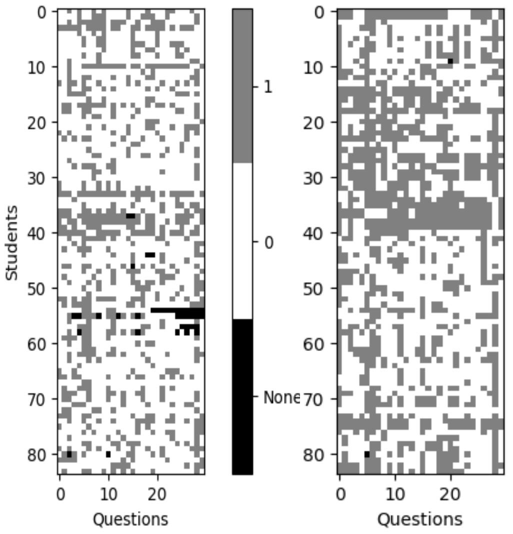
$$P_e(i,j) \equiv P(\beta_i, \delta_j) = \frac{\exp(\beta_i - \delta_j)}{1+\exp(\beta_i - \delta_j)}. \qquad (8)$$

The model defines the unit of measurement called a logit (log odds unit), which is used to measure both item difficulties and person abilities. The estimated measures are expressed on the logit scale, with the average item measure arbitrarily set at 0. These estimates are then adjusted for variance effects and iterated against each other until they meet a preset convergence criterion, resulting in a set of internally consistent item and person parameters (Planinic et al., 2010). The measures are linear, which is a crucial characteristic of the Rasch model. For example, a person with an ability of 3 logits has three times more ability than a person with an ability of 1 logit. This is fundamentally different from scores expressed as percentages, where it is impossible to say that a person who scores 30% on a test has three times more ability than a person who scores 10% on the same test. Percentages can reflect the correct ranking of persons or items but not the correct intervals between their abilities or difficulties.

In Figure 5, we present a grid of students' answers to 30 questions of the FCI pretest and post-test. The correct answers are indicated by grey squares, incorrect answers by white squares, and unanswered questions by black squares. Figure 6 shows the item-person map for students in the pre-test and post-test. After question and person calibrations are obtained, they are placed on a vertical ruler that measures person ability and item difficulty on the same logit scale. On the right-hand side of the ruler, the FCI questions are sorted by difficulty, with the most difficult items at the top and the easiest items at the bottom. On the left-hand side of the ruler, persons are sorted by their success on the FCI, with the most successful students at the top. It is evident from Figure 6 that the pre-test was more difficult for the students, as the distributions of item difficulties and person abilities are significantly more shifted relative to each other.

The mean item difficulty is 2 logits above the mean person ability in the pre-test. Ideally, the test should be centred on the target population. This plot also clearly shows the ordering of questions according to their difficulty. Questions with negative calibrations are easier, while those with positive calibrations are more difficult than the question average, which is set at zero. The spacing between questions is also very important. Questions should not be too close in difficulty, as this would make one question indistinguishable from the next. However, the separation between individual items should also not be too large to avoid significant gaps. Inspection of Figure 6 reveals that in the pre-test, the width of the items is about 2 logits, whereas the width of the person distribution is almost 4 logits. Most of the items are in the region between 0 logits and +2 logits, but only 1% of all students can be found in this range. For this sample of students, there are enough hard items but not enough easy items (see the Figure 7 for more details).

By employing the filtering capacity of Indexes analysis with Rasht evaluation and calibration technique, we have measured the students' ability in mechanics, perceived difficulty for mechanics in general, the efficacy of the university courses to repair conceptual shortcomings in physics inherited from the High School education etc. We considered in the first stance the raw data from the pre-test. We observed that the average of the FCI scores lies far under the 60% threshold of fundamental knowledge in physics. Also, the rate of guessing was critically high for the pre-test. More than 93% of the students have guessed at least one answer that is when the condition is fulfil: $\frac{\exp(\beta_i-\delta_j)}{1+\exp(\beta_i-\delta_j)} < 0.5 \wedge T(ij) = 1$. Considering these results we have qualified the FCI results as indicatory. However, recognising that no academic consequences would follow students' tests, we believe that this is an indicator of serios problems in understanding physics at High School. Notice that those result coincide with indexes findings by the which we de-qualify the outcomes



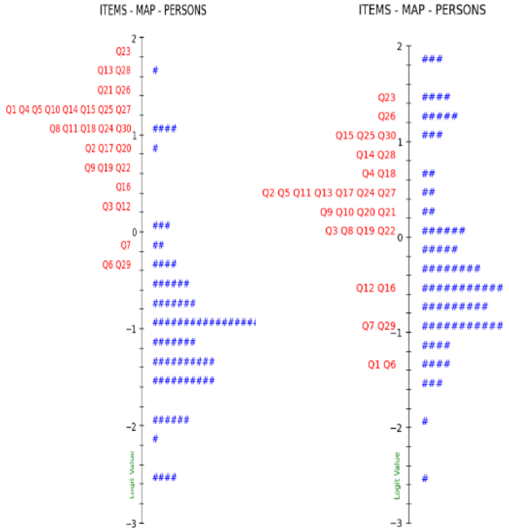


Figure 5. *Left:* The grid of student answers for the 30 questions on the pre-test, and *right:* the post-test. Correct answers are shown in gray, incorrect answers in white, and unanswered questions in black.

Figure 6. *Left:* Item-person map for students in the pre-test, and *right:* post-test. The right-hand vertical ruler displays the distribution of student abilities, while the left-hand side shows the distribution of item difficulties. Questions are labelled as Q1–Q30. Each # represents one student.

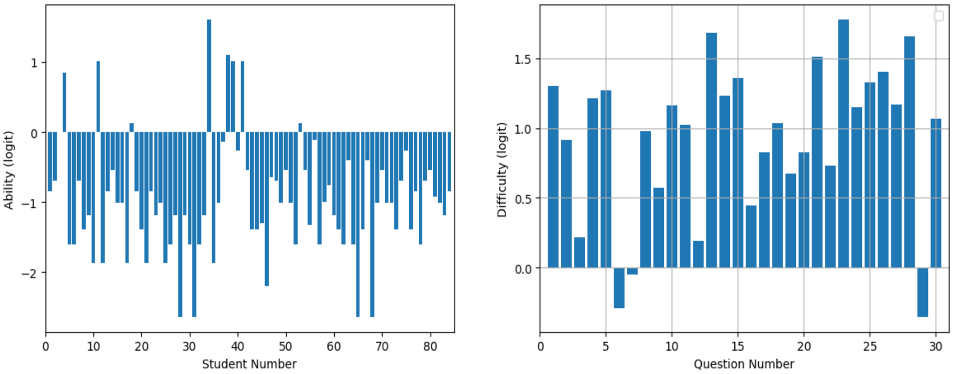


Figure 7: The student abilities and the question difficulty in the logit scale in the pretest.

As indicated above, in post-test, all indexes lie on the validity zone, that is a guarantee for trustworthy result and their reliability. In the first remark we observe that again, students' abilities and test difficulties do not match perfectly (see Figure 6 on the right). Despite the compliance of indexes, it resulted that regarding to sociometric approach of the findings, students' ability is lower than the difficulty of the test. It certifies that the reported perception of students that physics is very difficult referred in (Kushta et al., 2022), persists.

However, based on indexes' filtering the test by itself is conclusive. So, by this model estimating, the level of knowledge of physics for the group of students interviewed is evaluated around 29% which is significantly low than the desired threshold level at 60%. Despite the assessment of the FCI score for the pretest was not conclusive because of indexes un-compliances, we might estimate the improvement rate of conceptual knowledge,

$$G_{course} = \frac{\%CI_{after} - \%CI_{before}}{100 - \%CI_{before}} = \frac{44\% - 29\%}{1 - 29\%} \approx 21.1\% \,. \qquad (9)$$

This result testify that conceptual knowledge shortcomings do persist and make the advance very difficult. We observed that the distribution of the abilities regarding mechanics' knowledge, has improved in the sense that it becomes more symmetrical and regular (see Figure 6). It indicates that during the course, several basic concepts have been understood and clarified for most of the students which followed the course, smoothening original harsh differences. On the other side, guessing behaviour as a latent indicator of the conceptual knowledge failure has diminished. Also, the statistical power of the pre-test has improved. In Figure 8are shown outfit values of questions and the students and their admissible level is in the range [0.7-1.3]. So only the question 29 do not fulfil this condition. In principle, this question has not been understood clearly and more effort should be made to improve the transmission of knowledge embodied on this question.

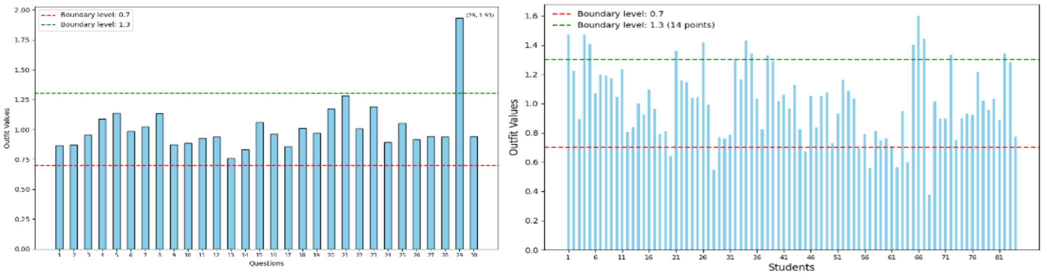


Figure 8: Qualitative analysis of the statistical significance of the findings in the pre-test.

After removing the question 29, in Figure 9 is presented a bubble chart of test questions, where each circle's size corresponds to the Rasch standard error of the question's calibration. Smaller circles indicate lower calibration uncertainty. Ideally, circles should be well-separated without large gaps, and close to the central axis, indicating good model fit. However, some circles overlap, and many are close in difficulty, making item ordering unclear. Larger circles for harder items at the top reflect fewer responses. Items far from the expected value of 1, such as questions 13 and 14 (more regular patterns) and questions 23, 21, and 20 (unpredictable patterns), indicate potential issues. While some irregularity is expected, too much regularity or unpredictability can threaten measurement validity. Questions 23 and 21's moderately large outfit values are likely due to lucky guesses, posing no serious issue.

We observe that the scores obtained individually and by the group on standard exams are higher than the results obtained herein by FCI test, that is a strong indicator for inherited defects form high school. So, in our standard exams students are checked by procedural exams, following the same methods utilized on previous stage of the education. By nature, we cannot check conceptual knowledge in full scale as the FCI did, because the branch where the test has been conducted has basically an engineering nature. But the finding is important, because there is a significant discrepancy between conceptual knowledge and procedural one. So, procedural knowledge can be improved significantly during a university course, whereas it seems very difficult to achieve remarkable improvement of conceptual knowledge if it has been damaged during precious stage of the studies

Conclusions

Our objective was to assess the level of conceptual understanding of mechanics among Albanian students in their final year of gymnasium. Although the tested group primarily consisted of students who chose physics in their last year of high school, we found that a significant portion of those students lack a Newtonian Physics understanding. Among several factors, mostly hidden, we believe that restricting physics learning in high school on algorithmic solution of the problems, has a significant effect.

As a result, students tend to focus on procedural and algorithmic problem-solving neglecting conceptual an in-depth knowledge. While traditional exams can measure progress during the course, the FCI test reveals persistent gaps in conceptual understanding from earlier educational stages. Therefore, our efforts should be directed towards enhancing conceptual knowledge to advance physics education.

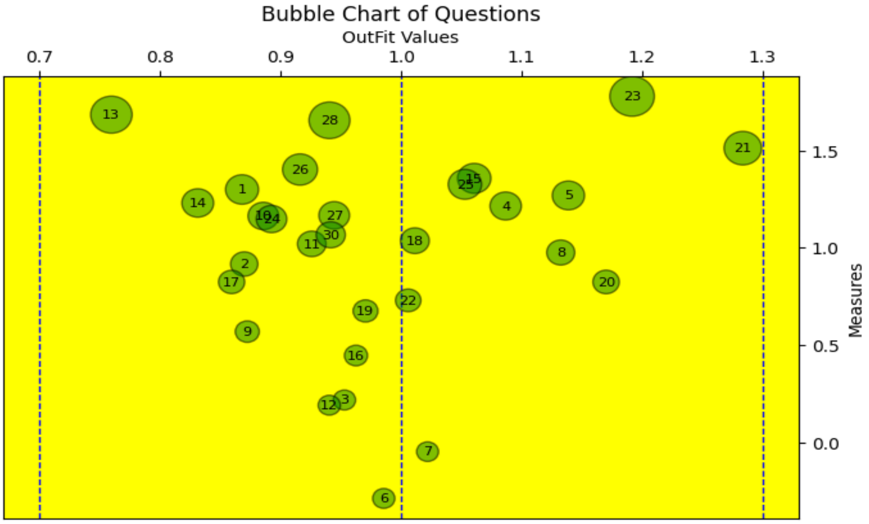


Figure 9: Bubble chart illustrating the relationship between outfit values and question measures for the pretest.

References

Aubrecht, G. J., & Aubrecht, J. D. (1983). Constructing objective tests. American Journal of Physics, 51(7), 613–620. doi:10.1119/1.13186.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? CBE Life Sciences Education, 15(4). doi:10.1187/cbe.16-04-0148.

Boçi, S and Prenga, D. (2022). Book-of-Abstracts-Second-International-Conference-of-NIP.pdf (ikf-akad.al).

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. Physical Review Special Topics - Physics Education Research, 2(1), 10105.

doi:10.1103/PhysRevSTPER.2.010105.

Doran, R. (1980). Basic Measurement and Evaluation of Science Instruction (NSTA, Washington, DC).

Embretson, S. E., & Reise, S. P. (2013). Item Response Theory. Psychology Press, London, United Kingdom. doi:10.4324/9781410605269.

Ghiselli, E., Campbell, J., and Zedeck, S. (1981). Measurement Theory for the Behavioural Sciences (Freeman, San Francisco).

Hafizi, M. et al., (2023). Prezantimet-e-Forumit-Fizika-ne-arsimin-parauniversitar.pdf (ikf-akad.al).

Hake, R. R. (1998). Interactive engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. 66, 64.

Hestenes, D., Wells, M. and Swackhamer, G. (1992). Force concept inventory. The physics teacher, 30(3), pp.141-158.

Hestenes, D. and Halloun, I. (1995). Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. 33, 502. https://modeling.asu.edu/research-and-evaluation.

Kline, P. (1986). A Handbook of Test Construction: Introduction to psychometric design (Methuen, London).

Kushta, E., Dode Prenga, S. M., & Dhoqina, P. (2022). Assessment of the Effects of Compulsory Online Learning During Pandemic Time on Conceptual Knowledge Physics. Mathematical Statistician and Engineering Applications, 71(4), 6382-6391. doi:10.17762/msea.v71i4.1228.

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), 1-11. doi:10.1103/physrevstper.6.010103.

Prenga, D., 2024. A Thematic Review on the Combination of Statistical Tools and Measuring Instruments for Analysing Knowledge and Students' Achievement in Science. European Modern Studies Journal, Vol 8, No 3

https://journal-ems.com/index.php/emsj/article/view/1173/1022

Prenga, D., Kushta, E., Mysli, F. (2023). Enhancing Concept Inventory Analysis by Using Indexes, Optimal Histogram Idea, and the Likert Analysis. Journal of Human, Earth, and Future. 2023 Mar 1;4(1):103-20. http://dx.doi.org/10. 28991/HEF-2023-04-01-08

Zaiontz, C. (2023). Building a Rasch Model. Real Statistics Using Excel. Available online: https://real-statistics.com/reliability/item-response-theory/building-rasch-model/ (accessed on February 2023).