POLITECNICO
MILANO 1863

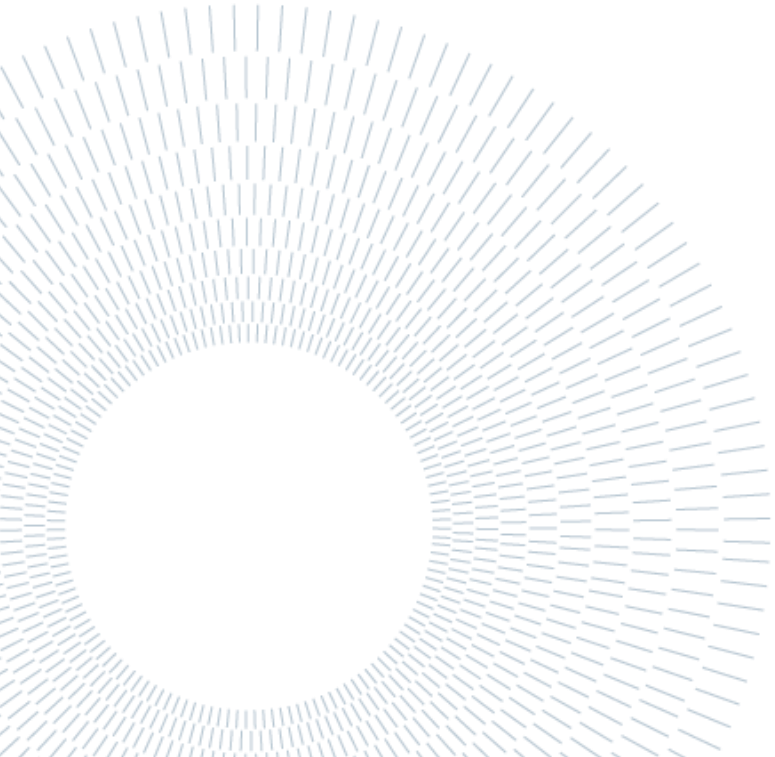
SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Item analysis of a physics multiple choice test designed for an orientation course

MASTER'S THESIS IN
ENGINEERING PHYSICS

Author: **Sara Pittini**

Student ID: 991764
Advisor: Maurizio Zani
Co-advisor: Matteo Bozzi
Academic Year: 2022-23



Abstract

This thesis work analyses a multiple-choice test consisting of 8 physics items. The test was administered to students doing their fourth and fifth year of high school who were participating in an orientation course offered by Politecnico di Milano and were interested in pursuing engineering studies after high school.

The items had a similar structure to the questions found on the engineering entrance exam, allowing students to gauge the level of knowledge required to pass the test. Additionally, the test aimed to address some common misconceptions and conceptual errors among students, enabling the professor teaching the class to emphasize the importance of grasping these concepts while studying physics.

The Classical Test Theory analysis helped us to detect the items or distractors that need revision. Some items needed little modifications to improve the distractors, while one item had a low ability to discriminate between high-archiving and low-archiving students probably due to the high item difficulty.

We used the chi-squared test to check the correlation among some factors. We discovered that students in their fourth year of high school tend to answer correctly to an item more often than students in their fifth year and that men tend to answer an item correctly more often than women.

Key-words: Orientation course, Classical Test Theory, chi-squared test, Physics

Abstract in italiano

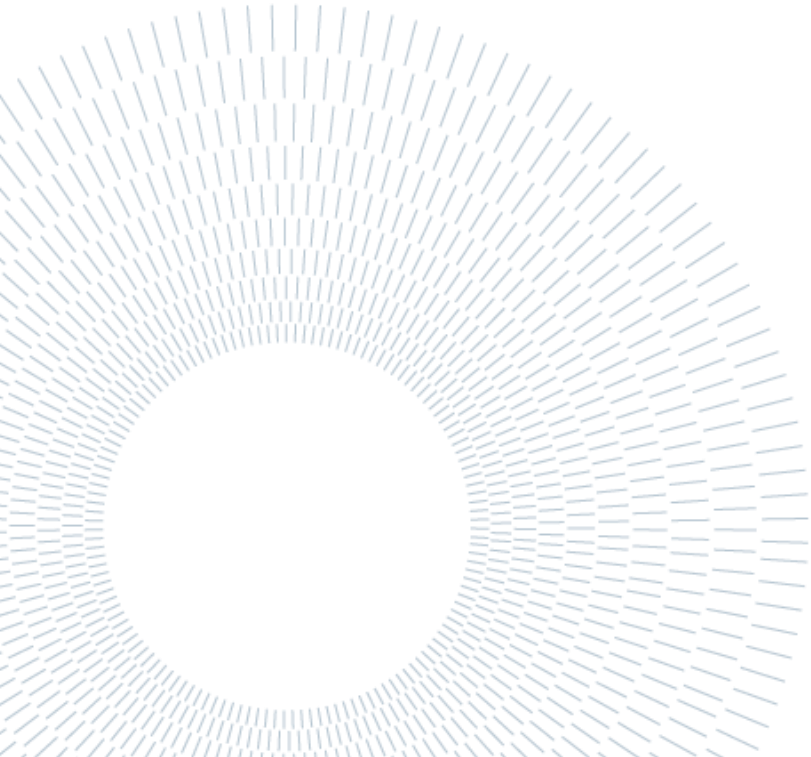
Questo lavoro di tesi analizza un test a risposta multipla composto da 8 quesiti di fisica. Il test è stato somministrato a degli studenti del quarto e quinto anno delle scuole superiori che hanno frequentato un corso di orientamento offerto dal Politecnico di Milano ed erano interessati a proseguire gli studi di ingegneria dopo la maturità.

Gli item avevano una struttura simile alle domande dell'esame di ammissione a ingegneria, consentendo agli studenti di valutare il livello di conoscenza richiesto per superare il test. Inoltre, il test serviva per affrontare alcuni misconcetti ed errori concettuali comuni tra gli studenti, consentendo al professore che insegnava in classe di sottolineare l'importanza di comprendere questi concetti durante lo studio della fisica.

L'analisi svolta utilizzando la Teoria Classica dei Test ci ha aiutato a rilevare gli elementi o i distrattori che necessitavano di revisione. Alcuni item necessitavano di piccole modifiche per migliorare i distrattori, mentre un item aveva una bassa capacità di discriminare tra studenti più e meno bravi, probabilmente a causa dell'elevata difficoltà dell'item.

Abbiamo utilizzato il test del chi quadrato per verificare la correlazione tra alcuni fattori. Abbiamo scoperto che gli studenti del quarto anno di scuola superiore tendono a rispondere correttamente a un item più spesso rispetto agli studenti del quinto anno e che gli uomini tendono a rispondere correttamente a un item più spesso delle donne.

Parole chiave: Corso di orientamento, Teoria Classica dei Test, test del chi quadrato, fisica



Contents

Introduction

The labor market is currently in high demand for specialists in science, technology, engineering, and mathematics (STEM). However, the number of graduates in STEM fields remains low. One of the contributing factors to this critical issue is the high dropout rates during the initial years of university. (Bozzi, 2021).

Educational guidance is considered one of the components needed to solve the problem because it helps build educational paths which develop student's needs and talents. Effective educational guidance also helps reduce early school dropout rates and bridges the gap between the skills required by employers and those acquired in schools. This, in turn, leads to a decrease in the number of NEETs (Not in Education, Employment, or Training) among individuals aged 15 to 29, contributing to an overall reduction in youth unemployment. Furthermore, it encourages students to continue learning throughout their lives(Magni, 2023).

This is why educational guidance is considered crucial, both by educators and political institutions. They believe that people require guidance to make decisions related to education, work, and life. When students' talents go unrecognized, they cannot fully develop, and as a result, they may fail to reach their maximum potential. (Magni, 2023).

Politecnico di Milano participates to a project called *Orientamento 2026-Orientamento attivo nella transizione scuola università* which was initiated by the Italian Ministry of University and Research. The project's objective is to offer orientation courses for students in their final three years of high school and provide guidance to help them prepare for the engineering entrance exam.

In addition to providing educational guidance, the courses offered by Politecnico di Milano included physics-related content and employed active teaching methods, which encouraged laboratory activities. These courses also featured a questionnaire designed to identify students' misconceptions and conceptual errors. Misconceptions refer to incorrect beliefs that students may hold about the physical world. Following the completion of these courses, the data was subjected to analysis using classical test theory, which aimed to pinpoint the most effective questionnaire items. Distractor evaluation was used to identify nonfunctioning distractors, and the chi-squared test was employed to derive insightful conclusions by exploring correlations between various factors.

The application of Classical Test Theory and distractor analysis was instrumental in identifying items in need of revision, ultimately leading to improvements in the test.

1 Orientation course structure

The Italian Ministry of University and Research promotes orientation courses for students in their last three years of high school, providing them with guidance on preparing for the engineering entrance exam. The courses had to last 15 hours each. Students who attended at least 70% of one course could get a participation certificate. Their schools were allowed to consider the classes as curricular or extracurricular activities and two-thirds of the classes had to be in presence. The courses promoted the participation of students with disabilities and learning difficulties and gender equality.

The orientation courses were designed to connect student's desires, skills in high demand in the job market, and education-related decisions. The goal is not to promote the university, but to:

- Getting a better understanding of the university environment, getting to know different options for personal growth;
- Learning about the scientific method through laboratory experience and active participation;
- Learning how to evaluate, check, and consolidate students' knowledge to reduce the difference between the skills owed and the ones needed at university;
- Getting to know the possible employment opportunities and their connection to the competencies learned during the course.

The courses could be conducted by high school and university professors, researchers, experts in the relevant course topics, as well as AFAM (*alta formazione artistica e musicale*) teachers. AFAM encompasses both public and private conservatories and academies of fine arts. The following are the regulatory references for the orientation courses:

- Ministerial Decree No 934 of the 3rd of August 2022;
- Directorial Decree No 1452 of the 22nd of September 2022.

Our 15-hour course was divided into three parts:

- 3 hours of presentation performed by IFOA, a training and consultancy center;
- 6 hours of math classes;
- 6 hours of physics classes.

The 6 hours of physics classes were divided into two sessions. The first session, lasting two hours, was conducted online, while the second session, lasting four hours, was held in person.

The first session was led by a physics university professor and focused on providing general advice to future students. In contrast, the second session was also led by the same university professor but with the assistance of three tutors. These tutors were all members of the ST2 research group, the Physics Education Research group at Politecnico di Milano.

During the first session, students were tasked with answering 8 multiple-choice questions that closely resembled those found on the engineering entrance test. In the second session, students were divided into groups of four and provided with a box for each group. Each box contained various everyday objects, and each group was assigned a different task. These tasks involved studying various physics phenomena or measuring physical quantities using the objects in their respective boxes. The instructions were intentionally broad, requiring students to either determine the solution independently or seek guidance from the tutors. Multiple potential solutions existed for each task, and some allowed for varying degrees of precision. Consequently, students could provide rough estimations before refining their measurements.

Following these activities, students were asked to complete a satisfaction questionnaire related to the physics component of the course. A total of 220 students registered for the orientation course. They were randomly divided into five groups, and the orientation course was conducted five times. Some of the 220 students registered but did not attend any classes, while others participated in only some of the sessions.

1.1. The first lesson of the physics orientation course

The first lesson of the physics part of the orientation course was held by a university professor and was about general advice needed by the students to get ready for the entrance exam. It covered the following topics:

- Advice for improving learning;
- Advice for moving from high school to university;
- The work of the ST2 lab, the laboratory that organized the physics part of the orientation course;
- The MOOCs (Massive Open Online Courses) made by Politecnico di Milano;
- Advice for learning physics.

During the first lesson of the physics orientation course, the students also had to answer a multiple-choice test made of 8 items regarding some specific physics topics.

These questions served two main purposes: firstly, to provide a rough assessment of the students' knowledge of physics, and secondly, to allow them to attempt questions similar to those found in the engineering entrance test. This enabled them to gain an understanding of the skills necessary to succeed on the test. Additionally, some of the questions were designed to address common misconceptions and conceptual errors prevalent among students, thereby increasing their awareness of these issues.

Students often hold a set of beliefs about the physical world based on their everyday experiences. These beliefs, known as misconceptions, may sometimes be incorrect or inconsistent with the formal concepts taught in classrooms.

Misconceptions have the potential to hinder students' learning outcomes by affecting the construction of their knowledge. Thus, one of the challenges for teachers is to help students overcome these misconceptions.

During this assessment, each student was required to answer each item individually, without assistance from peers or the professor. They were given a limited amount of time to answer each question.

The students completed the questionnaire using Socrative, an application designed for creating quizzes and collecting responses from participants. Socrative is among the audience response systems, commonly employed to enhance student engagement in the classroom. Many educators use this technology to foster active two-way communication and address issues related to students' concentration during lectures. By simulating one-to-one interactions, these systems facilitate prompt feedback. Such clicker systems have been shown to increase active learning, enjoyment, class participation, attendance, and retention.(Caldwell, 2007).

Using personal mobile phones, tablets or computers has been proven to be more enjoyable for the students and financially more sustainable than using clickers. Students prefer to use their devices because they are familiar with them (Ranieri et al., 2021).

Out of the 220 students who enrolled in the orientation course, only 113 actively participated in the first lesson and completed the questionnaire. These students were randomly divided into 5 groups, each of which had its lesson on a different day. The distribution of students in these groups was as follows: 21 students in the first group, 26 in the second, 19 in the third, 26 in the fourth, and 21 in the fifth. Of the 113 participating students, 89 were male, and 24 were female. Furthermore, 92 students were in their fourth year of high school, while 21 were in their fifth year.

The number of students coming from each high school is summarized in Figure 2, while the number of students coming from each type of high scholl is reported in Figure 3.

Figure 1: Number of students who attended the first physics lesson depending on the group

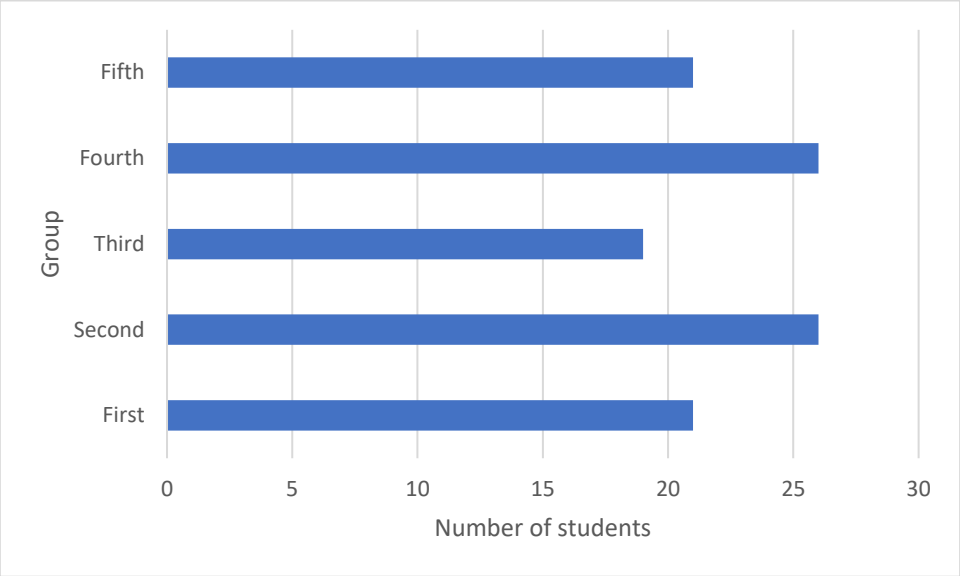


Figure 2: Number of students who attended the first physics lesson from each high school

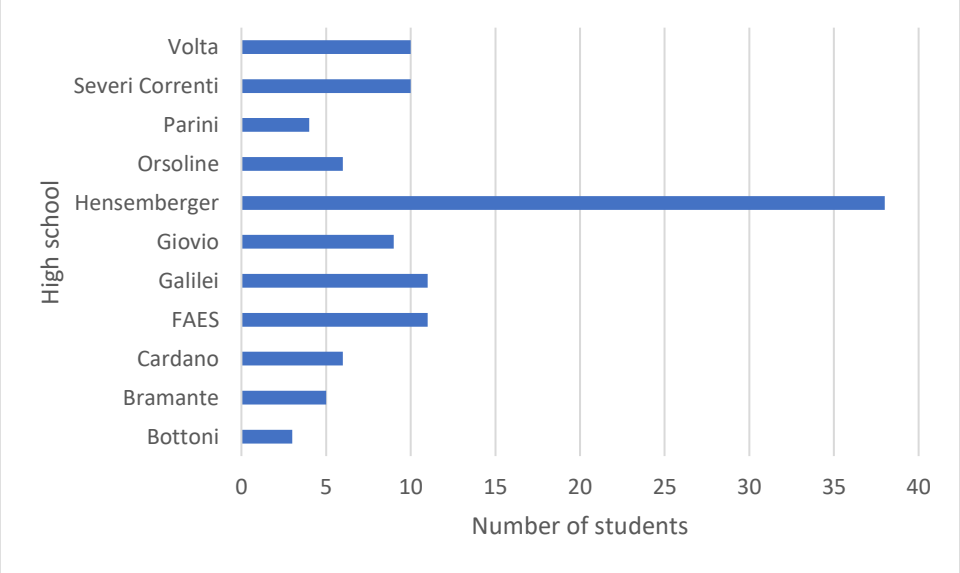
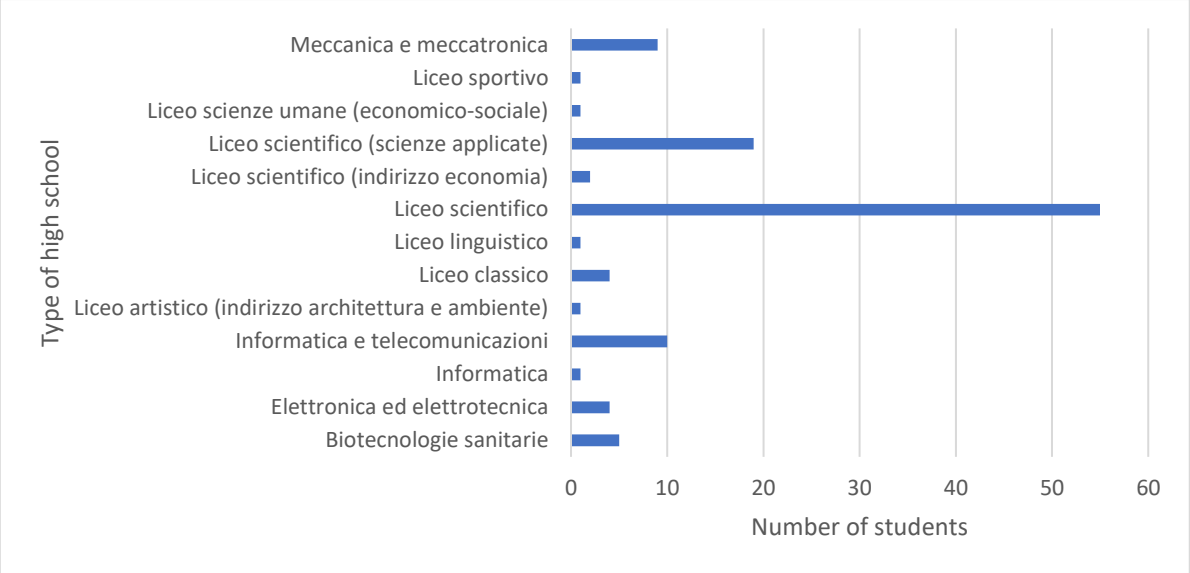


Figure 3: Number of students who attended the first physics lesson from each type of high school



1.1.1. Items of the questionnaire

The items of the questionnaire answered by the students during the orientation course were the following. The correct answers are bolded.

1. A boy throws a stone on the other side of a river to see whether the stone can get to the other side of the river or not. Which of the following pictures represents the forces acting on the stone (neglecting air resistance) when the stone is at the maximum height of its trajectory?

A. A
B. B
C. C
D. D

2. If you apply a force F_1=5N to an object on a horizontal plane, you notice that it does not move due to friction. If you apply a force F_2=10N, it still does not move. What is the intensity of the friction acting on the object?
   A. Greater than 10N
   B. 10N
   C. 5N in case you apply $\boldsymbol{F_1}$ and 10N in case you apply $\boldsymbol{F_2}$
   D. It is impossible to draw any conclusion using the available data

3. A body has a uniform circular motion of radius 1m with a speed of 5 m/s. Which of the following statements is correct?
   A. Since it is a circular uniform motion, the speed is constant and the acceleration is zero
   B. The tangential acceleration is zero and the centripetal acceleration is 25 $\boldsymbol{m/s^2}$
   C. The tangential acceleration is zero and the centripetal acceleration is 4 m/s^2
   D. None of the previous statements is correct

4. A body falls starting from a height H and a potential energy U. It takes time t to reach the height H=0, where it has kinetic energy K and zero potential energy. Which of the following statements is correct at time t/2?
   A. The kinetic energy is K/2
   B. The potential energy is U/4
   C. The kinetic energy is K/4
   D. The potential energy is U/2

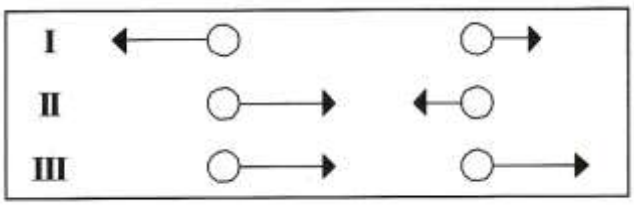
5. Two point charges with opposite signs are of 1nC each, they are at two opposite sides of a segment of length 2 m. What is the modulus of the electric field E and of the potential V in the middle of the segment?
   A. V=0 and E=0
   B. V=18V and E=18 V/m
   C. V=0 and E=18 V/m
   D. V=18V and E=0

6. The disk of the following picture rotates clockwise around an axis passing throw its center and perpendicular to the drawing. It makes 29 rounds per second. The disk is filmed using a movie camera which makes 30 frames per second. What is the motion of the black dot going to look like in the video?

A. It is going to move clockwise
B. It is going to move anticlockwise
C. It is going to move randomly
D. It is going to stay still

7. If you throw an object without applying any force, it falls with an acceleration of 9,8 m/s2, neglecting the air resistance. If you, instead, apply a force directed downwards, what is its acceleration right after you stop touching it neglecting the air resistance?
   A. Lower than 9,8 m/s^2
   B. Higher than 9,8 m/s^2
   C. Equal to 9,8 $\boldsymbol{m/s^2}$
   D. It depends on the initial velocity of the object

8. In the pictures I, II, and III below, you can find a couple of electrically charged spheres. The quantity and the sign of the charge on each sphere can be equal or different. The vectors represent the forces acting on the spheres. Considering the direction and the length of the vectors, indicate which pictures correspond to impossible situations.



A. Only III
B. Only I and II
C. Only II and III
D. All three of them: I, II, and II

Each of the items investigates student's knowledge about a different topic:

1. Very often students believe in the existence of two kinds of forces: impetus and active forces. Impetus is a force that keeps things moving. Impetus can be gained or lost in different ways. Every object is seen as a container that can store impetus. This misconception is evidence that the first law of dynamics is not understood (Hestenes et al., 1992). In the case of the first question, students might think that the stone stores impetus which decreases over time like fuel in a car.
2. This item focuses on students' comprehension of static friction. Frequently, students mistakenly believe that the force of static friction always equals the maximum static friction force.
3. It pertains to tangential and centripetal acceleration, the relationship between velocity and the intensity of velocity, and the tangential and normal components of acceleration.
4. It is based on the concepts of kinetic and potential energies and gravitation.
5. The question investigates the understanding of the superposition principle applied to electric fields and potentials.
6. It investigates the ability to conclude the case of different systems of reference.
7. It is based on the understanding of the second law of dynamics and the connection between force and acceleration, which is sometimes misunderstood by the students as between force and velocity (Hestenes et al., 1992).
8. It investigates the understanding of the action and reaction principle applied to the case of electrostatic interaction between point charges.

1.2. The second lesson of the physics orientation course

A qualitative investigation of students' perspectives on laboratory activities revealed that some students believed understanding the physics concepts related to a system was unnecessary when conducting experimental activities in a classroom involving that system. (Hu et al., 2017)

These students argued that the provided step-by-step instructions during the activities were sufficient to complete them without the need to grasp the equations and concepts. They felt it was possible to follow the instructions and finish the activities without a deeper understanding of the underlying principles.

On the other hand, students who emphasized the importance of comprehending the physical phenomena during the activities contended that equations guided experiments, and a deeper understanding was essential for high-quality work. They

believed that understanding made them more engaged in the experiment and helped them make sense of the procedure, thus making it easier to complete.

In our orientation course, we tackled the issue of students passively following instructions by adopting a different approach. We simply provided students with a final goal and asked them to find their solution, with the assistance of tutors but without step-by-step instructions.

Students were organized into groups of approximately four people, and each group was given a box containing everyday objects based on the task at hand. There were a total of 7 different experiments, each with multiple possible solutions.

For instance, one task involved estimating gravitational acceleration and the friction coefficient. The corresponding box contained small bricks, guides of various sizes and materials, paper tape, a meter, baking cups, and small balls. Students were permitted to use their smartphones to access internet information or use them as stopwatches or goniometers.

In this task, students could drop a ball from a known height and measure the time it took for the ball to reach the ground, allowing them to determine the gravitational constant. They could enhance the precision of the measurement by recording the ball's fall in slow motion. They could also repeat the experiment using baking cups to observe differences in motion due to friction.

Another task involved studying friction in fluids. The box included a transparent, long, thin vessel, transparent soap, small metallic balls of different sizes, magnets, paper tape, a funnel, rags, and baking cups. Students could investigate non-negligible air friction by throwing a baking cup and observing its descent. They could stick pieces of paper tape on the wall at equal distances, drop the baking cup, and record the fall in slow motion. They could use the video to study the cup's position over time.

To create a graduated cylinder, students could fill the vessel with soap, attach paper tape, and observe the balls moving slowly. This allowed them to measure the time it took for the balls to move from the top to the bottom of the vessel with precision. They could repeat the experiment using balls of different sizes and use magnets to extract the balls from the soap easily.

Out of the 220 students who registered for the course, only 106 attended the second physics class. These students were randomly divided into five groups before the first class, and this group division remained consistent for the second class. Consequently, the course was held five times in total, once for each group.22 students from the first group participated in the class, 25 from the second group, 13 from the third one, 24 from the fourth one, and, 22 from the fifth one. 85 students were male and 21 female, while 84 were doing their fourth year of high school and 22 were doing their fifth one.

The number of students coming from each high school is reported in Figure 5 and the number of students coming from each type of high school is summarized in Figure 6.

Figure 4: Number of students who attended the second physics lesson depending on the group

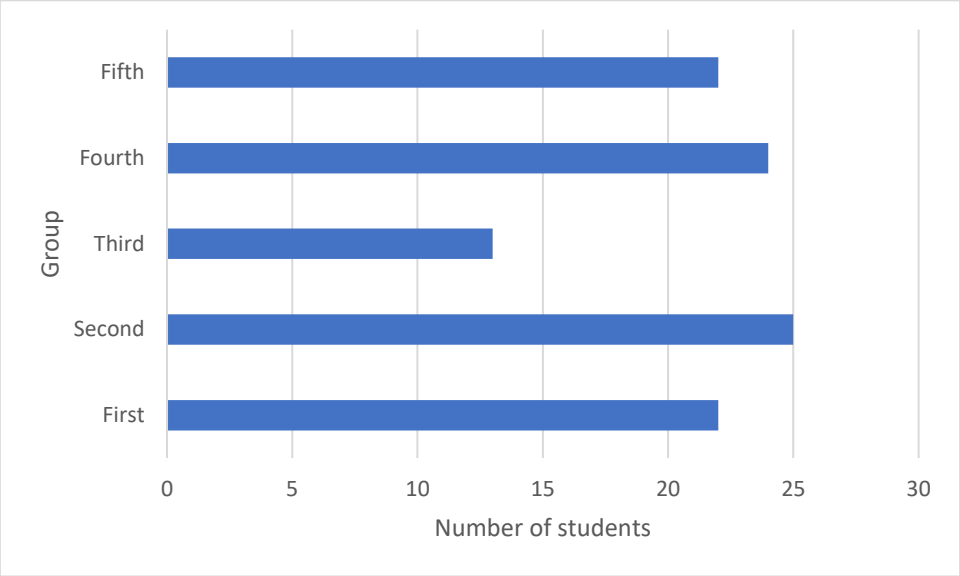


Figure 5: Number of students who attended the second physics lesson from each high school

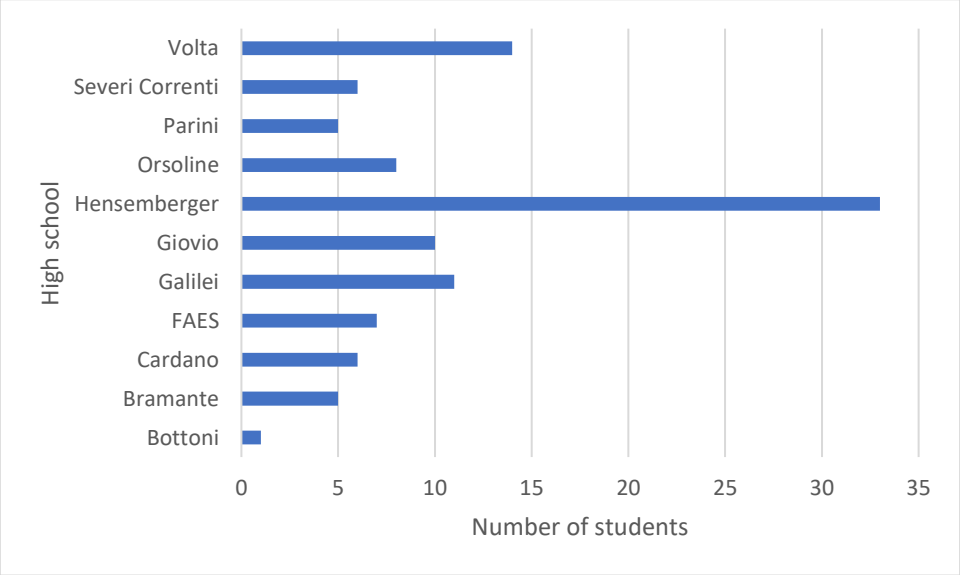
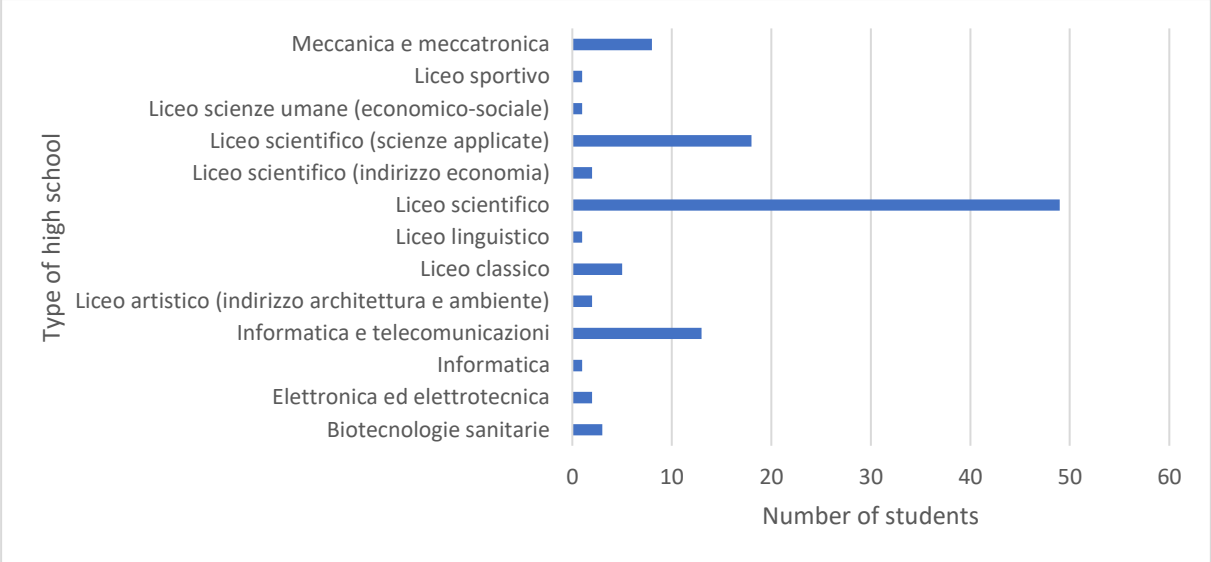


Figure 6: Number of students who attended the second physics lesson from each type of high school



2 Classical Test Theory

Item analysis involves a set of strategies used to select the best items from a pool of potential candidates. These items can be categorized into two groups: cognitive and noncognitive.

Cognitive items are designed to assess an individual's knowledge and measure cognitive constructs. Examples of cognitive items include true-false questions and completion tasks. These items are used to evaluate what a person knows.

Noncognitive items, on the other hand, are utilized to measure different subdimensions of a construct. A common example is a question with a Likert scale. Noncognitive items are used to gauge attitudes and opinions that cannot be adequately captured by a simple 'yes' or 'no' response. (Bandalos, 2018).

In our test, there are only cognitive items, so we will not focus on the description of the noncognitive ones. By using these tools, we can select the most suitable questions for our purposes (Bandalos, 2018).

2.1. Theoretical background of Classical Test Theory

Before delving into the calculations of the indexes defined in Classical Test Theory, we will present their definitions and theoretical interpretations.Classical Test Theory is instrumental in addressing questions such as 'Is an item difficult?' or 'Does an item effectively differentiate between students?' In practice, this theory involves the computation and interpretation of various indexes. In our analysis, we focused on item difficulty and discrimination, the point biserial coefficient, Kuder-Richardson reliability, and Ferguson's Delta, which will be clearly defined in the following section.

Subsequently, we conducted a distractors evaluation to assess the quality of response options for the questions. This evaluation serves as a valuable tool for scrutinizing item phrasing, as an ambiguously worded question or answer can lead students to make incorrect guesses.Since its understanding is based on the definition of discrimination coefficient and its definition does not require a long discussion, we discuss it only in chapter 2.1.2.

2.1.1. Item difficulty

Classical Test Theory introduces some indexes that are very useful to analyze a multiple-choice test. One of these indexes is the item difficulty (P) which is given by the ratio between the number of correct responses N_C and the total number of responses N (Ding & Beichner, 2009).

$$P = \frac{N_C}{N}$$

The higher P, the easier the item, which represents the percentage of correct responses. If you assign 1 point for each correct answer and zero points to each incorrect answer, as we did in our test, the item mean is equal to the item difficulty.

The item difficulty can be used to distinguish among items at different difficulty levels. The difficulty can be low (L), medium-low (ML), medium (M), medium-high (MH), or high (H). The link between the values of P and the labels is shown in the following table (Crocker & Algina, 1986).

Table 1: Meaning of the item difficulty

P	Difficulty
$1 \leq P < 0.75$	L
$0.75 \leq P < 0.50$	ML
$0.50 \leq P < 0.25$	M
$0.25 \leq P < 0$	MH
0	H

Usually, in standardized tests, the item difficulty for each question ranges between 0.3 and 0.7. Often just a few items have difficulty above or below this range. (Ding & Beichner, 2009)

Assuming that students' achievement levels follow a normal distribution, most students are typically situated within the middle range, with only a few students falling into the categories of high-achieving or low-achieving. Consequently, the questions should align with the knowledge of the majority of students and, therefore, should possess a difficulty index within the middle range. Another reason is connected to the variance of the total scores. The variance is given by

$$\sigma_i^2 = P_i(1 - P_i)$$

Where P_i is the difficulty of item i. You can derive that the variance is maximized for P_i equal to 0.5. This means that the best value of item difficulty to detect the differences among students is 0.5. If the items had all a very high or very low item difficulty, as you can calculate from the formula, the item variance would be very low and the test

would not be able to discriminate among students. In these cases, a revision of the items should be considered (Bandalos, 2018).

2.1.2. Discrimination coefficient

Another important index is item discrimination (D), which is based on the concept that low-achieving students are more likely to answer an item incorrectly, while high-achieving students are more likely to provide the correct answer. This index helps distinguish between these two groups.

At first, students are divided into an upper and lower group based on their total test scores. The two groups can correspond, for example, to the best and the worst 50%, 33%, and 25% of students. The discrimination coefficient is the proportion of students in the upper group who answered correctly P_u minus the proportion of students who answered correctly from the lower group P_l.

$$D = P_u - P_l$$

In our data analysis, we labeled the discrimination indexes corresponding to groups formed by the best and the worst 50%, 33%, and 25% with D50, D33, and D25. Instead, the proportion of students from the upper group who answered correctly an item is labeled with U50, U33, or U25 depending on the size of the upper group. In the same way, the proportion of students from the lower group who answered right is indicated by L50, L33, or L25. We chose these sizes for the groups because, according to the literature, these are the most commonly used (Bandalos, 2018; Ding & Beichner, 2009).

Item discrimination can be calculated both using an internal or an external criterion. In case you use an internal criterion, you use the procedure already described, instead, in case of an external criterion, you use the test results of another test to divide upper and lower groups. The internal criterion does not work, for instance, in case many items are miskeyed or total test scores are flawed (Bandalos, 2018). An example of an external criterion for the calculation of the discrimination coefficients is using the total test scores of the student's engineering entrance exam.

External criteria are often used for clinical evaluations and employment testing. Usually, tests built using an external criterion have fewer homogeneous items than the ones developed using the internal one. This is due to the lower homogeneity of the external criteria compared to the internal ones. For example, job testing requires the evaluation of many different skills such as job-specific abilities, time management, and social skills (Bandalos, 2018).

The discrimination index is considered acceptable if it is greater or equal to 0.3 (Ding & Beichner, 2009). Poor item discrimination might be caused by multiple reasons. One of them is that the item might be ambiguous or miskeyed. In this case, the problem can

be easily solved by correcting the item. Another option is that the item might have two correct answers.

The discrimination index can range from 1.0 to -1.0. Negative discrimination is a cause of concern because it indicates that low-archiving students tend to answer an item correctly more often than low-archiving students. A discrimination coefficient equal to zero, instead, indicates that an item is not able to discriminate between students. This happens in case an item is so difficult that nobody answers it correctly or, instead, is so easy that everyone gives the correct answer (Bandalos, 2018).

2.1.3. Point biserial coefficient, Kuder-Richardson coefficient, and Ferguson's Delta

The third index that we considered is the point biserial coefficient (r_{pbi}). It is a correlation coefficient that is used in the case of discrete dichotomies. A discrete dichotomy has no underlying continuum among categories. For instance, dead and alive are a discrete dichotomy because there are no options between being dead or alive. The point biserial coefficient is the correlation between the item scores and the test scores. It can be calculated as follows:

$$r_{pbi} = \frac{X_1 - X_0}{\sigma_x}\sqrt{P(1-P)}$$

X_1 is the average total score of the students who correctly answered the i-th item, while X_0 is the average total score of students who answered incorrectly the i-th item and σ_x is the standard deviation of the total scores. A low point biserial shows that an item does not test the same material as the others (Ding & Beichner, 2009). The point biserial coefficient should be greater or equal to 0,2 (Ding & Beichner, 2009).

The Kuder-Richardson coefficient and Ferguson's Delta are used to evaluate the entire test, instead of just focusing on items like the item difficulty, the discrimination coefficient, and the point biserial coefficient.

The Kuder-Richardson reliability (r_{test}) tells if the items measure the same ability. The higher the coefficient, the higher the correlation between the items. It is given by the following formula:

$$r_{test} = \frac{K}{K-1}\left(1 - \frac{\sum P_i(1-P_i)}{\sigma_x^2}\right)$$

K is the number of items in the entire test and P_i is the difficulty index of the i-th item (Ding & Beichner, 2009). The following table classifies the acceptability of the Kuder-Richardson coefficient based on its value (Doran, 1980).

Table 2: Meaning of Kuder-Richardson reliability

r_{test}	Meaning
$0.95 \leq r_{test} < 0.99$	Very high, rarely found
$0.90 \leq r_{test} < 0.95$	High, sufficient for the measurement of individuals
$0.80 \leq r_{test} < 0.90$	Fairly high, possible for the measurement of individuals
$0.70 \leq r_{test} < 0.80$	Okay, sufficient for group measurement, not individuals
$r_{test} < 0.70$	Low, useful only for group averages or surveys

Finally, Ferguson's Delta (δ) measures how well the final test scores are distributed over the possible range. It can be calculated as follows:

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - N^2/(K+1)}$$

Where f_i is the number of students with a score equal to i. An acceptable value of Ferguson's Delta has to be greater than 0,9 (Ding & Beichner, 2009).

2.2. Results from Classical Test Theory analysis

We calculated all the coefficients presented in the Classical Test Theory and used them to draw conclusions about the test to improve it.

2.2.1. Item difficulty

First, we calculated the item difficulty. It ranges from 0,19 to 0,58. Four items can be classified as medium difficulty (M), three as medium-high (MH), and one as medium-low (ML). The average item difficulty is 0,33. The results of the calculations of the difficulty indexes are reported in Table 3. Therefore, we can conclude that the test overall has the right difficulty.

Table 3: Item difficulties

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
P	0,31	0,38	0,19	0,27	0,25	0,50	0,58	0,19
difficulty	M	M	MH	M	MH	M	ML	MH

As proved in the theoretical part of Classical Test Theory, too easy or too difficult items do not discriminate between students. They can be included in tests with many questions to discriminate between very high and very low archiving students depending on the goal of the test. Since there are not any very complicated or simple questions and there are only 8 questions, there are no items that should be replaced based on considerations about the difficulty index.

The only two questions on electromagnetism (numbers 5 and 8) have a medium-high difficulty, while the other questions on mechanics have a medium-low, medium or medium-high difficulty. So, overall, in the test, the questions on electromagnetism are more difficult than the items on mechanics. Item number 8 is also based on the understanding of the action and reaction principle, so it might be difficult both because of the application of the concepts of electromagnetism and because of the action and reaction principle.

2.2.2. Discrimination coefficient

Both the discrimination index and the point biserial coefficient analyze the ability of the test to discriminate among students. The item discrimination index was calculated considering the top and bottom quartiles, the top and the bottom 33% of students, and the top and bottom half of students.

There are 31 students with a total score of 3. In case we consider groups made of 50% of students, 23 of them are part of the upper group and 7 of the lower group. In this case, the total number of possible divisions in the upper and lower group i is given by the formula of permutations with repetition.

$$i = \frac{31!}{23! * 7!} = 6{,}31 * 10^7$$

Instead, if we calculate D33, 4 of the students with a total score of 3 are in the upper group and zero in the lower group. The lower group is made up of only students with a total score of 2, 1, and 0. However, there are 17 students with a total score of 2 and only 6 of them are in the lower group. In this case, i is given by

$$i = \frac{31!}{4! * (31-4)!} * \frac{17!}{6! * (17-6)!} = 3{,}89 * 10^8$$

If we consider groups made of 25% of students, there are no students with a total score of 3 in the upper and lower groups. The upper group is made of 18 of the 32 students with total scores equal to 4 and all the students with total scores equal to 5, 6, and 7 because there are no students with a total score equal to 8. Instead, the lower group is made of 24 out of the 27 students with a total score equal to 1 and all the students with a total score equal to zero. In this case, we can calculate i as

$$i = \frac{32!}{18! * (32-18)!} * \frac{27!}{24! * (27-24)!} = 1{,}38 * 10^{12}$$

Since in all three cases, the number of possible divisions in upper and lower groups is very high, we can not calculate the discrimination coefficients for each case and then calculate the average. That is the reason why we chose to perform the calculation in only 30 cases. 30 is a number that is large enough to produce statistically relevant results and at the same time, i slow enough to make the computation fisible. The table below reports the maximum and minimum values obtained from all the calculations.

Table 4: Maximum and minimum values of the discrimination index obtained by repeating the calculation 30 times

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Min D25	0,39	0,46	0,39	0,50	0,21	0,43	0,46	0,21
Max D25	0,57	0,64	0,54	0,64	0,43	0,57	0,64	0,32
Min D33	0,32	0,43	0,38	0,46	0,30	0,32	0,43	0,22
Max D33	0,46	0,54	0,46	0,57	0,38	0,43	0,54	0,30
Min D50	0,24	0,23	0,23	0,26	0,17	0,13	0,34	0,10
Max D50	0,37	0,38	0,33	0,44	0,32	0,29	0,45	0,28

There is variability in the calculation of the discrimination indexes. For some questions sometimes they are acceptable, sometimes not. For example, item number 5 in one of the calculations had a discrimination index corresponding to 25% of 0,21, which is under the threshold, while in another calculation of 0,43, which is over the threshold of 0,3. For that reason, we calculated the averages and standard deviations and reported the results in Table 5. The cells with averages lower than 0,3 are marked in red.

Table 5: Averages (ave) and standard deviations (SD) of the discrimination coefficients.

		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
ave	**D25**	0,48	0,55	0,48	0,58	0,31	0,50	0,54	0,28
ave	**D33**	0,41	0,47	0,41	0,52	0,32	0,39	0,50	0,25
ave	**D50**	0,31	0,32	0,29	0,36	0,25	0,22	0,39	0,19
SD	**D25**	0,04	0,04	0,04	0,03	0,04	0,04	0,04	0,03
SD	**D33**	0,03	0,03	0,02	0,03	0,02	0,03	0,03	0,02
SD	**D50**	0,04	0,04	0,03	0,04	0,03	0,04	0,03	0,04

Overall the discrimination indexes have acceptable values except in item number 8, which has a discrimination index slightly below the threshold of 0.3 even when

considering groups containing 50% of students. All the other questions have better discrimination indexes.

When we look at the results of the calculation of the discrimination indexes, we notice that they are all positive, the discrimination indexes corresponding to 25% of students are higher than the ones corresponding to 33% and 50%, and the indexes corresponding to the 33% is higher than the one calculated using the top and bottom quartiles. This is because when you perform the calculations using the best and worst 25% or 33% of students, you neglect the performance of the students with a total score of around average. These students often have similar scores and might end up in different groups when you consider groups of 50% of students (Ding et al., 2006). In our case, this effect is not negligible at all. Since the test is made of only 8 questions, many students have the same final score. So, when we make the group division, many students end up in the upper or lower group based only on chance. Therefore, considering smaller upper and lower groups leads to higher discrimination indexes, however, this approach neglects the performance of the students with an average total score.

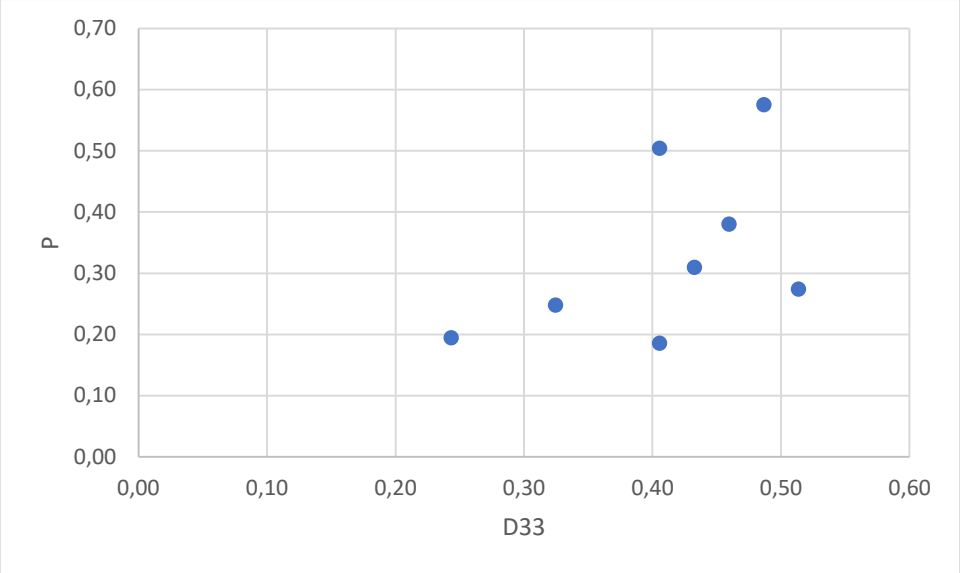
In summary, when forming two groups, we should consider both of the following factors:

- Smaller groups avoid the possibility that students with the same total score end up in different groups
- Small groups neglect the performance of students with a score around average

Groups composed of 33% of students strike a balance between these two considerations. In contrast, relying solely on quartiles neglects the performance of half of the students taking the test, while placing all students in a group amplifies the impact of the random division of students with scores around the average.

The following graph shows the relationship between item difficulty and the discrimination index calculated using groups formed by 33% of students each.

Figure 7: Dependence of the difficulty index from the item discrimination calculated using groups made of 33% of students



2.2.3. Point biserial coefficient, Kuder-Richardson coefficient, and Ferguson's Delta

The point biserial coefficients are all greater than 0.2, so the items have good reliability.

Table 6: Point biserial coefficients

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
rpbi	0,45	0,45	0,51	0,52	0,30	0,37	0,44	0,27

The Kuder-Richardson index is 0,31. The index is usually considered acceptable above 0,8, but when it comes to tests made of very specific questions about a very diverse range of topics, the values of the Kuder-Richardson index are usually low. In addition, since the test consists of only 8 items, it is not designed to comprehensively assess a student's knowledge across various physics topics. This limitation is intentional because the primary purpose of the test is not to evaluate the depth of a student's understanding. Instead, it aims to reveal potential misconceptions and conceptual errors that students may hold in physics, often without their awareness.

Due to the limited number of questions and the fact that they address disparate topics, it is expected that the test will exhibit low internal consistency. Consequently, the Kuder-Richardson coefficient, which measures the test's reliability, is likely to be low in this case. (Taber, 2018).

Ferguson's delta is 0,90 which corresponds exactly to the minimum acceptable value. So the tests discriminate the students well enough.

2.3. Distractors evaluation

The goal of the following analysis is to identify nonfunctioning distractors. The non-functioning distractors are incorrect answers to questions that have been chosen by less than 5% of students, or that have a positive discrimination coefficient. In other words, a distractor should represent a plausible alternative to the correct answer and, at the same time, attract more students from the lower group than those from the upper group (Quaigrain & Arhin, 2017).

Creating good distractors can be a challenging task since they should both sound reasonable to some students and not be misleading at the same time. The Tables 7 shows the percentage of students who chose each answer (tot), the total number of students in the upper or lower group made up 33% of the total (U33 and L33), and the corresponding discrimination coefficient D33. The cells corresponding to the correct answer are highlighted in grey, while the cells corresponding to nonfunctioning distractors are highlighted in red.

Tables 7: Percentage of students who choose each option (tot), U33, L33, and D33

Q1	A	B	C	D
tot	31,0	9,7	48,7	8,8
U33	19	0	17	1
L33	4	9	19	4
D33	0,4	-0,2	-0,1	-0,1

Q2	A	B	C	D
tot	41,6	2,7	38,1	16,8
U33	9	0	23	5
L33	23	0	6	8
D33	-0,4	0,0	0,5	-0,1

Q3	A	B	C	D
tot	62,8	18,6	2,7	15,9
U33	16	17	3	1
L33	31	1	0	6
D33	-0,4	0,4	0,1	-0,1

Q4	A	B	C	D
tot	19,5	12,4	27,4	39,8
U33	6	2	19	10
L33	10	8	0	19
D33	-0,1	-0,2	0,5	-0,2

Q5	A	B	C	D
tot	32,7	20,4	24,8	19,5
U33	5	9	17	5
L33	13	7	5	11
D33	-0,2	0,1	0,3	-0,2

Q6	A	B	C	D
tot	21,2	50,4	8,8	19,5
U33	3	26	1	7
L33	14	12	5	7
D33	-0,3	0,4	-0,1	0,0

Q7	A	B	C	D
tot	5,3	33,6	57,5	2,7
U33	1	6	29	0
L33	3	23	10	2
D33	-0,1	-0,5	0,5	-0,1

Q8	A	B	C	D
tot	65,5	8,0	6,2	19,5
U33	21	3	2	10
L33	29	5	3	1
D33	-0,2	-0,1	0,0	0,2

The distractors that have been chosen by less than 5% of students are:

- Answer B of question 2
- Answer C of question 3
- Answer D of question 7

The distractors with positive discrimination coefficients are:

- Answer C of question 3, which has also been chosen by less than 5% of students
- Answer B of question 5

So, there are 4 nonfunctioning distractors out of 24. There are 0 or 1 nonfunctioning distractors per item. This means that future revisions might involve just some distractors rather than the whole items.

The distractors analysis is useful also to detect some possible issues with the items, for example, if many students from the upper group consistently choose a wrong answer, it might be an indicator of an ambiguous item that should be rewritten. Not well-written items tend to affect more the performance of high-archiving students than one of the low-archiving ones. Or, for example, if there is no clear pattern, students might have answered randomly because the item is tricky or too difficult.

The Tables 8 indicate the number of students who chose each answer from the upper and lower groups. The correct answer is highlighted in grey.

Tables 8: It indicates the total number of students from the upper and lower groups who chose each option

Q1	A	B	C	D
U50	25	2	25	4
L50	9	9	30	6
U33	19	0	17	1
L33	4	9	19	4
U25	17	0	10	1
L25	2	7	14	4

Q2	A	B	C	D
U50	16	2	32	6
L50	30	1	11	13
U33	9	0	23	5
L33	23	0	6	8
U25	7	0	20	1
L25	18	0	5	5

Q3	A	B	C	D
U50	26	19	3	8
L50	44	2	0	10
U33	16	17	3	1
L33	31	1	0	6
U25	10	14	3	1
L25	24	0	0	5

Q4	A	B	C	D
U50	8	5	25	18
L50	14	9	6	26
U33	6	2	19	10
L33	10	8	0	19
U25	5	2	16	5
L25	7	7	0	14

Q5	A	B	C	D
U50	12	13	21	9
L50	24	10	7	13
U33	5	9	17	5
L33	13	7	5	11
U25	4	7	13	3
L25	9	6	5	8

Q6	A	B	C	D
U50	7	35	3	11
L50	17	21	7	11
U33	3	26	1	7
L33	14	12	5	7
U25	2	19	1	6
L25	13	6	4	6

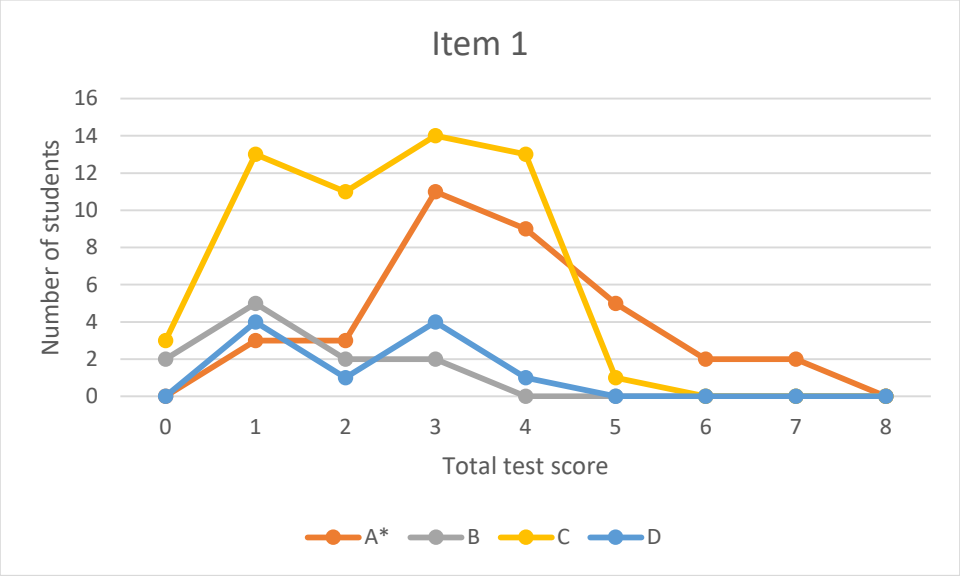
Q7	A	B	C	D
U50	2	8	45	0
L50	4	30	19	3
U33	1	6	29	0
L33	3	23	10	2
U25	1	6	20	0
L25	2	19	6	2

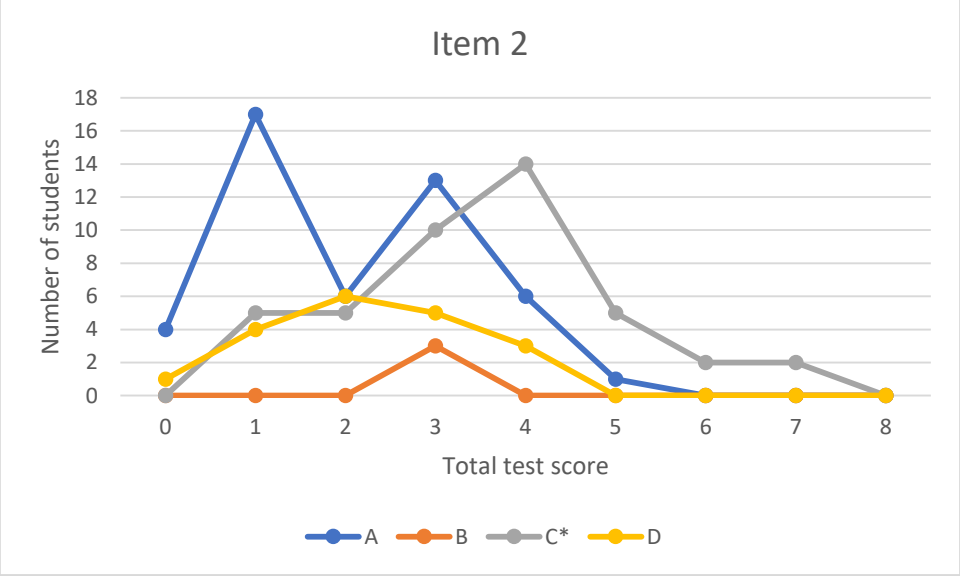
Q8	A	B	C	D
U50	34	4	2	15
L50	39	5	5	7
U33	21	3	2	10
L33	29	5	3	1
U25	15	2	1	9
L25	22	4	3	0

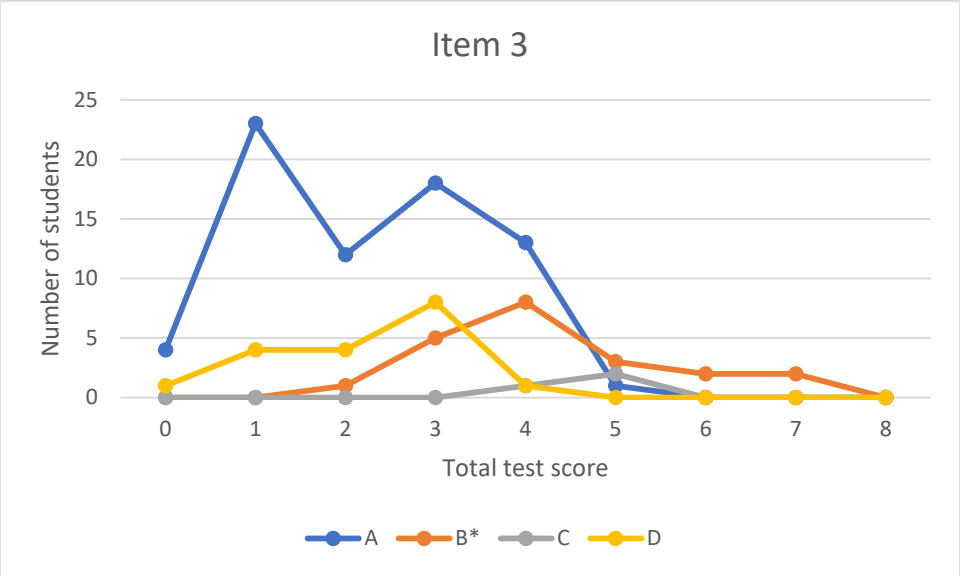
Many students chose the answer C instead of A in the first question. Answer A was chosen by many students for the third question, even among those in the top quartile.

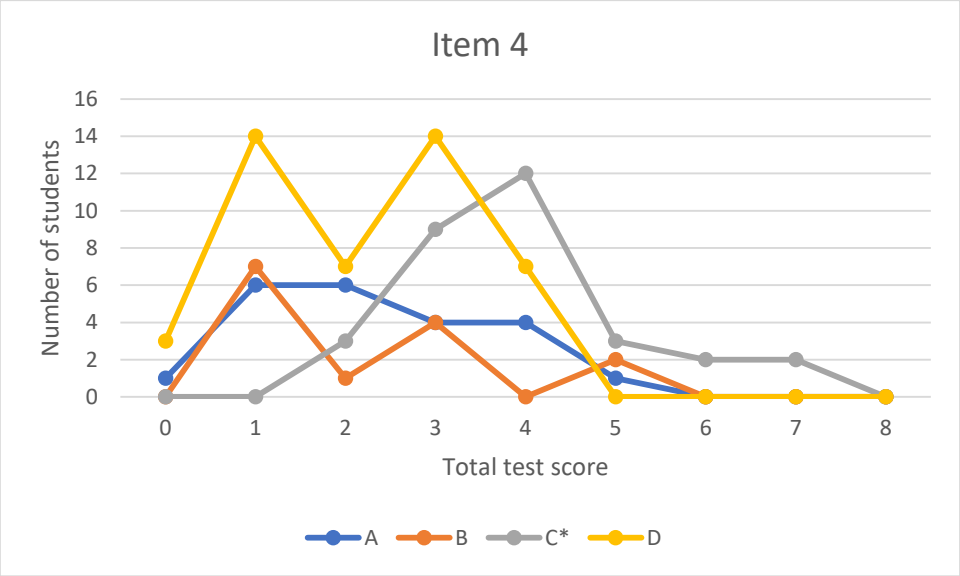
The response curves show the number of students who chose answers A, B, C, or D as a function of their total test scores. They have been plotted for each item of the test. The correct answer is indicated by the symbol *.

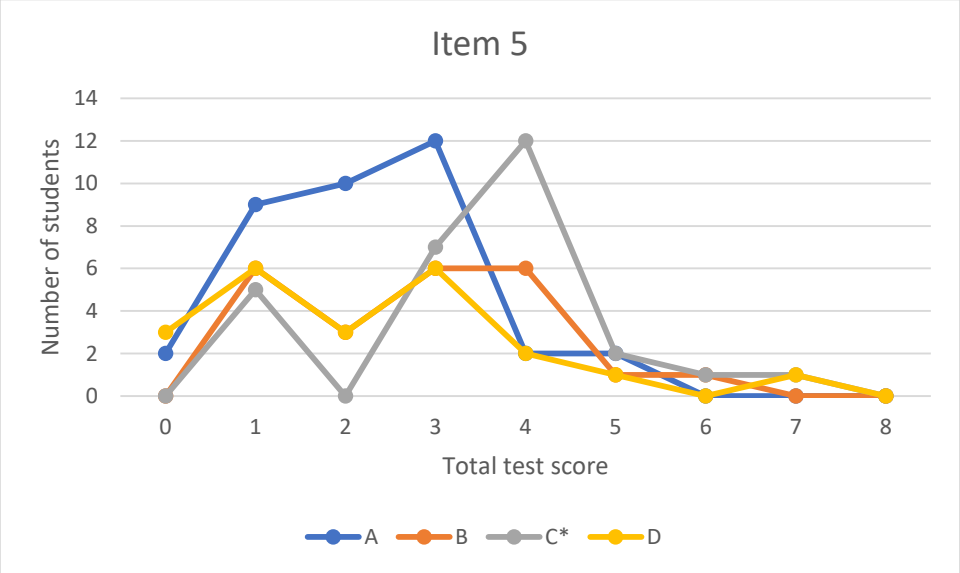
Figures 8: Response curves for all the items

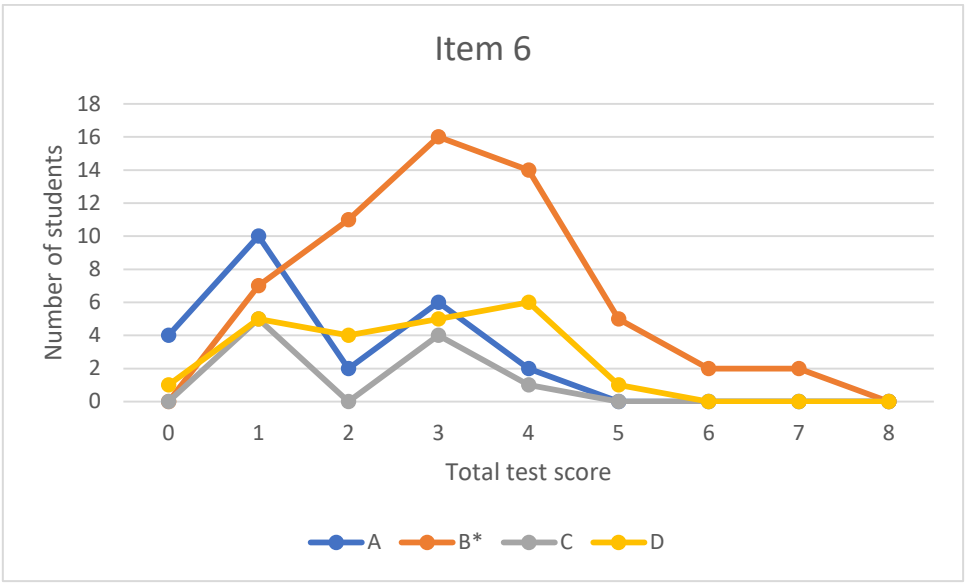


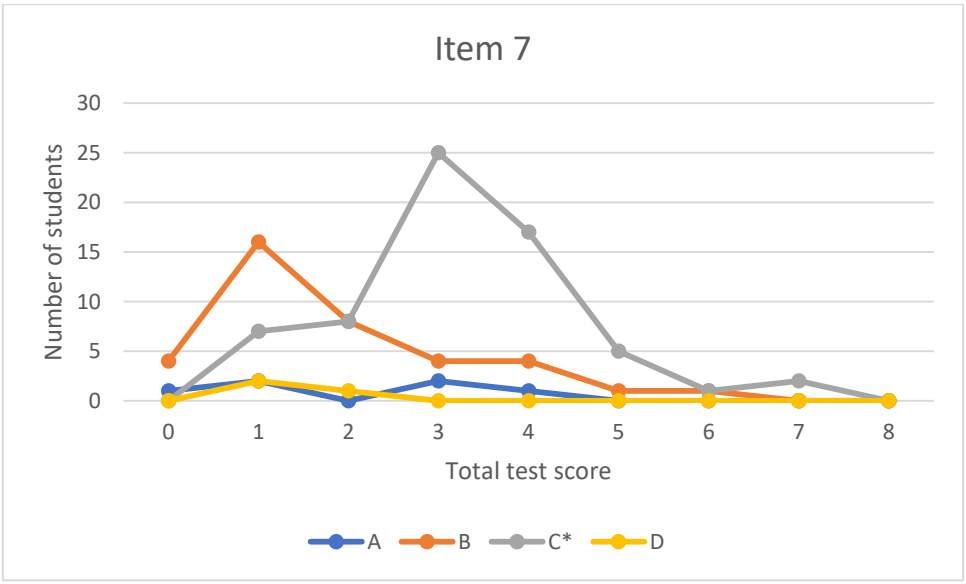


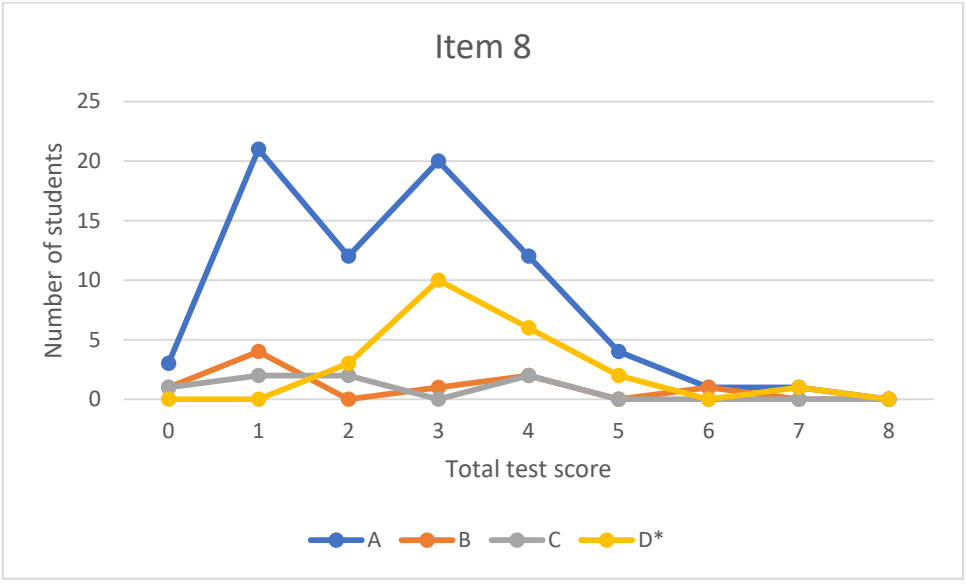

Item 3
Number of students
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8
Total test score
A B* C D


Item 4
Number of students
16
14
12
10
8
6
4
2
0
0 1 2 3 4 5 6 7 8
Total test score
A B C* D


Item 5
Number of students
14
12
10
8
6
4
2
0
0 1 2 3 4 5 6 7 8
Total test score
A B C* D


Item 6
Number of students
18
16
14
12
10
8
6
4
2
0
0 1 2 3 4 5 6 7 8
Total test score
A B* C D


Item 7
Number of students
30
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8
Total test score
A B C* D


Item 8
Number of students
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8
Total test score
A B C D*

In case all the distractors of an item are well functioning, for low total test scores, the curve corresponding to the distractors should be above the curve corresponding to the correct answer, while for high test scores, the curve of the right answer should be above the others.

In our test, the items that do not follow this behavior are:

- Item number 5. The non-ideal behavior might be caused by having a low number of students with a high total test score
- Item number 8. Many students chose option A instead of the right answer D even among the best students. The item also has a high difficulty and a low discrimination coefficient which might be causing the behaviour of the curves. So the item might need a revision

2.4. Classical Test Theory numerical results summary

All the coefficients from Classical Test Theory related to each item can be summarized in Table 9. It contains the values and classification of the item difficulties, the averages of the discrimination coefficients calculated considering different upper and lower group sizes, the point biserial coefficients and the number of non functioning distractors per item n.

Table 9: Item difficulties, discrimination coefficients, and point biserial coefficients

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
P	0,31	0,38	0,19	0,27	0,25	0,5	0,58	0,19
difficulty	M	M	MH	M	MH	M	ML	MH
D25	0,48	0,55	0,48	0,58	0,31	0,5	0,54	0,28
D33	0,41	0,47	0,41	0,52	0,32	0,39	0,5	0,25
D50	0,31	0,32	0,29	0,36	0,25	0,22	0,39	0,19
rpbi	0,45	0,45	0,51	0,52	0,3	0,37	0,44	0,27
n	0	1	1	0	1	0	1	0

3 χ^2 test

The chi-squared test is a method used to compare frequencies and proportions. Its primary application is in assessing the goodness of fit, which means understanding if an observed distribution is statistically different from an expected or theoretical distribution based on mathematical, biological, or physical laws. In this case, the test is used to tell whether the observed frequencies or proportions are similar to the expected ones, or they are so different that the differences cannot be due to statistical fluctuations. The test can be also used in the case of contingency tables to evaluate the independence between two factors (Soliani, 2015). We used the test to find out if the answers were correlated to factors such as grade or gender.

3.1. χ^2 test theory

In case we apply the chi-squared test to contingency tables, the null hypothesis states that the two factors are independent and their differences are due only to statistical fluctuations. Instead, the alternative hypothesis is that the two factors are dependent, so the null hypothesis can be rejected.

When we want to study the dependency or independence of two factors, first, we build a table with the observed frequencies n_{ij} and calculate the sum of the frequencies of each row and each column, and the total number of frequencies. Then we build the table of the expected frequencies $\hat{n}_{ij}$. Each element is given by

$$\hat{n}_{ij} = \frac{f_i * f_j}{N}$$

Where f_i is the sum of the frequencies of the ith row, f_j is the sum of the frequencies of the j-th column and N is the total number of frequencies (Soliani, 2015).

The χ^2 is given by

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Where R and C are the numbers of rows and columns of the contingency tables before calculating the sum of the frequencies (Soliani, 2015).

The number of degrees of freedom of the table is given by

$$(R-1)*(C-1)$$

In the case of a 2×2 contingency table, the number of degrees of freedom is 1.

Given the number of degrees of freedom and the probability of making the first type error α (probability of wrongly rejecting the null hypothesis), you can obtain the critical value and compare it to χ^2 (Soliani, 2015).

Cochran's rule states that the chi-squared test is valid if in each cell of the contingency table

- No expected frequency is lower than 1
- At least 80% of the expected frequencies are equal to or higher than 5

When we have too many low frequencies, we can reduce the number of categories.

In the case of a 2×2 contingency table, each expected frequency cannot be lower than 5, and the total number of observations N has to be high enough. If N is higher than about 100 or 200, the test is considered reliable. Instead, if it is smaller but above 30, we need the so-called Yates's correction for continuity. If it is smaller than 30, the test is not applicable (Soliani, 2015).

The power of a statistical test is defined as the probability of rejecting a null hypothesis when it is false. It increases with the total number of observations.

The value of the chi-squared depends on:

- The difference between the observed frequencies and the expected ones;
- The total number of observations T;
- The size of the contingency table.

When we investigate an effect, we want to know its magnitude. Statistical test results are considered significant if they meet some statistical standards. However statistical significance has a different meaning from the word significance used in an everyday language. A statistically significant result is very likely not due to chance, while a practically significant result is meaningful in real life. Sometimes a statistically significant result is trivial and sometimes a statistically nonsignificant result is important (Ellis, 2010).

Therefore, in a study, it is very important to introduce the effect sizes to determine the meaningfulness of a result. The effect size is the magnitude of a result and can be expressed in many different ways. Two of them are the Cramer's V and the odds ratio.

So we need some methods that can determine which part of data contributes the most to the goodness of a fit. One of them is the Cramer's V. It is defined as

$$V = \sqrt{\frac{\chi^2}{T*(k-1)}}$$

Where k is the smaller number between R and C (Soliani, 2015).

Cramer suggested his formula because the maximum value of χ^2 is given by

$$\chi^2_{max} = T * (k - 1)$$

So, the maximum value of V is 1 (Soliani, 2015). The size effect can be classified based on the value of V following Table 10 (Rea & Parker, 2014).

Table 10: Classification of Cramer's V values

V	classification
0≤V<0,1	Negligible
0, 1 ≤V<0,2	Weak
0, 3 ≤V<0,4	Moderate
0,4≤V<0,6	Relatively strong
0,6≤V<0,8	Strong
0,8≤V≤1	Very strong

Another method is the odds ratio which is defined starting from the odds. Let's assume that p is the probability that an event A happens. Odds is defined as

$$odds = \frac{p}{1-p}$$

So, the odds is the ratio between the probability that an event A happens and the probability that A does not happen (Espa & Micciolo, 2012). The odds ratio is the ratio between two odds. In the case of a 2×2 table is given by

$$OR = \frac{ad}{bc}$$

Where a, b, c, and d are the elements of the contingency table indicated in the example below (Soliani, 2015). In the case of a 2×C table, you can calculate multiple odds ratios by choosing 2 different columns every time (Espa & Micciolo, 2012).

Table 11: Example of contingency table

	Factor 2	
Factor 1	**C**	**D**
A	a	b
B	c	d

The effect size can be classified based also on the value of the odds ratio. The classification is reported in Table 12 (Rosenthal, 1996).

Table 12: Classification of odds ratio values. The indicated values are just qualitative.

OR	classification
1,5	Weak
2,5	Moderate
4	Strong
10	Very strong

3.2. Results of the χ^2 test

Using the chi-squared test, for each item we compared the frequencies of

1. Grade and correctness of the answer
2. Gender and correctness of the answer
3. Gender and distractor
4. Type of high school and correctness of the answer

The orientation course was designed for students doing their fourth and fifth years of high school. So when we performed the test considering the grade, we considered only these two years.

When we tested the answer, we considered that all the right answers were in one group and all the wrong answers were in another group.

When we performed test number 4, we considered only two types of high schools: *liceo scientifico* and *liceo scientifico opzione scienze applicate.*

The students who did not answer an item were not considered in the analysis.

3.2.1. Grade and correctness of the answer

All the students who participated in the orientation courses were doing their fourth or fifth year of high school. The students doing their fifth year should have learned slightly more advanced physics topics than the students doing their fourth year, but at the same time, they might remember less about the topics covered during the first years of high school. So it is interesting to compare their results on the test.

The total number of students who answered the multiple-choice test was 113. Out of these 113 students, 55 did *liceo scientifico* and 19 *liceo scientifico opzione scienze applicate*. So students doing these two types of high schools are 65.5% of the total number of students. There were 10 students studing *informatica e telecomunicazioni* and 9 *meccanica e meccanotrica*. So the students of *Informatica e telecomunicazioni* make 8,8% of the total and the ones of *meccanica e meccanotrica* the 8,0%. The number of students doing the other types of high schools was lower or equal to 5.

The topics covered in each year of high school during physics classes by *liceo scientifico* and *liceo scientifico opzione scienze applicate* students are the same. During the last year of high school, they learn about magnetism and some of the physics discoveries of the XX century. The topics that have to be covered are described in the Ministerial Decree No 211 of the 7th of October 2010 *Indicazioni Nazionali*, annex F.

We compared the frequencies of the grade and correctness of the answer and obtained the following results reported in Table 13.

Table 13: Chi-squared test results considering the grade and the correctness of the answer (test number 1)

item	p-value	V	OR
1	0,17	0,18	2,23
2	0,33	0,09	0,63
3	0,95	0,00	0,96
4	0,04	0,40	4,44
5	0,61	0,03	0,75
6	0,84	0,00	0,91
7	0,37	0,07	0,64
8	0,20	0,16	2,68

Item number 4 has a p-value lower than 0,05, so the factors are probably not independent. In all the other items grade and correctness of the answer are independent. Table 14 reports the contingency table of the observed frequencies of item number 4. p is the probability of having observed a correct answer in each of the two grades considered. We can conclude that, according to the test, students doing their fourth year of high school tend to answer item number 4 more frequently than students doing their fifth year.

Table 14: Observed frequencies in case of test 1 item number 4

	grade	
answer	4	5
1	29	2
0	62	19
p	0,32	0,10

Item number 4 is based on the knowledge of kinetic and potential energies. These topics in *liceo scientifico* and *liceo scientifico opzione scienze applicate* shoud be learned at the end of the second year of high school. Students doing *informatica e telecomunicazioni* and *meccanica e meccatronica* should learn these topics at the beginning of the second

year. Since students of *liceo scientifico*, *liceo scientifico opzione scenze applicate*, *informatica e telecomunicazioni* and *meccanica e meccatronica* make together the 82,3% of the total, most of the students who took the test should have learned them in the second year.

3.2.2. Gender and correctness of the answer

Over the past century, the relevance of Women's Rights in international policies has grown significantly, resulting in numerous advancements. However, despite the progress made, there remain pressing issues that demand attention. One such concern is the need to enhance women's participation in STEM fields. There still exists a persistent misconception that certain areas, such as STEM, should be exclusive to one gender. (Montecinos & Anguita, 2015). Women have 57% of all bachelor's degrees in the U.S., however, they are underrepresented in most of the STEM fields (Koch et al., 2022). This idea led to a small participation of women in fields like physics research and education and reinforced the belief that physics is a male subject (Montecinos & Anguita, 2015).

Because of the strong interest shown by research to investigate the gender gap and the governments being willing to take action to reduce it, it was interesting to investigate if there were any gender differences in the answers to the multiple-choice test.

We performed a chi-squared test considering the gender and the correctness of the answer. The results are reported in Table 15.

Table 15: Chi-squared test results considering gender and the correctness of the answer (test number 2)

item	p-value	V	OR
1	0,20	0,15	2,00
2	0,71	0,01	0,84
3	0,79	0,01	1,18
4	0,74	0,01	0,84
5	0,05	0,37	4,19
6	0,33	0,09	1,57
7	0,37	0,08	1,51
8	0,32	0,09	1,93

Gender and correctness of the answer seem independent except for item number 5 which has a p-value of 0,05. The observed frequencies of item number 5 are reported in Table 16. According to the test, men tend to answer correctly item number 5 more frequently than women.

Table 16: Table of the observed frequencies in case of test 2 item number 5

	gender	
answer	M	F
1	26	2
0	62	20
p	0,30	0,09

30% of men answered correctly to item number 5, while only 9% of women gave the right answer. We also used the chi-squared test to check if there are any distractors that tend to be chosen mainly by men or women.

3.2.3. Gender and distractor

In test number 2, we tested the independence of gender and correctness of the answers and we found that men answered to item number 5 more frequently than women. So we tested the independence of gender and distractors to find out if men and women tend to choose different distractors. If this is the case, we might have a possible explanation of the results of test number 2. The results are reported in Table 17.

Table 17: Chi-squared test results considering gender and distractor (test number 3)

item	p-value	V
1	0,47	0,18
2	0,53	0,15
3	0,77	0,06
4	0,56	0,13
5	0,31	0,26
6	0,34	0,29
7	0,30	0,35
8	0,52	0,14

There is no p-value below 0,05, so we cannot reject the null hypothesis stating that gender and distractor are independent. The test showed that no distractor is chosen more frequently by women or men, not even item number 5, which was answered correctly more frequently by men than by women.

3.2.4. Type of high school and correctness of the answer

We compared the type of high school considering only *liceo scientifico* and *liceo scientifico opzione scienze applicate* and the correctness of the answer because these two

types of high schools are both oriented towards teaching the sciences, but *liceo scientifico opzione scienze applicate* has more hours of science-related subjects than *liceo scientifico*. But, at the same time, the topics covered during physics classes are the same in the two types of high schools.

The results are summarized in Table 18. Some values are missing in the table because in the case of some items the expected frequency is smaller than 5, so the test could not be used.

Table 18: Chi-squared test results considering the type of high school and correctness of the answer (test number 4)

item	p-value	V	OR
1	0,46	0,06	0,67
2	0,28	0,14	0,56
3	-	-	-
4	0,67	0,02	0,79
5	0,45	0,07	1,61
6	0,23	0,17	0,52
7	0,37	0,10	0,65
8	-	-	-

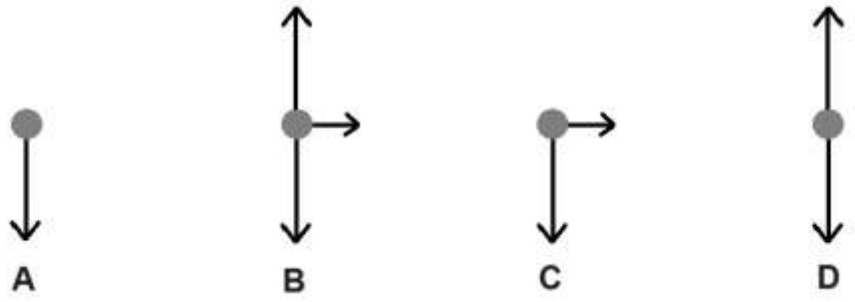
The test does not indicate that students coming from the two types of high schools that we considered tend to answer the items differently.

4 Possible item modifications of the multiple-choice test and future work

Our analysis led to some conclusions about the item quality based on the Classical Test Theory and suggested some possible improvements to the items. Some of them can be implemented as follows:

1. The statistical analysis did not suggest any changes to item 1, however, distractors C and D seam equivalent. The only difference was the system of reference because the original text did not say if the boy threw the stone to the left or the right. To avoid having two equivalent answers, we changed slightly the text and the distractors in the following way:

   A boy throws a stone on the other side of a river to see whether the stone can get to the other side of the river or not. Which of the following pictures represents the forces acting on the stone when the stone is at the maximum height of its trajectory assuming that the stone moves from left to right? Neglect air resistance.



   A. *A*
   B. B
   C. C
   D. D

   We wrote in the text that the stone moves from left to right. Answer D, in the modified item, corresponds to no force acting on the stone because the two

forces in the picture cancel out. Answer B was also changed to have an extra distractor with a force pointing to the right.

2. Answer B was identified as a non-functioning distractor because it was chosen by a low number of students, so the item was changed in the following way:

   If you apply a force F_1=5N to an object on a horizontal plane, you notice that it does not move due to friction. If you apply a force F_2=10N, it still does not move because of friction. Which statement about the intensity of friction acting on the body is correct?

   A. It is greater than 10N
   B. It does not depend on the applied horizontal force
   C. It is equal to 5N in case you apply F_1 and 10N in case you apply F_2
   D. You can not determine it using these data

3. Answer C is a non-functioning distractor because it was not chosen by enough students. The value of the centripetal acceleration of answer C was chosen assuming that some students might think that the acceleration a is given by the ratio of the velocity v divided by the period of rotation T. The period of rotation can be calculated as

$$T = \frac{2 * \pi * r}{v} = 1{,}26\ s$$

   Where r is the radius. So, the acceleration is

$$a = \frac{v}{T} = 3{,}9\ m/s^2$$

   Since the calculation is not based on the formula of the centripetal acceleration, probably very few students chose this answer. So we changed the item in the following way:

   A body has a uniform circular motion of radius 1m with a speed of 5 m/s. Which of the following statements is correct?

   A. Since it is a circular uniform motion, the speed is constant and the acceleration is zero
   B. The tangential acceleration is zero and the centripetal acceleration is 25 m/s^2
   C. The centripetal acceleration is 3,9 m/s^2
   D. None of the previous statements is correct

   The distractor C has been changed in such a way that it is not too similar to the correct answer and that it is not based on the concepts of tangential and centripetal accelerations.

4. The item was not changed because it was not needed.

5. The item was not changed despite having distractor B with a positive discrimination coefficient. Usually when high school students learn the superposition principle they apply it to electric fields and not to potential. So maybe students from the lower group chose mainly answers A and D because they struggled to apply the superposition principle to electric fields and students from the upper group chose answer B because they struggled to apply the principle mainly to the potential.

6. The item was not changed because no improvements were needed.

7. Distractor D was chosen by very few students, so it was changed in the following way:

   If you throw an object without applying any force, it falls with an acceleration of 9,8 m/s^2, neglecting the air resistance. If you, instead, apply a force directed downwards, what is its acceleration right after you stop touching it neglecting the air resistance?

   A. It changes during the fall
   B. Higher than 9,8 m/s^2
   C. Equal to 9,8 m/s^2
   D. It depends on the applied force

   Distractor D was just slightly changed and distractor A was completely changed because in the old version of the test answers A, B, and C all in the same item made it seem not plausible for the students.

8. The item was not changed. It has a discrimination coefficient below 0,3 probably because it is a very difficult item. It seems like many students got it wrong because they did not consider the action and reaction principle.

We assessed the quality of the items, but we did not examine whether the orientation course genuinely influences students' approach to physics. Enhancing the methods students use to study physics should correlate with improved performance. As the multiple-choice test shares the same question format as the engineering entrance exam, one potential avenue for exploration is comparing students' performance in the orientation course multiple-choice test with their results in the engineering entrance exam. This could be a prospective research direction undertaken by the ST2 research group.

Conclusion

The statistical analysis of the multiple choice test of the orientation course led to different results for each item. We can summarize them item per item:

1. It had medium difficulty, good discrimination, and no non-functioning distractors, however, since two distractors seemed equivalent, the item was modified;
2. The difficulty was medium, the discrimination was good, and had a distractor which was chosen by less than 5% of the students, so we introduced some modifications to the item;
3. It was one of the two hardest items, it had good discrimination and one distractor had both positive discrimination and was chosen by too few students, so the question was slightly modified;
4. The difficulty was classified as medium, the discrimination was acceptable and all the distractors were functioning, so the item was not modified;
5. It had a medium-high difficulty and good discrimination. One of the distractors had a positive discrimination coefficient, but it was not changed because, probably, even good students struggle with applying the superposition principle to potentials. The response curve does not have a clear ideal behaviour probably because not many students got high test scores;
6. The difficulty was medium and the discrimination acceptable. In addition, it had no non-functioning distractors, so no changes were needed;
7. It was the easiest item of the test and had a good discrimination coefficient. One of the distractors was chosen by less than 5% of students probably because the other answers made it seem not plausible to most of them, so the item was modified;
8. It was one of the two hardest items of the test and had a discrimination coefficient below the threshold of 0,3 probably because of the high difficulty. Therefore the item was not changed. The response curve does not have the expected behavior because many students chose a distractor even among the best-performing students in the overall test.

The correlation between item scores and test scores is good for all the items. The internal consistency of the test is low because it is made of very specific questions and the final test scores are well distributed over the possible range. So, we can conclude that, overall, the test was well-designed for our purposes.

In addition, using the chi-squared test, we obtained some interesting results which need further investigation:

- Students doing their fourth year of high school answered item number 4 correctly more frequently than students doing their fifth year;
- Men answered correctly item number 5 more frequently than women.

Bibliography

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, OBandalos, D. (2018). *Measurement Theory and Applications for the Social Sciences*. Guilford Press. https://www.guilford.com/books/Measurement-Theory-and-Applications-for-the-Social-Sciences/Deborah-Bandalos/9781462532131

Bozzi, M. (2021). *Improving the learning experience in STEM programmes: Peer learning as a key factor of an integrated approach in large size classes*. https://www.politesi.polimi.it/handle/10589/177093

Caldwell, J. E. (2007). Clickers in the Large Classroom: Current Research and Best-Practice Tips. *CBE—Life Sciences Education*, *6*(1), 9–20. https://doi.org/10.1187/cbe.06-12-0205

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics - Physics Education Research*, *5*(2), 020103. https://doi.org/10.1103/PhysRevSTPER.5.020103

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), 010105. https://doi.org/10.1103/PhysRevSTPER.2.010105

Doran, R. L. (1980). *Basic Measurement and Evaluation of Science Instruction*. National Science Teachers Association, 1742 Connecticut Ave. https://eric.ed.gov/?id=ED196733

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press. https://doi.org/10.1017/CBO9780511761676

Espa, G., & Micciolo, R. (2012). *Analisi esplorativa dei dati con R* (1° edizione). Apogeo Education.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). *PhysPort Assessments: Force Concept Inventory*. PhysPort. https://www.physport.org/assessments/assessment.cfm?A=FCI

Hu, D., Zwickl, B. M., Wilcox, B. R., & Lewandowski, H. J. (2017). Qualitative investigation of students' views about experimental physics. *Physical Review Physics Education Research*, *13*(2), 020134. https://doi.org/10.1103/PhysRevPhysEducRes.13.020134

Koch, A. J., Sackett, P. R., Kuncel, N. R., Dahlke, J. A., & Beatty, A. S. (2022). Why women STEM majors are less likely than men to persist in completing a STEM

degree: More than the individual. *Personality and Individual Differences*, *190*, 111532. https://doi.org/10.1016/j.paid.2022.111532

Magni, F. (2023). *A challenge for school autonomy*.

Montecinos, A., & Anguita, E. (2015). Being a Woman in The World of Physics Education: Female Physics Student Teachers' Beliefs About Gender Issues, in the City of Valparaiso, Chile, from a Qualitative Perspective. *Procedia - Social and Behavioral Sciences*, *197*, 977–982. https://doi.org/10.1016/j.sbspro.2015.07.286

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*. https://doi.org/10.1080/2331186X.2017.1301013

Ranieri, M., Raffaghelli, J. E., & Bruni, I. (2021). Game-based student response system: Revisiting its potentials and criticalities in large-size classes. *Active Learning in Higher Education*, *22*(2), 129–142. https://doi.org/10.1177/1469787418812667

Rea, L. M., & Parker, R. A. (2014). *Designing and Conducting Survey Research*.

Rosenthal, J. A. (1996). Qualitative Descriptors of Strength of Association and Effect Size. *Journal of Social Service Research*, *21*(4), 37–59. https://doi.org/10.1300/J079v21n04_02

Soliani, L. (2015). *Statistica di base*. Piccin-Nuova Libraria.

Taber, K. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*, 1–24. https://doi.org/10.1007/s11165-016-9602-2

rlando, FL 32887 ($44.

Appendix

The appendix reports the results of the 30 calculations of the discrimination coefficients D25, D33, and D50 corresponding to the calculation using groups made of 25%, 33%, and 50% of students respectively.

n is the number of the calculation.

The discrimination coefficient is considered acceptable above 0.3. The values below this threshold are highlighted in red.

At the end of the table, there are reported the average, maximum, and minimum values of all 30 calculations.

Table 19: Calculations of the discrimination coefficients for each item and the averages (ave), maximum (max) and minimum values (min)

n	D	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
1	D25	0,46	0,54	0,43	0,57	0,43	0,54	0,50	0,25
1	D33	0,43	0,46	0,41	0,51	0,32	0,41	0,49	0,24
1	D50	0,34	0,23	0,30	0,39	0,32	0,23	0,43	0,16
2	D25	0,50	0,46	0,54	0,57	0,25	0,50	0,57	0,32
2	D33	0,32	0,46	0,43	0,49	0,32	0,43	0,54	0,27
2	D50	0,24	0,31	0,26	0,40	0,28	0,29	0,36	0,19
3	D25	0,43	0,57	0,54	0,57	0,32	0,46	0,50	0,32
3	D33	0,43	0,43	0,43	0,51	0,30	0,38	0,54	0,24
3	D50	0,28	0,36	0,30	0,38	0,17	0,27	0,40	0,17
4	D25	0,50	0,54	0,50	0,57	0,32	0,46	0,54	0,29
4	D33	0,46	0,51	0,41	0,51	0,32	0,32	0,46	0,27
4	D50	0,35	0,26	0,26	0,33	0,28	0,24	0,38	0,24
5	D25	0,46	0,64	0,39	0,54	0,32	0,54	0,57	0,25
5	D33	0,38	0,46	0,41	0,54	0,35	0,41	0,46	0,27
5	D50	0,33	0,29	0,26	0,33	0,30	0,26	0,34	0,23
6	D25	0,46	0,54	0,46	0,54	0,32	0,57	0,54	0,29
6	D33	0,38	0,46	0,41	0,57	0,30	0,41	0,51	0,24
6	D50	0,33	0,33	0,30	0,40	0,26	0,24	0,34	0,14
7	D25	0,46	0,54	0,39	0,61	0,32	0,54	0,64	0,21
7	D33	0,41	0,43	0,41	0,57	0,32	0,32	0,54	0,27
7	D50	0,28	0,26	0,32	0,40	0,28	0,22	0,38	0,21
8	D25	0,46	0,54	0,50	0,57	0,32	0,50	0,54	0,29

n	D	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
8	D33	0,41	0,49	0,43	0,51	0,30	0,38	0,51	0,24
8	D50	0,30	0,31	0,30	0,37	0,26	0,22	0,38	0,21
9	D25	0,43	0,57	0,46	0,57	0,36	0,46	0,57	0,29
9	D33	0,46	0,43	0,41	0,49	0,32	0,43	0,46	0,27
9	D50	0,31	0,33	0,33	0,35	0,23	0,19	0,36	0,24
10	D25	0,46	0,54	0,46	0,57	0,32	0,50	0,57	0,29
10	D33	0,43	0,43	0,41	0,49	0,35	0,38	0,51	0,27
10	D50	0,37	0,29	0,30	0,35	0,24	0,22	0,34	0,23
11	D25	0,43	0,54	0,54	0,61	0,32	0,43	0,54	0,32
11	D33	0,41	0,49	0,41	0,54	0,30	0,38	0,51	0,24
11	D50	0,33	0,29	0,26	0,38	0,28	0,22	0,34	0,23
12	D25	0,46	0,50	0,54	0,61	0,32	0,54	0,46	0,29
12	D33	0,43	0,49	0,41	0,51	0,38	0,38	0,46	0,22
12	D50	0,31	0,36	0,33	0,30	0,28	0,13	0,45	0,17
13	D25	0,46	0,50	0,46	0,61	0,32	0,54	0,57	0,25
13	D33	0,43	0,43	0,43	0,51	0,32	0,41	0,46	0,27
13	D50	0,30	0,29	0,28	0,37	0,28	0,20	0,38	0,24
14	D25	0,50	0,61	0,50	0,57	0,32	0,46	0,46	0,29
14	D33	0,41	0,51	0,43	0,51	0,35	0,41	0,43	0,22
14	D50	0,30	0,29	0,30	0,35	0,28	0,24	0,41	0,17
15	D25	0,46	0,54	0,50	0,57	0,29	0,46	0,57	0,32
15	D33	0,41	0,49	0,41	0,49	0,30	0,41	0,51	0,27
15	D50	0,35	0,36	0,26	0,40	0,21	0,17	0,41	0,17
16	D25	0,50	0,57	0,46	0,57	0,32	0,54	0,50	0,25
16	D33	0,43	0,49	0,43	0,46	0,30	0,38	0,54	0,24
16	D50	0,24	0,33	0,26	0,33	0,30	0,26	0,43	0,19
17	D25	0,54	0,54	0,50	0,61	0,25	0,46	0,57	0,25
17	D33	0,41	0,49	0,38	0,54	0,32	0,38	0,49	0,27
17	D50	0,30	0,36	0,30	0,38	0,28	0,19	0,38	0,16
18	D25	0,54	0,54	0,54	0,61	0,25	0,46	0,46	0,32
18	D33	0,38	0,43	0,41	0,51	0,32	0,41	0,54	0,27
18	D50	0,24	0,33	0,33	0,30	0,28	0,17	0,41	0,28
19	D25	0,46	0,57	0,46	0,57	0,29	0,50	0,57	0,29
19	D33	0,46	0,46	0,41	0,57	0,30	0,38	0,46	0,24
19	D50	0,31	0,31	0,26	0,40	0,24	0,19	0,38	0,24
20	D25	0,46	0,57	0,50	0,54	0,29	0,54	0,54	0,29
20	D33	0,43	0,46	0,43	0,51	0,32	0,38	0,49	0,24
20	D50	0,35	0,33	0,28	0,33	0,28	0,19	0,38	0,21
21	D25	0,54	0,50	0,46	0,61	0,29	0,43	0,57	0,32
21	D33	0,41	0,43	0,41	0,57	0,30	0,43	0,51	0,22
21	D50	0,28	0,33	0,30	0,33	0,21	0,27	0,41	0,21
22	D25	0,50	0,57	0,46	0,64	0,25	0,46	0,54	0,29
22	D33	0,43	0,51	0,38	0,51	0,32	0,35	0,51	0,24
22	D50	0,35	0,36	0,30	0,40	0,21	0,24	0,38	0,10
23	D25	0,39	0,61	0,46	0,61	0,36	0,50	0,54	0,25
23	D33	0,46	0,51	0,38	0,51	0,30	0,38	0,51	0,22

n	D	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
23	D50	0,31	0,38	0,26	0,33	0,24	0,22	0,41	0,17
24	D25	0,46	0,50	0,50	0,54	0,32	0,57	0,50	0,32
24	D33	0,38	0,51	0,38	0,51	0,38	0,41	0,46	0,24
24	D50	0,30	0,33	0,23	0,37	0,26	0,26	0,38	0,23
25	D25	0,50	0,54	0,46	0,64	0,29	0,46	0,50	0,32
25	D33	0,43	0,43	0,41	0,46	0,32	0,38	0,54	0,30
25	D50	0,37	0,31	0,33	0,26	0,26	0,15	0,45	0,21
26	D25	0,46	0,57	0,46	0,61	0,32	0,50	0,54	0,25
26	D33	0,41	0,54	0,38	0,54	0,30	0,41	0,49	0,22
26	D50	0,37	0,29	0,30	0,40	0,23	0,24	0,38	0,14
27	D25	0,46	0,46	0,54	0,61	0,36	0,43	0,57	0,29
27	D33	0,38	0,43	0,46	0,49	0,32	0,43	0,51	0,24
27	D50	0,28	0,36	0,30	0,28	0,21	0,27	0,43	0,21
28	D25	0,57	0,54	0,50	0,61	0,21	0,46	0,54	0,29
28	D33	0,38	0,46	0,43	0,51	0,30	0,38	0,51	0,30
28	D50	0,33	0,36	0,32	0,30	0,24	0,13	0,45	0,21
29	D25	0,46	0,61	0,46	0,50	0,32	0,54	0,57	0,25
29	D33	0,41	0,49	0,38	0,54	0,30	0,41	0,54	0,22
29	D50	0,33	0,33	0,32	0,44	0,21	0,20	0,41	0,10
30	D25	0,43	0,57	0,43	0,61	0,36	0,54	0,54	0,25
30	D33	0,41	0,46	0,43	0,49	0,30	0,43	0,51	0,24
30	D50	0,33	0,36	0,30	0,35	0,24	0,22	0,38	0,16
ave	D25	0,48	0,55	0,48	0,58	0,31	0,50	0,54	0,28
ave	D33	0,41	0,47	0,41	0,52	0,32	0,39	0,50	0,25
ave	D50	0,31	0,32	0,29	0,36	0,25	0,22	0,39	0,19
max	D25	0,57	0,64	0,54	0,64	0,43	0,57	0,64	0,32
max	D33	0,46	0,54	0,46	0,57	0,38	0,43	0,54	0,30
max	D50	0,37	0,38	0,33	0,44	0,32	0,29	0,45	0,28
min	D25	0,39	0,46	0,39	0,50	0,21	0,43	0,46	0,21
min	D33	0,32	0,43	0,38	0,46	0,30	0,32	0,43	0,22
min	D50	0,24	0,23	0,23	0,26	0,17	0,13	0,34	0,10

List of Figures

List of Tables

List of symbols

Variable	Description	SI unit
P	Item difficulty	-
N_C	Number of correct responses	-
N	Total number of responses	-
σ	Standard deviation	-
D	Discrimination coefficient	-
P_u	Proportion of students who answered correctly an item from the upper group	-
P_l	Proportion of student who answered correctly an item from the lower group	-
r_{pbi}	Point biserial coefficient	-
X_1	Average total score of the students who correctly answered an item	-
X_0	Average total score of the students who incorrectly answered an item	-
K	Total number of item in a test	-
δ	Ferguson's Delta	-
f_i	Total number of students following a certain criterioni	-
ave	Average of the discrimination coefficients	-
SD	Standard deviation of the discrimination coefficients	-
tot	Percentage of students who chose each answer	-
n	Number of non functioning distractors per item	-
n_{ij}	Observed frequency	-

Variable	Description	SI unit
$\hat{n}_{ij}$	Expected frequency	-
R	Number of rows in a contingency table	-
C	Number of columns in a contingency table	-
χ^2	Chi-squared	-
T	Total number of observations	-
k	Smallest numbet between R and C	-
OR	Odds ratio	-
V	Cramer's V	-