POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

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

TESI MAGISTRALE IN ENGINEERING PHYSICS – INGEGNERIA FISICA

AUTHOR: SARA PITTINI

ADVISOR: MAURIZIO ZANI

ACADEMIC YEAR: 2022-2023

1. Introduction

Politecnico di Milano is actively involved in a project called *Orientamento 2026-Orientamento attivo nella transizione scuola università*, an initiative launched by the Italian Ministry of University and Research. The primary goal of this project is to provide orientation courses for high school students in their last three years, guiding them in preparation for the engineering entrance exam.

Beyond the standard educational support, Politecnico di Milano's course incorporates physics-related content and employs dynamic teaching methods, emphasizing hands-on laboratory activities. These sessions also integrate a questionnaire designed to uncover and address students' misconceptions and conceptual errors, which are incorrect beliefs they may have about the physical world.

After the completion of the courses, the collected data underwent analysis using Classical Test Theory to identify the most effective questionnaire items. Distractor evaluation was employed to weed out non-functioning distractors, and the chi-squared test was used to draw insightful conclusions by exploring correlations among various factors.

2. Orientation course structure

Our 15-hour course comprised three segments:

- A 3-hour presentation conducted by IFOA, a training and consultancy center.
- 6 hours dedicated to math classes.
- 6 hours focused on physics classes.

The 6-hour physics classes were split into two sessions. The initial two-hour session took place online, while the subsequent four-hour session occurred in person.

During the online session, students were tasked with responding to 8 multiple-choice questions closely resembling those featured in the engineering entrance test. These questions served dual purposes: firstly, to provide a preliminary assessment of the students' physics knowledge, and secondly, to allow them to tackle questions similar to those in the engineering entrance test.

This facilitated an understanding of the requisite skills for success on the exam. Moreover, some questions aimed to address prevalent misconceptions and conceptual errors among students, enhancing their awareness of these issues.

Each student participated in the assessment individually, without assistance from peers or the professor, and had a limited timeframe for each question. The students utilized Socrative, an application designed for creating quizzes and collecting responses from participants, to complete the questionnaire.

3. Classical Test Theory analysis

Item analysis encompasses a range of strategies employed to choose the most suitable items from a pool of potential candidates. The numerical results of the analysis are summarized in Tables 1.

In the context of multiple-choice tests, Classical Test Theory introduces several useful indices for analysis. Among these is the item difficulty (P), calculated as the ratio of the number of correct responses (N_C) to the total number of responses (N) (Ding & Beichner, 2009).

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

The item difficulty serves as a tool to differentiate items across various difficulty levels, which can be categorized, according to one of the classifications, as low (L), medium-low (ML), medium (M), medium-high (MH), or high (H) (Crocker & Algina, 1986). The difficulty scale ranges from 0.19 to 0,58, ensuring that the test maintains an appropriate level of challenge for students without being overly easy or difficult.

Another crucial index is item discrimination (D), grounded in the idea that low-achieving students are more prone to answering an item incorrectly, while high-achieving students are more likely to provide the correct response. Initially, students are divided into upper and lower groups based on their total test scores, with these groups typically representing the top and bottom 50%, 33%, or 25% of students. The discrimination coefficient is calculated as the proportion of students in the upper group who answered correctly (P_u) minus the proportion of students who answered correctly in the lower group (P_l).

$$D = P_u - P_l$$

Depending on the group size, discrimination coefficients can be defined as D25, D33, and D50, calculated using the 25%, 33%, and 50% of students, respectively. The discrimination index is considered acceptable if it is greater than or equal to 0,3 (Ding & Beichner, 2009).

When forming two groups, two crucial factors should be taken into account:

- Smaller groups prevent the possibility of students with the same total score ending up in different groups.
- Small groups may overlook the performance of students with scores around the average.

Groups consisting of 33% of students strike a balance between these considerations. Relying solely on quartiles neglects the performance of half of the students taking the test while placing all students in a group increases the impact of the random division of students with scores around the average.

The discrimination indexes exhibit acceptable values except for item number 8, which has a discrimination index of 0,25 when considering groups composed of 33% of students, slightly below the 0,3 threshold.

The third index considered is the point biserial coefficient (r_{pbi}), representing 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 represents the average total score of students who correctly answered the i-th item, while X_0 signifies the average total score of students who answered the i-th item incorrectly, and σ_x is the standard deviation of the total scores. A low point biserial coefficient indicates that an item does not assess the same material as the others (Ding & Beichner, 2009). It is recommended that the point biserial coefficient be greater than or equal to 0,2 for satisfactory performance,

All point biserial coefficients in this analysis surpass 0,2, indicating that the items exhibit good reliability.

The Kuder-Richardson reliability (r_{test}) assesses whether the items measure the same ability. The coefficient's magnitude reflects the correlation between the items and is determined by the following formula:

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

K represents the total number of items in the entire test, and P_i denotes the difficulty index of the i-th item (Ding & Beichner, 2009).
The Kuder-Richardson index is reported as 0.31. While an index above 0.8 is typically considered acceptable, in the case of tests consisting of highly specific questions covering a diverse range of topics, Kuder-Richardson values tend to be lower. Additionally, given that the test comprises only 8 items, it is not intended to comprehensively evaluate a student's knowledge across various physics topics. This intentional limitation is because the primary goal of the test is not to gauge the depth of a student's understanding but rather to uncover potential misconceptions and conceptual errors in physics.
Finally, Ferguson's Delta (δ) assesses how effectively the final test scores are distributed over the possible range and can be calculated as follows:

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

Where f_i represents the number of students with a score equal to i, an acceptable value for Ferguson's Delta should be greater than 0,9 (Ding & Beichner, 2009).
Ferguson's delta is determined to be 0,90, precisely meeting the minimum acceptable value. Therefore, the tests effectively discriminate among students.
Non-functioning distractors are incorrect answers chosen by less than 5% of students or those with a positive discrimination coefficient. Essentially, a distractor should present a plausible alternative to the correct answer while attracting more students from the lower group than from the upper group (Quaigrain & Arhin, 2017).
The distractors chosen by less than 5% of students are:

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

Distractors with positive discrimination coefficients are:

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

In total, there are 4 non-functioning distractors out of 24, translating to 0 or 1 non-functioning distractor per item. This suggests that future revisions might involve addressing specific distractors rather than overhauling entire items.

	Q1	Q2	Q3
P	0,31	0,38	0,19
difficulty	M	M	MH
D25	0,48	0,55	0,48
D33	0,41	0,47	0,41
D50	0,31	0,32	0,29
rpbi	0,45	0,45	0,51
n	0	1	1

	Q4	Q5	Q6
P	0,27	0,25	0,5
difficulty	M	MH	M
D25	0,58	0,31	0,5
D33	0,52	0,32	0,39
D50	0,36	0,25	0,22
rpbi	0,52	0,3	0,37
n	0	1	0

	Q7	Q8
P	0,58	0,19
difficulty	ML	MH
D25	0,54	0,28
D33	0,5	0,25
D50	0,39	0,19
rpbi	0,44	0,27
n	1	0

Tables 1: Summary of the numerical results obtained using Classical Test Theory. n is the total number of non-functioning distractors per item.

4. χ^2test

The chi-squared test serves as a method for comparing frequencies and proportions and can also be applied to contingency tables to assess the independence between two factors (Soliani, 2015). In our case, we utilized the test to investigate potential correlations between answers and factors such as grade or gender.

The null hypothesis states that the two factors are independent, with any differences attributed only to statistical fluctuations. On the other hand, the alternative hypothesis suggests that the two factors are dependent, allowing for the rejection of the null hypothesis.

In research studies, it is crucial to introduce effect sizes to gauge the significance of a result. The effect size represents the magnitude of a result and can be expressed in various ways, with two common measures being Cramer's V and the odds ratio.

Cramer's V is defined as:

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

Where χ^2 is the value calculated while performing the χ^2 test, T is the total number of observations and k is the smallest number between the number of rows and columns. The maximum value of V is 1 (Soliani, 2015).

Instead, the odds ratio, in the case of a 2×2 contingency table is defined as

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

Where a, b, c, and d are the elements of the contingency table defined as in Table 2.

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

Tables 2: Example of contingency table.

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.

All the tests gave a p-value smaller than 0,05, except for:

- test number 1. in the case of item number 4. The p-value was 0,04, the Cramer's V of 0,40, and the odds ratio of 4,44. So we concluded that students doing their fourth year of high school tend to answer item number 4 more frequently than students doing their fifth year;
- test number 2. in the case of item number 5. The p-value was 0,05, the Cramer's V of 0,37, and the odds ratio of 4,19. According to the test, men tend to answer correctly item number 5 more frequently than women.

5. Conclusions

The analysis of the multiple-choice test in the orientation course yielded varied results for each item. We can summarize the findings item by item:

1. It exhibited moderate difficulty, good discrimination, and no non-functioning distractors.
2. The difficulty level was moderate, discrimination was good, and it included a distractor chosen by fewer than 5% of students.
3. Identified as one of the two most challenging items, it demonstrated good discrimination, and one distractor showed positive discrimination but was chosen by too few students.
4. Classified with a medium difficulty, acceptable discrimination, and all distractors functioning.
5. Presented with a medium-high difficulty and good discrimination. One distractor displayed a positive discrimination coefficient, suggesting even proficient students struggled with the question's topic.
6. Displayed a medium difficulty and acceptable discrimination, with the added benefit of having no non-functioning distractors.

7. Ranked as the easiest item in the test, featuring a good discrimination coefficient. One distractor was chosen by less than 5% of students.
8. Identified as one of the two most challenging items in the test, with a discrimination coefficient below the 0,3 threshold, likely due to its high difficulty.

The correlation between item scores and test scores was strong for all items. However, the test's internal consistency was low, given its composition of highly specific questions and the well-distributed final test scores across the possible range. Overall, it can be concluded that the test was well-designed for its intended purposes.

Furthermore, employing the chi-squared test revealed intriguing results warranting further investigation:

- Fourth-year high school students answered item number 4 correctly more frequently than their fifth-year counterparts.
- Men outperformed women in correctly answering item number 5.

References

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887 ($44.

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

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

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