Validation of a Multiple-Choice Physics Test Using Classical Test Theory: Toward a More Up-to-Date International Item Database

Andrea Fusari, Politecnico di Milano, Italy
Matteo Bozzi, Politecnico di Milano, Italy
Roberto Virzi, Politecnico di Milano, Italy
Marco Costigliolo, Politecnico di Milano, Italy
Gwenaëlle Guillerme, T.I.M.E. Association, France
Maurizio Zani, Politecnico di Milano, Italy

The Paris Conference on Education 2026
Official Conference Proceedings

Abstract

This study examines the psychometric properties of a multiple-choice physics questionnaire using the framework of Classical Test Theory (CTT), with the aim of validating individual items and contributing to the development of a reliable item bank for research and assessment. The instrument was designed to investigate students' conceptual understanding of fundamental physics topics and to identify common misconceptions or conceptual errors that may persist after formal instruction. The questionnaire is being developed within the T.I.M.E. project by Politecnico di Milano in collaboration with Doshisha University, Bauman Moscow State Technical University, and the University of Trento. The dataset includes responses from approximately 1,000 university students enrolled in engineering and science programs. This internationally diverse sample includes students from Italian, Japanese, and Russian institutions, enabling the analysis of response patterns across different educational contexts and exploring how prior schooling systems influence the persistence of physics misconceptions. Item analysis was conducted using standard CTT indicators, including item difficulty, discrimination indices, and distractor functioning. Particular attention was devoted to distractor performance to identify functional alternatives that capture specific incorrect reasoning patterns. Results indicate that ten of the twelve analyzed items include effective distractors and are well-constructed, while two items exhibit weak distractors that may require revision. These results represent a first step toward the construction of an updated international physics item bank, which could support both formative and summative assessment and enable comparative studies of physics learning across different educational systems.

Keywords: misconceptions, distractor, test, stem, physics

iafor
The International Academic Forum
www.iafor.org

Introduction

Multiple-Choice Questions (MCQs) are among the most popular evaluation tools in education and psychometrics. Their widespread adoption across standardized assessments and academic contexts is primarily driven by their structural efficiency, which allows for the objective, fair, and replicable scoring of large student cohorts in reduced timeframes, while ensuring an extensive coverage of curriculum topics within a single testing session (*Effective Assessment and the Improvement of Education*, 2017; Popham, 2020). However, the utility of MCQs extends far beyond mere administrative convenience or the assessment of low-level rote learning. When built with proper scientific rigor, they do much more than just test factual memory (Haladyna, 2004): they can effectively measure higher-order cognitive skills—such as critical thinking, analytical reasoning, and the application of conceptual knowledge to novel scenarios—and help diagnose deep-seated student misconceptions (Haladyna et al., 2002). By meticulously embedding specific incorrect reasoning patterns into functional alternative options, well-designed MCQs transform a summative grading tool into a powerful diagnostic instrument capable of mapping the intricate and flawed mental models that students often harbor after formal instruction (Bozzi et al., 2020).

A standard multiple-choice question relies on a clear three-part structure: a stem (the question or problem statement), a key (the single correct and scientifically sound answer), and the distractors (the incorrect options). However, writing these options is far from a mechanical task; as Haladyna (2004) wrote, *"item writing is an art form that is informed by science."* While statistical analysis evaluates a test *after* it has been taken, initially designing a question that truly taps into a student's thinking requires a mix of pedagogical intuition and deep subject expertise (Haladyna et al., 2002).

Nevertheless, to turn a collection of questions into a truly reliable research instrument, ex-post statistical validation is essential. Using frameworks like Classical Test Theory (CTT) (Crocker & Algina , 2006) or Item Response Theory (IRT) (Lord, 1980), researchers evaluate the quality of MCQs through specific psychometric parameters. These include the difficulty index (p), which represents the proportion of students who answered the item correctly. Monitoring this index is crucial, as it helps researchers maintain an overall balance in the test's difficulty. Additionally, the evaluation relies on the discrimination index, with the point-biserial correlation serving as a primary indicator. This parameter measures how effectively a question differentiates between high-performing students and those who struggle with the material.

Finally, distractor analysis looks at how incorrect answers are distributed across different performance groups (Haladyna & Downing, 1993). This last parameter is particularly crucial. Analyzing distractors does not just prove that the alternative options are plausible and clear; it fundamentally allows us to track and identify persistent misconceptions within the student sample. Indeed, in well-crafted tests, distractors are never chosen at random; instead, they are designed around common reasoning errors and flawed mental models, making them invaluable for diagnostic analysis and for mapping student misconceptions (Bozzi et al., 2020).

This study focuses specifically on the distractor analysis of a physics questionnaire developed by Politecnico di Milano within the T.I.M.E. (*Top International Managers in Engineering*) project, a collaborative effort with the Bauman Moscow State Technical University (Russia), Doshisha University (Japan) and the University of Trento.

Research Questions

RQa: How can a distractor efficiency analysis guide the specific revision and optimization of the test items?

RQb: What distinct response patterns emerge when evaluating these items across different student nationalities?

Theoretical Framework and Item-Level Analysis

This research is situated within the framework of Classical Test Theory (CTT). Within this paradigm, psychometric evaluation involves both item-level parameters, which assess the performance of individual tasks, and global parameters, which evaluate the instrument as a whole.

At the individual item level, a primary metric evaluated is the Item Difficulty Index (p), defined as in Equation 1.

$$p = \frac{N_1}{N} \tag{1}$$

Where N_1 represents the number of correct responses and N denotes the total sample size. Under this formulation, lower values of p indicate greater item difficulty. Based on their empirical values of difficulty index, items are stratified into distinct operational categories, ranging from difficult and moderately difficult to medium (optimal) and moderately easy.

To evaluate the capacity of an item to distinguish between high-performing and low-performing examinees, the Item Discrimination Index (D) is computed. The sample is first ranked by total score and partitioned into four equal-sized groups representing score quartiles. Equation 2 expresses the mathematical formulation of the D index.

$$D = \frac{N_H - N_L}{n} \tag{2}$$

Where N_H and N_L represent the number of correct responses within the highest and lowest scoring quartiles, respectively, and n is the total number of respondents within a single quartile.

Because identical total scores (ties) can lead to arbitrary student placement at the boundaries of the quartiles, the initial random ordering of the dataset could artificially skew the resulting discrimination index. To mitigate this potential bias, an iterative resampling procedure was implemented. Specifically, the dataset was subjected to 100,000 random re-orderings (shuffling) within identical score strata prior to quartile assignment, and the discrimination index was subsequently recalculated for each iteration. The final classification of the items, following the standard psychometric thresholds established by Crocker and Algina (Crocker & Algina, 2006), was ultimately determined using the mean discrimination value obtained across these iterations.

An additional robust metric employed to evaluate item discrimination and option efficiency is the Point-Biserial Correlation Coefficient (r_{pb}). In the context of distractor analysis, this coefficient measures the correlation between a student's choice of a specific response option (correct or incorrect) and their overall score on the assessment. It is defined by Equation 3.

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{\sigma_x}\sqrt{p(1-p)} \tag{3}$$

Where $\bar{X}_1$ represents the mean total score of examinees who selected the specific option, $\bar{X}_0$ is the mean total score of examinees who did not select that option, σ_x is the standard deviation of the total test scores, and p is the difficulty index of the item.

The diagnostic value of the point-biserial coefficient varies depending on whether the analyzed option is the correct answer (the key) or an incorrect alternative (a distractor). For the correct key, a coefficient of $r_{pb} \geq 0.20$ is considered indicative of an acceptable-to-strong positive correlation, confirming that higher-performing students are successfully selecting the correct answer. Conversely, for functional distractors, an ideal coefficient falls within the negative range of $-0.20 \leq r_{pb} < 0$, demonstrating that the option successfully draws lower-performing students away from the key (Attali & Fraenkel, 2000).

Furthermore, a strongly negative coefficient where $r_{pb} < -0.20$ highlights a highly effective distractor that likely aligns with a prevalent student misconception or a systematic conceptual error (Aslanides & Savage, 2013). On the other hand, if a distractor yields a positive correlation where $r_{pb} > 0$, it implies that high-performing students are preferentially selecting this incorrect option. This statistical anomaly strongly suggests that the item phrasing is ambiguous or flawed, thereby signaling the need for immediate qualitative revision.

To evaluate the overall reliability and internal consistency of the assessment instrument, the Kuder-Richardson Formula 20 (KR-20) was computed. Since the items are scored dichotomously as correct or incorrect, the KR-20 index serves as a robust measure to determine the degree to which all tasks within the questionnaire consistently measure the same underlying conceptual construct. The formula is expressed by Equation 4.

$$KR\text{-}20 = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} p_i(1-p_i)}{\sigma_X^2}\right) \tag{4}$$

Within this formulation, the parameters are defined as follows: K represents the total number of items comprising the instrument, p_i denotes the Item Difficulty Index of the i-th item, and σ_X^2 signifies the variance of the observed total test scores. In line with psychometric literature, values equal to or greater than 0.70 are considered acceptable for low-stakes diagnostic research tools, while values equal to or exceeding 0.80 indicate high internal consistency, ensuring that the observed test scores are stable and largely unaffected by measurement error (Bland & Altman, 1997).

While item-level discrimination parameters evaluate individual tasks, Ferguson's Delta (δ) is employed as a global psychometric index to assess the comprehensive discriminatory power of the entire assessment instrument. Specifically, it measures the degree to which the test can successfully differentiate among examinees across the full spectrum of abilities, maximizing the variance of the total score distribution. Ferguson's Delta is formulated Equation 5:

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - \frac{N^2}{K+1}} \tag{5}$$

Where N denotes the total sample size, K represents the total number of items in the test, and f_i is the frequency of examinees obtaining each possible total score from $i = 0$ to K.

The value of δ ranges from 0 to 1, where a value of 1.0 indicates an ideal, uniform distribution in which every possible test score is achieved by an equal number of respondents, thereby providing maximum discrimination. In accordance with psychometric standards established in educational measurement, a test is considered to possess a highly satisfactory discriminatory capacity if $\delta \geq 0.90$. A high Ferguson's Delta ensures that the instrument does not merely group students into broad performance categories but rather offers a fine-grained differentiation of their conceptual understanding.

Option Response Curves (ORCs)—frequently designated as *Distractor Response Curves*—constitute a non-parametric, graphical diagnostic tool designed to evaluate the performance of multiple-choice items. Unlike standard CTT indices, which condense distractor performance into a single, static correlation coefficient, ORCs model the empirical probability of selecting each individual response option—both the correct key and the incorrect distractors—as a continuous function of the examinee's overall proficiency. By analyzing the full spectrum of options simultaneously, ORCs provide a fine-grained, visual assessment of how effectively each alternative attracts or repels students across different ability levels.

In an ORC plot, the data are mapped across two principal axes. The horizontal axis, or abscissa, represents the examinees' proficiency level. Within a CTT framework, this is typically operationalized by partitioning the sample into ordered performance intervals based on their raw total test score, such as deciles. The vertical axis, or ordinate, denotes the conditional probability from 0 to 1, or the relative frequency of examinees within a given ability stratum who selected that specific option. Each response alternative yields its own distinct trajectory, which is expected to align with specific psychometric profiles.

The geometric behavior of an ORC allows researchers to evaluate the structural integrity of a physics item and uncover underlying cognitive patterns through distinct curve typologies (Morris et al., 2012).

The curve for the correct key must be monotonically increasing, a profile that confirms that as a student's conceptual understanding and overall test performance increase, the probability of selecting the correct option approaches unity. Conversely, functional distractors exhibit an ideal incorrect profile characterized by a monotonically decreasing trend. These alternatives should possess high attractivity for low-performing students to capture random guessing or superficial reasoning and systematically drop to zero as student proficiency increases.

Beyond these monotonic trends, non-monotonic curves provide critical value for misconception identification. A distractor that yields a bell-shaped curve peak within the intermediate ability ranges. This specific profile is highly valuable for Physics Education Research (PER) because it visually captures a transitional cognitive state where students have moved past random guessing but are strongly drawn to a structured, systematic misconception before achieving full scientific mastery.

Finally, ORCs are instrumental in identifying flawed or ambiguous items. If a distractor curve trends upward or remains stable within the high-ability strata, it indicates a critical item anomaly. A positive slope for an incorrect option suggests that top-performing students are preferentially selecting it, signaling that the option may be scientifically ambiguous, poorly phrased, or double-keyed, thereby necessitating immediate qualitative revision.

Methodology

Research Design

The diagnostic test was developed through a collaborative effort between the Politecnico di Milano and the University of Trento. This questionnaire comprises 12 multiple-choice items, systematically categorized into three macro-domains of physics: Mechanics, Thermodynamics, and Electromagnetism (Bozzi et al., 2019). Each item features a single correct response (the key) and three incorrect alternatives (distractors). Crucially, the distractors were deliberately engineered based on physics education literature to isolate and identify specific student misconceptions or systematic conceptual errors, as delineated in Table 1. No penalties or negative markings were applied for incorrect answers; consequently, the observed total test scores are distributed within a range of 0 to 12. To facilitate the tracking of items across the different conceptual areas of the test, a standard coding system was adopted: items focusing on Mechanics are labeled as IM, those on Thermodynamics as IT, and those covering Electromagnetism as IE.

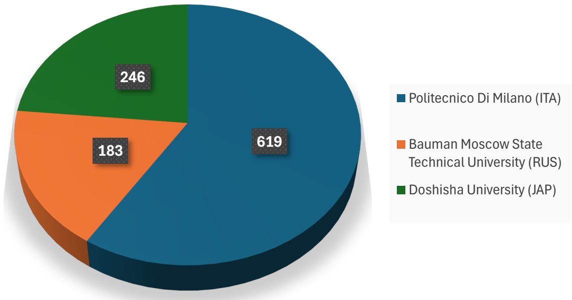
Table 1
Classification of Test Items and Mapping of the Investigated Misconception

Quiz	Misconception area	Specific heading	Macro-area
IM1	Displacement and distance travelled	Kinematics	Mechanics
IM2	Static friction	Force	Mechanics
IM3	Impulse, momentum law	Linear momentum	Mechanics
IM4	Conservation of energy, linear and angular momentum	Gravitation	Mechanics
IT1	Phase transition and heat exchanged	Thermodynamics processes	Thermodynamics
IT2	Relation between heat and energy of a Thermodynamicssystem	Heat	Thermodynamics
IT3	Quasi-static and no quasi-static adiabatic process	Thermodynamics processes	Thermodynamics
IT4	Efficiency of a heat engine and comparison with Carnot engine	Heat engine	Thermodynamics
IE1	Source of an electric field	Electric field	Electromagnetism
IE2	Electrical induction	Conductors	Electromagnetism
IE3	Electric field in the presence of a dielectric	Dielectrics	Electromagnetism
IE4	Equilibrium of a current loop lying in a magnetic field	Magnetostatics	Electromagnetism

Participants and Procedure

The initial international administration of the survey targeted undergraduate students across three distinct institutions: the Politecnico di Milano (Italy), the Bauman Moscow State Technical University (Russia), and Doshisha University (Japan) divided as shown in Figure 1. To ensure the internal validity of the study and isolate the target population, a strict filtering criterion was applied to include exclusively first-year undergraduate students (*freshmen*). Following this sample reduction procedure, the final validated dataset comprised a total sample size of $N = 1048$ students. The test items were originally drafted in English to establish a baseline conceptual framework. To ensure linguistic equivalence and cross-cultural validity across the cohorts, the questionnaire was subsequently translated into the respective native languages of the participating universities. All psychometric analyses, statistical computations, and data visualizations were executed using the R statistical computing environment (RStudio).

Figure 1
Distribution of the Sample by Nationality



Results

As a first step, the classical CTT parameters were calculated. The item difficulty index was computed for the overall sample and for each national subgroup (*Table 2*), while the discrimination index (*Table 3*) and the point-biserial correlation coefficient (*Table 4*) were calculated for the overall sample. As shown in Table 3, D_{min} and D_{max} denote the minimum and maximum values of the discrimination index D observed over the entire set of permutations.

Table 2
Calculated Item Difficulty Index for Each Item and Nationality

Item	ALL	ITA	JAP	RUS
IM1	0.31	0.14	0.46	0.70
IM2	0.25	0.24	0.24	0.32
IM3	0.36	0.34	0.38	0.41
IM4	0.18	0.13	0.29	0.20
IT1	0.63	0.60	0.62	0.77
IT2	0.51	0.58	0.42	0.37
IT3	0.18	0.18	0.20	0.17
IT4	0.48	0.54	0.42	0.37
IE1	0.41	0.44	0.31	0.44
IE2	0.12	0.10	0.15	0.14
IE3	0.45	0.50	0.31	0.48
IE4	0.16	0.11	0.19	0.30

Table 3
Mean Item Discrimination Index Obtained From 100,000 Permutations

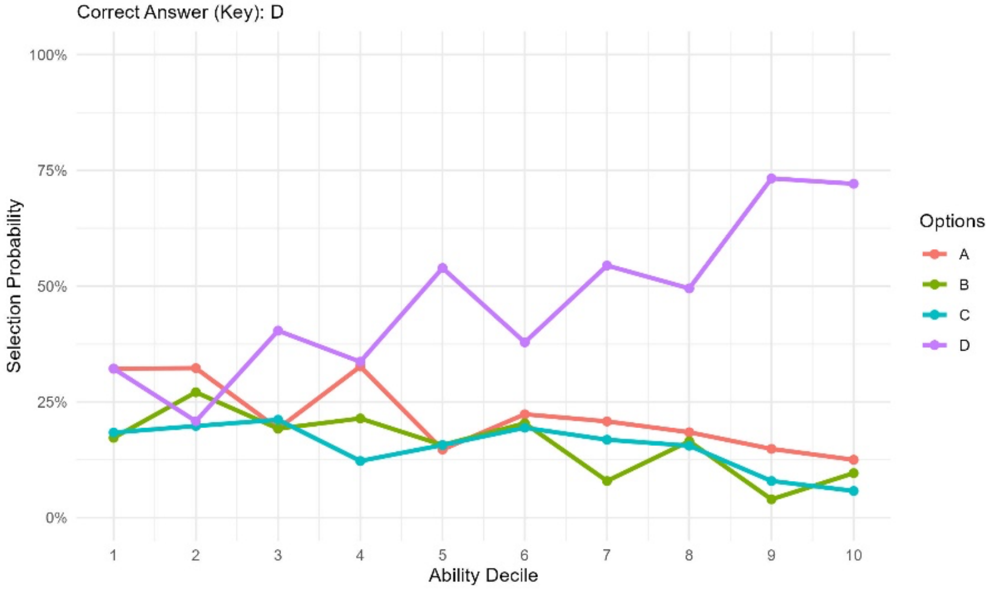
Items	Dmean	Dmin	Dmax	Classification
IM1	0.36	0.28	0.42	Reasonably good item (RG)
IM2	0.28	0.22	0.35	Marginal item (M)
IM3	0.37	0.30	0.44	Reasonably good item (RG)
IM4	0.19	0.13	0.25	Poor item (P)
IT1	0.42	0.35	0.50	Good Item (G)
IT2	0.41	0.34	0.50	Good Item (G)
IT3	0.18	0.13	0.24	Poor item (P)
IT4	0.35	0.28	0.43	Reasonably good item (RG)
IE1	0.44	0.38	0.51	Good Item (G)
IE2	0.18	0.13	0.23	Poor item (P)
IE3	0.43	0.35	0.50	Good Item (G)
IE4	0.19	0.13	0.24	Poor item (P)

Table 4
RPBI Values and Response Option Selection Percentages for Each Item

Item	Option	Type	rpbi_option	%		Item	Option	Type	rpbi_option	%
IE1	A	Distractor	-0.14	15.5		IM3	A	Distractor	-0.23	40.6
IE1	B	Distractor	-0.13	19.0		IM3	B	Key	0.35	36.0
IE1	C	Key	0.37	40.9		IM3	C	Distractor	-0.05	11.3
IE1	D	Distractor	-0.13	21.6		IM3	D	Distractor	-0.05	9.4
IE2	A	Distractor	-0.06	26.0		IM4	A	Key	0.22	17.7
IE2	B	Distractor	-0.02	37.3		IM4	B	Distractor	-0.07	38.5
IE2	C	Distractor	-0.06	22.2		IM4	C	Distractor	-0.15	18.0
IE2	D	Key	0.25	12.1		IM4	D	Distractor	0.06	23.3
IE3	A	Distractor	-0.11	20.7		IT1	A	Distractor	-0.07	2.7
IE3	B	Distractor	-0.11	15.1		IT1	B	Distractor	-0.15	9.6
IE3	C	Distractor	-0.08	14.5		IT1	C	Key	0.38	63.5
IE3	D	Key	0.30	45.0		IT1	D	Distractor	-0.23	21.9
IE4	A	Distractor	-0.03	29.7		IT2	A	Distractor	-0.24	24.3
IE4	B	Key	0.21	16.1		IT2	B	Key	0.35	50.7
IE4	C	Distractor	0.00	16.5		IT2	C	Distractor	-0.11	19.8
IE4	D	Distractor	-0.09	34.9		IT2	D	Distractor	-0.06	3.7
IM1	A	Distractor	-0.16	44.8		IT3	A	Key	0.21	18.2
IM1	B	Key	0.34	31.4		IT3	B	Distractor	-0.07	21.0
IM1	C	Distractor	-0.07	8.3		IT3	C	Distractor	-0.02	18.9
IM1	D	Distractor	-0.12	13.1		IT3	D	Distractor	-0.02	36.1
IM2	A	Distractor	-0.10	24.6		IT4	A	Distractor	-0.12	23.5
IM2	B	Distractor	-0.03	35.1		IT4	B	Distractor	-0.11	17.5
IM2	C	Distractor	-0.13	12.5		IT4	C	Key	0.29	48.4
IM2	D	Key	0.28	25.4		IT4	D	Distractor	-0.07	6.1

Option Response Curves (ORCs) were generated for all items. Figure 2 presents a representative example of an ORC. Regarding this particular item. option D was the keyed response. whereas the other choices functioned as distractors.

Figure 2
Option Response Curves for Electromagnetism Item IE3



Finally, the KR-20 coefficient and Ferguson's delta were calculated for the overall sample. yielding values of 0.10 and 0.90. respectively.

Discussion

The findings of this study offer valuable insights into both the psychometric quality of the instrument and the cultural factors influencing physics education. The overall distribution of the item difficulty index reveals a balanced assessment tool. consisting of four medium-high difficulty items. six medium difficulty items. and two low-difficulty items (*Figure 3*).

Shifting the focus to the second research question (RQb). the analysis of cross-national response patterns reveals distinct variations when the items are disaggregated by nationality. The most prominent cross-national divergence is observable in the Item Difficulty Indices (*Figure 4*). Specifically. mechanics items such as IM1 and IM2 were found to be substantially less difficult for the Russian student cohort compared to their Italian and Japanese peers. Conversely. the Italian cohort demonstrated greater proficiency in specific thermodynamics items like IT2 and IT4. These localized fluctuations in item difficulty underscore that student vulnerability to specific physical misconceptions is not universally uniform. Instead. it is modulated by national curricula and instructional methodologies. making cross-cultural validation a requirement for the creation of a standardized international item bank.

Figure 3
Difficulty Level

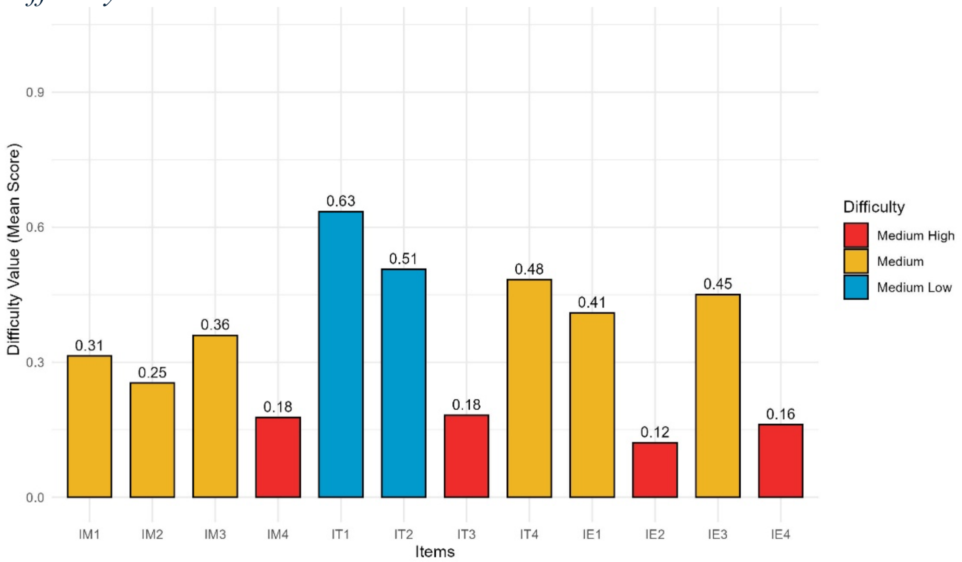
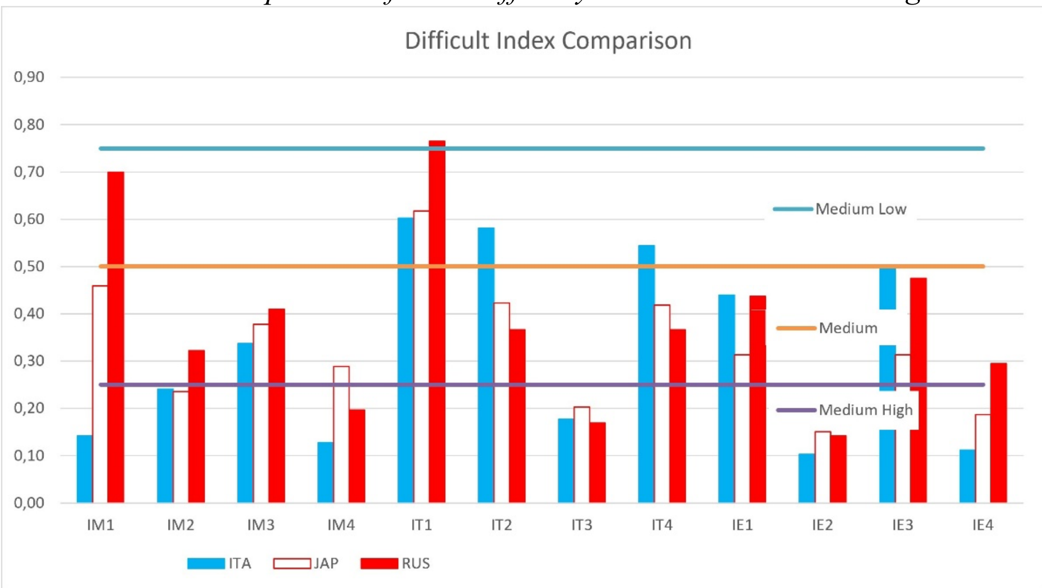
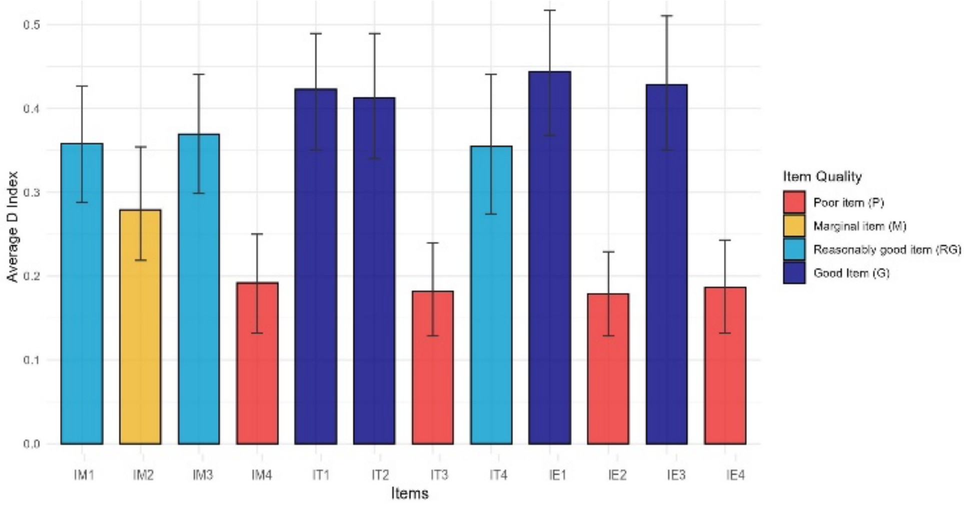


Figure 4
Cross-National Comparison of Item Difficulty Index Across the Investigated Countries



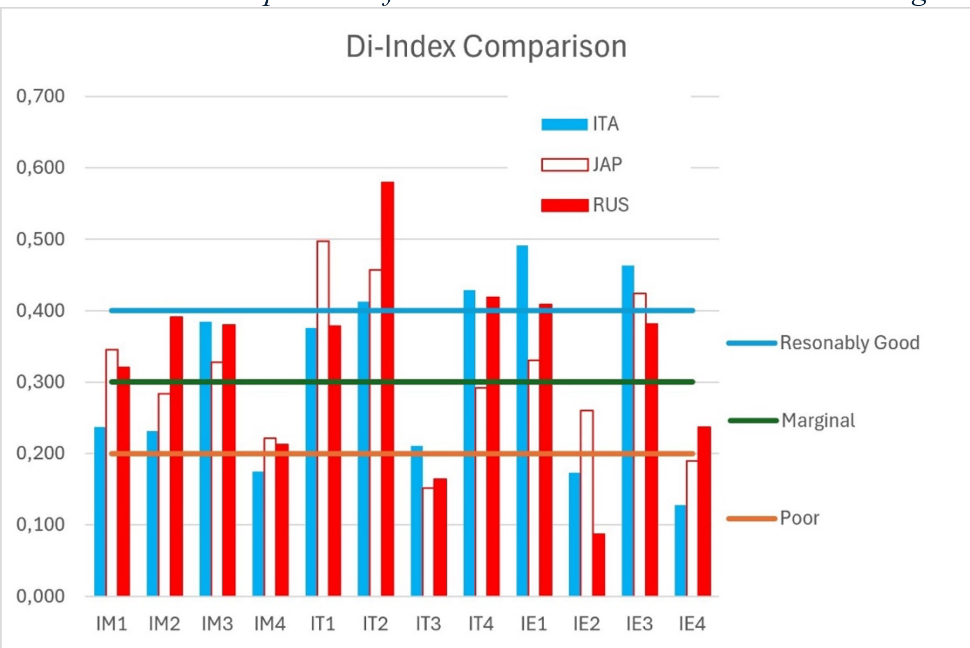
As illustrated in Figure 5. the overall discrimination index demonstrates a well-balanced profile. featuring four items with strong discriminatory power and three items that discriminate reasonably well (Ebel & Frisbie, 1991). The bars represent the Discrimination Index range as generated during the permutation cycles.

Figure 5
Overall Means Discrimination Index Across All Nationalities (100.000 Permutations)



A similar cross-national variation is observable in the discrimination index. as shown in Figure 6. although the differences among the participating nationalities are less pronounced than those found in the item difficulty indices.

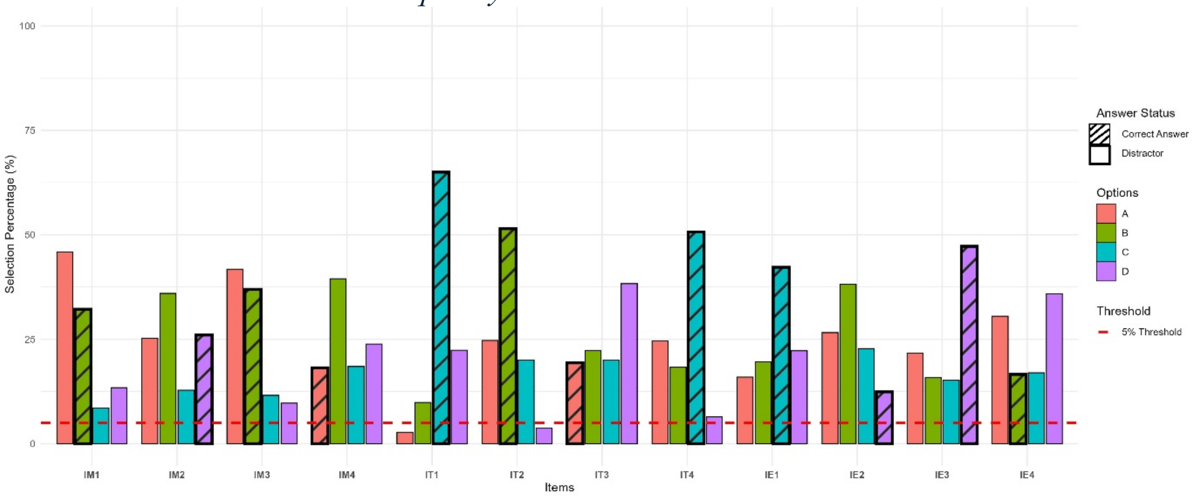
Figure 6
Cross-National Comparison of Discrimination Index Across the Investigated Countries



Regarding the global reliability indices. the KR-20 coefficient was found to be particularly low (0.10). well below the commonly accepted threshold of 0.70. The low KR-20 coefficient observed globally does not necessarily imply poor item quality; rather. as demonstrated by Cortina (1993). internal consistency estimates are strictly dependent on the unidimensionality assumption and significantly underestimate reliability when an instrument taps into multiple distinct dimensions. In the present study. this result is an expected statistical consequence of the highly heterogeneous and multidimensional content covered by the inventory. The instrument was deliberately designed as a diagnostic inventory aimed at identifying students' misconceptions rather than as a traditional unidimensional summative assessment.

Regarding the first research question (RQa). the distractor efficiency analysis provides an empirical roadmap for targeted qualitative item optimization. By utilizing frequency thresholds. it was possible to successfully isolate specific items requiring immediate structural refinement. For instance. items IT1 and IT2. as shown in Figure 7. were found to contain non-functional distractors. with alternative choices selected by fewer than 5% of the examinees. This signals that these distractors fail to act as plausible options and should be replaced to maximize overall option efficiency.

Figure 7
Overall Item and Distractor Frequency Distribution



An analysis of the point-biserial correlation coefficients (r_{pbi}) in Table 5—derived by condensing Table 4—reveals several noteworthy dynamics. Specifically. items IM3. IT1. and IT2 display options with strongly negative correlations. serving as indicators of specific student misconceptions (Bozzi et al., 2020). Regarding item IM4. a positive coefficient is observed for a distractor. which may reflect structural ambiguity in its wording. Similarly. item IE4 presents an interesting case with a correlation coefficient of zero; this lack of correlation not only places the option on the critical threshold for reclassification but may also point to ambiguous phrasing. It should be noted that these statistical findings do not deterministically mandate revisions but rather highlight the value of a qualitative peer review of the items.

Table 5
RPBI Values and Response Option Selection Percentages

Item	Option	rpbi_option	Type	%	Classification
IT2	A	-0.24	Distractor	24.3	Strong Misconception
IT1	D	-0.23	Distractor	21.9	Strong Misconception
IM3	A	-0.23	Distractor	40.6	Strong Misconception
IE4	C	0.00	Distractor	16.5	Functional Distractor
IM4	D	0.06	Distractor	23.3	Ambiguous/ Atypical

Examining the Option Response Curves (ORCs) for these items provides further insight. For instance. Figure 8 shows that option A of Item IT2 attracts a remarkably high number of students from the low- and mid-ability brackets. suggesting a strong underlying misconception. Furthermore. distractor D is selected by an insufficient number of students. rendering it non-functional. Conversely. Figure 9 reveals that distractor D of item IM4 tends to attract a greater proportion of students from the high and upper-middle ability levels than from lower tiers. Consistent with the positive correlation noted earlier. this unexpected pattern points to potential ambiguity in the wording. signaling that the option should be reviewed for potential revision.

Figure 8
Option Response Curve. Item IT2

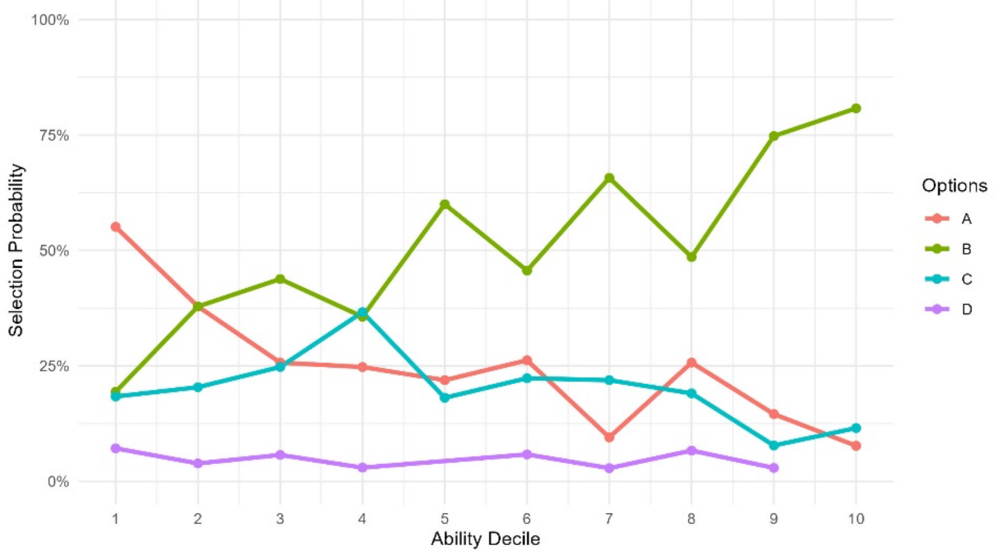
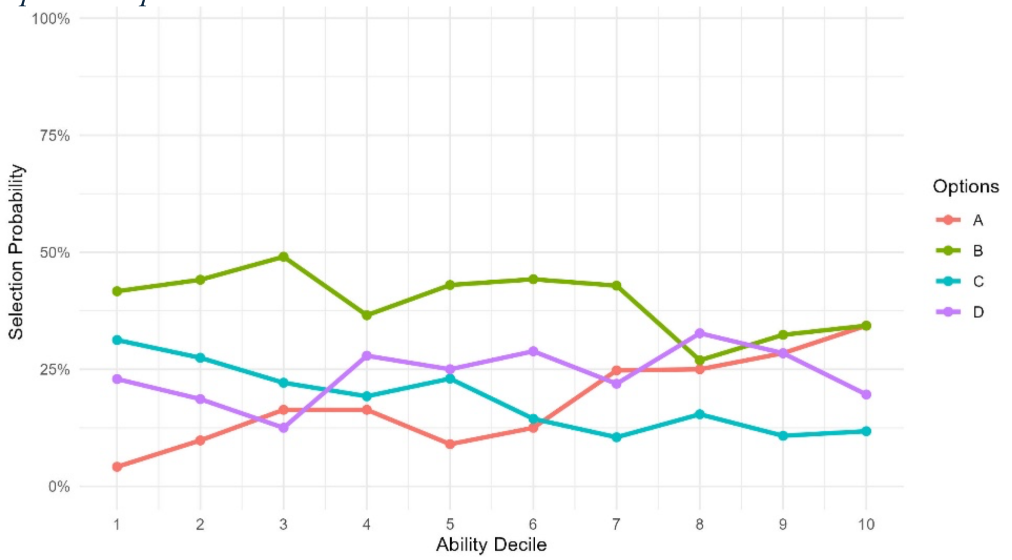


Figure 9
Option Response Curve. Item IM4



Conclusion

This study investigated the optimization and potential integration of the Politecnico-developed physics test with established international standards. specifically the Force Concept Inventory (FCI). to construct a robust evaluative assessment tool.

Through a systematic distractor efficiency analysis (RQa). we identified specific items—such as IT1 and IT2—with low-performing options falling below the critical 5% selection threshold. as well as items showing anomalous point-biserial correlations (r_{pbi}) or unexpected Option Response Curves (ORCs). like IM2. IM4. and IE4. Rather than mandating immediate. arbitrary changes. these psychometric indicators provided an empirical roadmap for a targeted qualitative peer review. isolating instances of structural text ambiguity and deep-seated student misconceptions.

Furthermore. cross-examining these items across international cohorts (RQb) revealed distinct. nationality-specific response patterns. These variations underscore the critical importance of cross-cultural validation when designing a standardized diagnostics item bank.

Ultimately. refining the non-functional distractors and mapping these cross-national differences aims not only to improve the internal validity of the Politecnico test. but also seeks to establish a more calibrated framework. When paired with the FCI. this optimized instrument may provide a comprehensive tool with the potential to yield deep. rigorous evaluative insights into university-level physics education across diverse instructional contexts.

Acknowledgments

The authors would like to sincerely thank the T.I.M.E. Association and. in particular. its Secretary General. Gwenaëlle Guillerme. for funding and supporting this project.

Special thanks are due to Patrizia Ghislandi (University of Trento. Italy) for her invaluable support and guidance regarding the psychological aspects of the questionnaire.

We also deeply appreciate the collaboration and availability of our international partners: Kazuhiko Tsukagoshi. Mami Matsukawa. Motoi Wada. and Naoto Nagaoka from Doshisha University (Japan). as well as Alexey B. Pnev and Andrey A. Zhirnov from Bauman Moscow State Technical University (Russia).

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The author declares that Grammarly. an AI-assisted writing software. was used in proofreading and refining the language used in the manuscript. The usage was limited to correcting grammatical and spelling errors and rephrasing statements for accuracy and clarity. The author further declares that. apart from Grammarly. no other AI or AI-assisted technologies have been used to generate content in writing the manuscript. The ideas. design. procedures. findings. analyses. and discussion are originally written and derived from careful and systematic conduct of the research.

References

Aslanides, J. S., & Savage, C. M. (2013). Relativity concept inventory: Development. analysis. and results. *Physical Review Special Topics - Physics Education Research, 9*(1), 010118. https://doi.org/10.1103/PhysRevSTPER.9.010118

Attali, Y., & Fraenkel, T. (2000). The Point-Biserial as a Discrimination Index for Distractors in Multiple-Choice Items: Deficiencies in Usage and an Alternative. *Journal of Educational Measurement, 37*(1), 77–86. https://doi.org/10.1111/j.1745-3984.2000.tb01077.x

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha: Table 1. *BMJ, 314*(7080), 572. https://doi.org/10.1136/bmj.314.7080.572

Bozzi, M., Ghislandi, P., Tsukagoshi, K., Matsukawa, M., Wada, M., Nagaoka, N., Pnev, A. B., Zhirnov, A. A., Guillerme, G., & Zani, M. (2019). Highlight misconceptions in Physics: A TIME project. *INTED2019 Proceedings,* 2520–2525. https://library.iated.org/view/BOZZI2019HIG

Bozzi, M., Ghislandi, P., & Zani, M. (2020). Misconceptions in Physics: An uphill climb. *INTED2020 Proceedings,* 2162–2170. https://library.iated.org/view/BOZZI2020MIS

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed). Prentice Hall.

Haladyna, T. M. (Ed.). (2004). *Developing and validating multiple-choice test items* (3rd ed). Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (1993). How Many Options is Enough for a Multiple-Choice Test Item? *Educational and Psychological Measurement, 53*(4), 999–1010. https://doi.org/10.1177/0013164493053004013

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. L. Erlbaum Associates.

Morris, G., Harshman, N., Branum-Martin, L., Mazur, E., Mzoughi, T., & Baker, S. (2012). An Item Response Curves Analysis of the Force Concept Inventory. *American Journal of Physics, 80,* 825–831. https://doi.org/10.1119/1.4731618

Murphy, R., & Broadfoot, P. (2017). *Effective Assessment and the Improvement of Education: A Tribute to Desmond Nuttall*. Routledge. https://doi.org/10.4324/9780203732618

Popham, W. J. (2020). *Classroom assessment: What teachers need to know* (Ninth edition). Pearson.

Contact email: andrea.fusari@polimi.it