11th International Conference on Higher Education Advances (HEAd'25)
Universitat Politècnica de València, València, 2025
DOI: https://doi.org/10.4995/HEAd25.2025.20062

Comparison between Rasch analysis and Classical Test Theory for a Physics questionnaire validation

Roberto Virzi, Matteo Bozzi, Marco Costigliolo, Maurizio Zani
Department of Physics, Politecnico di Milano, Italy.

How to cite: Virzi, R.; Bozzi, M.; Costigliolo, M.; Zani, M. (2025). Comparison between Rasch analysis and Classical Test Theory for a Physics questionnaire validation. In: 11th International Conference on Higher Education Advances (HEAd'25). Valencia, 17-20 June 2025. https://doi.org/10.4995/HEAd25.2025.20062

Abstract

In academic year 2022/2023, an orientation course was proposed to high school students. The aim of the course was to help students to prepare for an admission test to engineering faculties. In this context, the students took a physics propaedeutical test, and their responses were used to validate the questionnaire by applying the Classical Test Theory. Consequently, a new analysis has been performed using the Rasch model instead. In this paper, a comparison between the results of these two different approaches is presented. The outcomes of these methodologies seem to be in a good accordance, and the combined use of them can provide more reliable information about either the test and the sample of students. This is particularly important for small samples, such as the one in our analysis.

Keywords: *Rasch; CTT; Test; Orientation; Physics.*

1. Introduction

In the last years, the importance of orientation activities proposed to high school students has grown up. As a consequence, Italian universities improved their offer of courses to help students deciding about their future academic path. In this context, during the academic year 2022/2023 the Politecnico di Milano offered a very wide range of orientation courses. The size of the catalogue was made possible by funds related to the National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) (Decreto Ministeriale 934/22, 2022). According to PNRR requests, the courses had to span 15 hours and were to be designed with specific objectives. The main aim was to help students explore and understand the university environment, while also fostering awareness of their expectations and aspirations for future studies.

In this catalogue, there was a course run by the research groups ST2 and FDS of Politecnico di Milano named *Introduzione metodologica alla preparazione al test di Ingegneria*

(Methodological Introduction to Engineering Test Preparation). This course was structured in 3 different activities. 3 hours were dedicated to a frontal lesson about motivation, 6 hours was dedicated to Mathematics activities and 6 hours were dedicated to Physics activities. More specifically, our research group (ST2) provided the physics contents, and this paper is focused only on this part.

The course addressed students who intended to take the Engineering Test of Politecnico di Milano named *Test OnLine* (TOL) and who wanted to prepare to pass it. The main objective of this activity were not to provide physics contents in a traditional way. On the contrary we aimed to develop skills in order to make students able to study and improve autonomously.

Our 6 hours were structured as follow. 2 hours were dedicated to an online introductory meeting about study methods during which students were asked to answer an 8-item propaedeutic physics test. The remaining 4 hours were dedicated to a laboratory session. Globally, 113 students joined the proposal.

Concerning the laboratory session, we decided to design our activities following an open-ended style (Trumper, 2003). This approach allows students to feel more protagonist and to discover that Physics has a strong experimental nature (Wilcox & Lewandowski, 2016). In addition, the open-ended style laboratory can foster the science self-efficacy (Hu et al., 2022) that provide several positive aspects to the learning process (Alhadabi, 2021; Bandura, 1997; Hazari et al., 2010). Other details about the propaedeutical test will be discussed in further paragraphs.

The test proposed in the introductory meeting has been validated using 2 different methodologies: the Classical Test Theory (CTT) and the Rasch Analysis. The main aim of this work is to compare the results of these validations to highlight similarities and differences

In this paper we present in Section 2 the objectives of the propaedeutical test. In Section 3 we explain the theoretical background concerning the Rasch Analysis and we list our research questions. In Section 4 we describe the structure of the test and the methods of analysis. In section 5 we present the results. Eventually, in section 6 we discuss our results.

2. The Propaedeutical Test

During the introductory online lesson, students were asked to answer an 8-item multiple choice questionnaire and immediately after this test, the teacher provided a feedback. The main purpose of the test was not to evaluate the student or to rank them based on their scores. Indeed, the aim was to conduct a survey to make students aware of some gaps or misconceptions. In this context, the use of feedback is very important because helps students to reflect on their knowledge and learning methods exploiting a meta-cognitive approach (Romainville, 2006). The feedback given by the teacher was useful to make students able to recognize their errors and avoid them

in the future (Hattie & Clarke, 2018). In the opposite direction, feedback provided by students was useful for the teacher to understand why they chose a specific incorrect answer.

In addition, there is another aim that is not strictly related to the specific course presented in the Introduction. Designing and analysing this test allowed us to validate it and to build a database of good items to use even in other courses or situations. In fact, our research group is often engaged in experimental teaching activities. In this context, having access to a reliable item bank is very useful. An example of these activities is the peer-learning strategy we experimented in last years (Bozzi et al., 2021) but also other researches like (Bozzi et al., 2019, 2020, 2021, 2023; Gondoni et al., 2021)

In a previous work (accepted and waiting for publication in a Springer book) we validated this questionnaire using the CTT. More specifically, we calculated the Difficulty Level, the Discrimination Index and the Point-Biserial Coefficient (Ding & Beichner, 2009). The results of this analysis will be reported in Section 5 to make a comparison with the results of the Rasch Analysis

3. Rasch Analysis

If the Classical Test Theory assigns at each item of a test a coefficient that describes different characteristics (difficulty level, discrimination index and others), the Rasch Analysis is based on a completely different approach. As suggested by its name, this methodology was proposed by Georg Rasch (Rasch, 1960). In a nutshell, the Rasch model is a psychometric framework used for analyzing data from assessments, particularly in educational and psychological testing. It is part of the broader family of Item Response Theory (IRT) models but stands out due to its simplicity and strict mathematical properties.

At its core, the Rasch model evaluates the interaction between two factors: the person's ability (or trait level) and the item difficulty. It assumes that the probability of a correct response to an item depends on the difference between the person's ability and the item's difficulty. The model uses a logistic function to express this probability (Wright & Stone, 1979):

$$P(X_{ij}) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}} \quad (1)$$

Where $P(X_{ij})$ is the probability that person i answers item j correctly, θ_i represents the person's ability and $\beta_j$represents the item's difficulty. The Rasch model has several important features (Boone et al., 2014). First the *unidimensionality* that assumes that the performance depends on a single latent trait (e.g., ability). Another one is the *invariance*. This means that the item difficulty is independent of the sample and person ability is independent of the item set, if the statistics is sufficiently large. In addition, thanks to the *additivity*, both abilities and difficulties are expressed on a logit scale, enabling meaningful comparisons. The logit scale is a linear scale

used in the Rasch model, representing the natural logarithm of odds (log-odds) for probabilities, enabling comparisons of person ability and item difficulty.

Applications of the Rasch model include test design, item calibration, and detecting biases in items or groups. Its strict assumptions make it valuable for ensuring data quality and validity. When the data fit the model, it provides robust and interpretable measures that are invariant across contexts, enhancing the fairness and comparability of assessments.

In this paper we will answer the following research questions. The first (RQ1) is: Can the Rasch model be applied to our test? Then (QR2) is: Is the item difficulty level of the CTT similar to the Rasch model one?

4. Structure and Methods

The test was proposed to 113 students attending our orientation course in academic year 2022/2023. The 8 items were multiple choice questions with one right answer and three distractors. The questions cover different topics as shown in Table 1.

Table 1. Topics covered by each item.

Item	Topic
Q1	Force and movement
Q2	Friction
Q3	Circular motion
Q4	Kinetic and potential energy
Q5	Electric field and electric potential
Q6	Period and angular velocity
Q7	Free fall
Q8	Electrostatic force and Newton's third law

All these questions were designed to highlight possible misconceptions or gap in preparation. Indeed our items have the structure and the aim that could be similar to items in more famous tests like CSEM(Maloney et al., 2001) or FCI (Hestenes et al., 1992).

To validate this questionnaire, we applied first the CTT calculating some coefficients including Difficulty Level, Discrimination Index and Point-Biserial coefficient. The Difficulty Level measures how much a question has been difficult dividing the number of right answers by the total number of answers. This means that the frequency of correct answers to each item is used as an indicator for the difficulty of that item. The Discrimination Index is useful to evaluate how much an item is able to make differences among high-level students and low-level students. More precisely, it is calculated by sorting the students based on their final scores and comparing the number of students at the top of the leaderboard who answered correctly with the number of students at the bottom who answered correctly. The Point-Biserial coefficient is another measure

of the discrimination of an item. It provides an evaluation of the correlation between the score in a single question and the total score. We made all our calculations using the software R.

Consequently, we took into account the Rasch Analysis. In this case, to proper calculate the indices we had to rely on a specific R package named "eRm". This package contains functions that calculate all the parameters useful for the Analysis.

We started estimating how much the data fit with the Rasch model. First, we performed a Chi-square test observing that almost all the p-values were greater than 0.05. The only exception was 0.045 for the item Q7. Then we went deeper finding the values of the indicators named *Infit* and *Outfit.* The first, similarly to the Chi-square, provides a measure of how data fit with the model excluding the outliners points. The second one provides the same information but considering all the data. As suggested by (Bond & Fox, 2001; Karabatsos, 2000) the values of Infit and Outfit should be between 0.7 and 1.3 to state that there is a good accordance between data and Rasch model. The results will be presented in next section.

Then we wanted to compare the β values with the difficulty coefficients obtained using the CTT. To make this comparison we relied on the classification of Difficulty levels proposed by (Crocker & Algina, 1986) and we found the corresponding upper and lower bounds of β for each level. We did it using the (1) in which P is equal to the bound values of difficulty levels in CTT and θ is assumed to be zero. This means that in this calculation we approximated the ability of all students to the theoretical average ability of a group of persons. We also considered the standard error of β to find a minimum and a maximum difficulty level.

5. Results

As far as the CTT concerns, we report in Table 2 the results of the analysis we performed in our previous work.

Table 2. Coefficient of Classical Test Theory: Difficulty level (P), Discrimination Index (D) and Point-Biserial Coefficient (r).

Item	P	P-level	D	D-level	r
Q1	0.32	Medium	0.48	Good	0.45
Q2	0.38	Medium	0.56	Good	0.45
Q3	0.19	Medium-High	0.47	Good	0.51
Q4	0.28	Medium	0.58	Good	0.52
Q5	0.25	Medium-High	0.33	Reasonably Good	0.30
Q6	0.50	Medium-Low	0.50	Good	0.37
Q7	0.58	Medium-Low	0.55	Good	0.44
Q8	0.20	Medium-High	0.27	Marginal	0.27

To assign a category to each difficulty level and discrimination index we followed the structure proposed by (Crocker & Algina, 1986; Ebel & Frisbie, 1972). Concerning the Point-Biserial coefficient, all the values are above 0.2 so they can be considered "acceptable" (Kline, 1986). An additional information is that, considering all the students that answered the test, the average final score in this test was 2.67 up to 8.

Table 3 shows the Item Difficulty levels obtained with the Rasch model with the standard error of each item, the converted β-level scale and the values of Infit and Outfit for each item.

Table 3. Summary of results related to the Rasch model. The columns contain Item number, Difficulty Level, Standard Error, Minimum of converted β-level scale, Maximum of converted β-level scale, Infit indicator, Outfit indicator.

Item	β	SE_β	β-level (min)	β-level (max)	Infit	Outfit
Q1	0.07	0.20	Medium-Low	Medium	0.95	0.95
Q2	-0.27	0.19	Medium-Low	Medium-Low	0.99	1.01
Q3	0.78	0.23	Medium	Medium-High	0.80	0.64
Q4	0.25	0.20	Medium	Medium	0.87	0.74
Q5	0.39	0.21	Medium	Medium	1.06	1.23
Q6	-0.81	0.19	Medium-Low	Medium-Low	1.09	1.07
Q7	-1.11	0.19	Easy	Medium-Low	0.97	1.01
Q8	0.72	0.22	Medium	Medium	1.07	0.92

6. Discussion and conclusion

To answer the RQ1 we observed that in Table 3 all the Infit values are between 0.7 and 1.3. The same happens with the Outfit values with a single exception in Q3. This means that the model is applicable even if the number of students involved is not particularly high, as would theoretically be required.

Concerning the RQ2, we compared the Difficulty levels of CTT reported in third column of Table 2 and the ones of Rasch model reported in fourth and fifth column of Table 3. We notice that, in most cases, the categories overlap. We also observe that when β-level is different from P-level, the difficulty estimated with the Rasch model is always lower than the one obtained with the CTT. We think that this is probably due to the low level of preparation of students in our sample. Indeed, knowing that the average final score was 2.67 we thought that our students had a bad performance. Nevertheless, according to the validation we performed in our previous work, this test is not suitable for evaluating students' level, so we could not confirm this perception. From our point of view, it is therefore interesting to find that β, which does not depend on the sample, is actually lower than P. This can confirm the low level of our students.

Considering the results presented in this paper we can add another consideration. A sample of 113 students is not enough to completely trust neither the outcomes of CTT nor the Rasch model.

On the other hand, we found that in our work the data fit the Rasch model and the results are comparable with the CTT. This means that, in this case, running both analysis is useful to find more reliable information and come to stronger conclusions. We are waiting to verify this aspect more precisely using data from further editions, which involved a larger number of students.

References

Alhadabi, A. (2021). Science Interest, Utility, Self-Efficacy, Identity, and Science Achievement Among High School Students: An Application of SEM Tree. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.634120

Bandura, A. (1997). *Self-efficacy: The exercise of control* (pp. ix, 604). W H Freeman/Times Books/ Henry Holt & Co.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Psychology Press. https://doi.org/10.4324/9781410600127

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6857-4

Bozzi, M., Balossi, B., Sieno, L. D., Ganzer, L., Gondoni, P., Genco, I., Minnai, C., Pini, A., Rezoagli, F., Zanoletti, M., & Zani, M. (2019). Securing Freshmen's Learning Through A Physics Refresher Course: A Breakthrough Experience At Politecnico Di Milano. *ICERI2019 Proceedings*, 2237–2243. 12th annual International Conference of Education, Research and Innovation. https://doi.org/10.21125/iceri.2019.0610

Bozzi, M., Ghislandi, P., & Zani, M. (2020). Misconceptions in physics: an uphill climb. *INTED2020 Proceedings*, 2162–2170. 14th International Technology, Education and Development Conference. https://doi.org/10.21125/inted.2020.0670

Bozzi, M., Ghislandi, P., & Zani, M. (2021). Misconception in fisica: Un'opportunità di collaborazione tra università e scuola superiore. *Nuova Secondaria*, *XXXVIII*(5), 81–85.

Bozzi, M., Mazzola, R., & Zani, M. (2023). Peer learning in higher education: An example of practices. *Il Nuovo Cimento C*, *46*(6), 1–11. https://doi.org/10.1393/ncc/i2023-23206-7

Bozzi, M., Raffaghelli, J. E., & Zani, M. (2021). Peer Learning as a Key Component of an Integrated Teaching Method: Overcoming the Complexities of Physics Teaching in Large Size Classes. *Education Sciences*, *11*(2), Article 2. https://doi.org/10.3390/educsci11020067

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887 ($44.75). https://eric.ed.gov/?id=ed312281

Decreto Ministeriale 934/22, Pub. L. No. 934/2022 (2022). https://www.mur.gov.it/it/atti-e-normativa/decreto-ministeriale-n-934-del-03-08-2022

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

Ebel, R. L., & Frisbie, D. A. (1972). *Essentials of educational measurement*. https://www.academia.edu/download/62436844/Essentials_of_Educational_Measurement20200321-35262-g74v5m.pdf

Gondoni, P., Bozzi, M., Balossi, B., Calisesi, G., Ferocino, E., Genco, I., Molteni, L. M., Pini, A., Rezoagli, F., Zanoletti, M., & Zani, M. (2021). A novel approach to online physics refresher courses at Politecnico di Milano. *INTED2021 Proceedings*, 1471–1475. 15th International Technology, Education and Development Conference. https://doi.org/10.21125/inted.2021.0336

Hattie, J., & Clarke, S. (2018). *Visible Learning: Feedback*. Routledge. https://doi.org/10.4324/9780429485480

Hazari, Z., Potvin, G., Tai, R. H., & Almarode, J. (2010). For the love of learning science: Connecting learning orientation and career productivity in physics and chemistry. *Physical Review Special Topics - Physics Education Research*, *6*(1), 010107. https://doi.org/10.1103/PhysRevSTPER.6.010107

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

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, *1*(2), 152–176.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design* (pp. xi, 259). Methuen.

Maloney, D. P., O'Kuma, T. L., Hieggelke, C. J., & Van Heuvelen, A. (2001). Surveying students' conceptual knowledge of electricity and magnetism. *American Journal of Physics*, *69*(S1), S12–S23. https://doi.org/10.1119/1.1371296

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedagogiske Institut.

Romainville, M. (2006). *Awareness of cognitive strategies: The relationship between university students' metacognition and their performance: Studies in Higher Education: Vol 19, No 3*. https://www.tandfonline.com/doi/abs/10.1080/03075079412331381930

Trumper, R. (2003). The Physics Laboratory – A Historical Overview and Future Perspectives. *Science & Education*, *12*(7), 645–670. https://doi.org/10.1023/A:1025692409001

Wilcox, B. R., & Lewandowski, H. J. (2016). Open-ended versus guided laboratory activities:Impact on students' beliefs about experimental physics. *Physical Review Physics Education Research*, *12*(2), 020132. https://doi.org/10.1103/PhysRevPhysEducRes.12.020132

Wright, B., & Stone, M. (1979). Best test design. *Measurement and Statistics*. https://research.acer.edu.au/measurement/1