UNIVLORA SCIENTIFIC JOURNAL

Kushta, E. ,Kadena, F. , & Prenga, D. DOI: 10.63871....
UniVlora Scientific Journal 2025, no.I, volume II

Original research

Assessment of Difficulty Levels in Understanding Science Among Undergraduate Students Using the Rasch Analysis

Elmira Kushta [1], Florida Kadena[2], Dode Prenga[3]

[1] Department of Mathematics and Physics , Faculty of Technical and Natural Sciences, University "Ismail Qemali" , Vlorë, Albania
[2] Department of Physics, Faculty of Natural Sciences, University of Tirana, Albania
[3] Department of Physics, Faculty of Natural Sciences, University of Tirana, Albania
Corresponding author: elmira.kushta@univlora.edu.al

Article Type: Original research
Received: November 2025 | Accepted: December 2025 | Published: December 2025

Abstract

The assessment of students' understanding of the difficulty levels in physics, provided by Rasch analysis, is carried out here using two different tests as measurement tools. We used a polytomous version of the standard FCI test by selecting items that maintain the core of this conceptual inventory. To do this, we randomly divided the 380 responses collected during a recent FCI analysis into five groups and then restructured six items into six separate evaluation scale items. At this stage, we carefully selected items to highlight the dominance of one common-sense error in mechanics. Next, we created our own test with items designed to cover six levels of difficulty, alternating simpler calculation steps within physics-based problems. For both cases, the threshold parameter was estimated using the polytomous Rasch model. The results were analysed and discussed.

Keywords: *Concept Inventory, Physics, Knowledge, Index theory, The Rasch Model, Sociometric Measuring and Instruments.*

Introduction

The assessment of conceptual knowledge by using the Rasch model is a well-known and straightforward technique, see (Hestenes, 1992), (Planinic, 2010; 2019), etc. In the corresponding polytomous model, the threshold parameter measures the theoretical default level (Planinic, 2010) and can be seen as a key parameter of the conceptual knowledge. However, for a successful analysis, besides several statistical drawbacks for your system discussed in (Prenga et al, 2022), inevitable because of the heterogeneities in the system, a more general challenge consists of the

difficulties in drafting a Concept Inventory question with more than a binary assessment opportunity. Let motioned for clarity view aspect of the Conceptual knowledge tests are known also as Concept Inventory (CI) test, which have been initiated with the Force Concept Inventory (FCI) by (Hestenes et al., 1992), and advanced through numerous applications and theoretical works, (Crooks et al., 2014), (Andrich, 2005), (Savinainen, 2002)., etc. The original FCI test consists of a set of 30 multiple-choice items that are drafted in such a way as to measure the depth of knowledge in mechanics, contrasting to classical exercises that investigate students' understanding of a certain part of the subject (Prenga and Boçi, 2025). Also, the typical CI tests aim for investigating knowledge shortcomings, errors, and common-sense interpretation, and problems in learning and teaching in general, see (Hestenes et al., 1992) for the thorough analysis of the 'six common sense errors in mechanics, for example. Following those prerequisites, a polytomous variable in a Rasch model application must be based on a test constructed by a set of multiple-score items that can assess concept inventory knowledge and must include the potential for common sense errors. The first requirement can be concluded by setting up problems that involve students' critical thinking and active involvement, in terms of discussion in (Vilia et al., 2017), (Kim & Sunderman, 2005), (Kumar et al., 2023), etc. This is because we want to avoid calculations. After all, the problems would converge to a procedural exam, which is not typically conceptual knowledge. In this case, composing multiple-stage solution questions could be a choice, but in this case, we run into the problem of a knowledge report assimilation, because failing to pass a stage means a full failure, without having the chance to try to resolve the next step. This choice would contradict the CI idea that includes probability to resolve more difficult problems than students' abilities. Therefore, the realisation of a composite problem that keeps the CI is not a straightforward procedure. Instead, we might gather for this purpose some CI responses that are contaminated distinctively with at least one common-sense error, to produce a clone-item with several scales of evaluation. For this process, we have considered our recent FCI measurements referred to (Prenga et al., 2023), (Kushta et al, 2023), (Hamolli & Prenga, 2024), which have been conducted in the framework of proper CI analysis. The other opportunity was the mixed CI and procedural version, which poses several difficulties, mostly as a subjective or uncertified instrument. For this case, we considered a pilot version and properly checked the CVR index by asking several colleagues to evaluate before concluding the useful draft.

Data and Methods

Instrument Design and Content Validation

For drafting the test with n-score items, we have extended CI questions by asking basic calculations as well. We started initially with a 3-point evaluation for problems with very little effort, like the following

A. Given spheres (1) and (2), leaving the roof with zero and nonzero initial horizontal velocity, after leaving the roof,
(a) sphere (1) will reach the floor earlier
(b) sphere (2) and (1) will reach the floor at the same time
(c) sphere (2) will reach the floor a little later, just because of the air resistance
(d) sphere (2) will reach the floor little earlier, because the air resistance will slow it down due to the initial velocity, when sphere (1) has moved somewhat vertically.

B. Answer and fill the needed missing conceptual word. Which force is strongest when two objects interact? We made a real situation in the scene as follows: Which one is correct?
(a) The Earth attracts the Moon with a stronger force, because the Earth weighs more (0 point)
(b) The attracting force Earth-Moon F_{12} is exactly -F_{21} . It is related to the fact that both bodies are closed systems, but being a closed system is not a precondition. (1)

Original research

(c) The attracting force Earth-Moon F_{12} is exactly - F_{21} because according to Newton's third law of motion, the forces of the action and reaction are (complete) (2)
(d) The attracting force Earth-Moon F_{12} is exactly - F_{21} because according to Newton's third law of motion, the forces of the action and reaction are always equal and opposite (2). As a closed system, both moon and earth are falling toward a common centre with acceleration (complete)
(e) The attracting force Earth-Moon F_{12} is exactly - F_{21} because according to Newton's third law of motion, the forces of the action and reaction are always equal and opposite (2). Therefore, the moon is falling toward Earth, and Earth is falling toward the Moon with (Specify) acceleration

Initially, we sent our ad-hoc test to a group of 11 colleagues made up of experienced teachers, university lecturers of general physics, and professors of didactics for evaluation. It was a low-scale check, but we decided to count their assessment on the proposed instrument. It resulted that the Content Validity Ratio calculated after their review was low, at 0.4. In this first attempt, we have realised 5 items like question A above and 5 similar question B, belonging to different scores per item. For the 5-level item's test, the CVR was lower than the item based on the 4-level item, like A. This suggests considering also the difficulty level of the question. In the meantime, we collected the remarks of our colleagues and realised an improved version of both types of tests for further check, this time by a group of teachers. In this case, we let the teacher suggest their own scores. The test has been sent to several teachers for evaluations, and we gathered responses from a couple of them, not statistically significant, but suitable for addressing the problem. In this stage, we realised that teachers assigned different scores to the alternatives of the items, and the situation was worse as the complexity of the items increased. We have highlighted the facts that there are difficulties in reaching a consensus on how to distribute points in a CI-based question with multiple scores. It suggests that, aside from many factors, the general tendency of the moment would point to the practicality of teaching and learning, which is mirrored differently in different schools of the country. The heterogeneity of the evaluation indicates ana heterogeneity in the conceptual knowledge focusing when teaching. Therefore, the assessment of conceptual knowledge with a polytomous variable based on an ad hoc test is quite unlikely to succeed for the moment. This is why we have tuned back to the recent idea suggested by (Prenga 2025) to produce a fictitious test based on recent FCI tests. In this application, we have considered a full FCI set of 380 answers and initially divided it into 5 groups. Next, we have analysed the taxonomy of common-sense error and identified the diamond error type based on the analysis and tables provided in (Hestenes, 1992). By randomly choosing six items of different errors, we produce a clone question of 6 difficulty levels in the sense that summing up the scores would range from 0-6. This clone question bears the ADN and philosophy of the FCI test. Collecting all clones produced so far, we reached 30 items to mimic the FCI size, but it was not necessary for sure. We believe that this last is adequate for measuring and consider it for final use.

Rasch Model and Statistical Analysis

The Rasch analysis is a well-known technique for analysing and calibrating sociometric measurement. Considering the result of the CI test consisting of N_{items} conducted $N_{responders}$, initial answers are organised in a matrix $T(i,j) =$ (0,1,2,3,4,5). The procedure of data elaboration is similar to the binary variable addressed in a large literature theoretically, see (Rasch, 1961), (Andrich, 2005), (Wright, 1982; 1977), and by measurements, see (Savinainen, aninic, 2010) etc. So, students' success is calculated initally $\frac{SuccessProbability}{FailureProbability} = \frac{p_i}{1-p_i}$ and similar is the easiness $\frac{p_j}{1-p_j}$ of the whole test, and next based on the IRT the probability of success given by the logistic

Original research

function $p(1,y) = logit(y) \equiv \frac{e^y}{1+e^y}$, so the student's ability to resolve the test and item's dif-ficulty for dichotomous variable are given by the following student's ability to resolve the test and item's difficulty for dichotomous variable are given by the following

$$\alpha_i = ln\frac{\mathrm{p}(i)}{1-\mathrm{p(i)}} \equiv \ln\frac{a_i}{1-a_i};\ \delta_j = ln\frac{1-\mathrm{p(j)}}{\mathrm{p(j)}} \equiv ln\frac{1-e_j}{e_j} \quad (1)$$

and therefore, the estimated probability:

$$P_e(i,j) \equiv P(\alpha_i, \delta_j) = \frac{exp\,(\alpha_i-\delta_j)}{1+exp\,(\alpha_i-\delta_j)} \quad (2)$$

Similarly, but with some effort in defining difficulties and abilities that are now related also to the inner structure of the item alongside the whole test responses, the polytomous Rasch model analysis. The analogues of (2) above would evaluate the probability that a student would achieve the score level h, given the threshold parameter $\{\tau_j\}$ that measures the difficulty of obtaining scores h relative to h-1, by using the following estimate probability for success

$$P(x_{ij} = h|\beta_i, \delta_j, \tau) = \frac{exp[k\left(\beta_i^{(k)}-\delta_j^k\right)-\sum_{j=0}^{h}\tau_j]}{\sum_{k=0}^{m} exp[k\left(\beta_i^{(k)}-\delta_j^k\right)-\sum_{k}^{h}\tau_k]} \quad (3)$$

Here in the model parameter τ it is very important because it presents the structure of knowledge of the population under study. If statistical prerequisites were fulfilled, the results are intriguing because they report the knowledge structure on a larger scale, indicating problems and challenges for the science education system itself. Being aware that our clone test merits some legitimacy arguments based on the discussion provided in (Prenga et al, 2025), etc, which we are not reproducing herein, the following analysis reveals interesting findings and interpretations.

Results

After analysing the ability and optimising the histogram, we have observed that mostly there are 5-6 levels of the measured ability. In this logic, we estimated that also, perceived difficulties are also in the same structure. On the other side, we have decided the levels of difficulty would match the grades used in our system, so a choice of 5-6 levels of the default is logically supported. Also, we approached our system of knowledge evaluation based on ten grades, and considering that very low levels covering grades 1-4 might be assimilated in a unique level, the 6-level approach seems logical. Basically, we are based in this last idea to construct 6-score items for our test. Notice that for the whole set of FCI test results, we have found a large magnitude between difficulty levels [−1.4450 ; 4.4269 ;10.2989;16.1708 ; 22.0427], which does not match with ability histogram, indicating additional problems with sampling. To shed some light on the heterogeneity indicated by this finding, we explored several sampling trails seeking significant signals regarding the consistency of the finding. It resulted that, at last, qualitatively, the results obtained were stable, Table 1.

Table 12. Threshold parameter measured by the FCI-cloned test

Difficulty threshold	Sample:100 random records from 280 responders (2022)	Whole interviews, 280 Students	Sample: 100 random records 100 from a set of 360 responses, (2023)	Sample: 180 students (2023)	Sample: 53 students (2024)
	0	0	0	0	0
	-3.1083	-4.1788	-2.9319	-3.2483	-3.3252
	-0.733	-0.4317	0.1203	-0.6582	-0.5847
	0.5574	1.53	1.4429	0.9692	1.6345
	3.2839	3.0804	1.3687	2.9373	2.2754

We observed that for all groups selected, the first element of the threshold vector has a very high magnitude compared to the others, and so does the last one. The consistency of these features suggests that the measurement is acceptable and mirrors some representative characteristics of the whole system. By contemplating formal interpretation of the threshold difficulty on the polytomous Rasch model, we concluded that, from a general perspective, the effort needed to bring very basic students to the admissible level or to push good students to the excellent level is much higher than moving between other knowledge levels. Notice that this estimator is not the same as the difference between difficulty stages $\Delta_i = \delta_{i+1} - \delta_i$, because the definition of the difficulty parameter on the polytomous model is local and depends on the overall estimation of each item, neglecting its subdivisions in n-levels. This said, natural levels of the overall knowledge must be determined by referring to the abilities measured, because one cannot make statistics with a very small number of items used for testing. As a result, natural difficulty levels do not represent the 10 grades or 55-gradesystem. By optimising corresponding histograms, we can analyse this feature following the same idea as in (Prenga, 2024), and occasionally those levels matched in our case, but the results are not interchangeably symmetrical, nor identical. This is basically a result of the nature of output quantities of the Rash model, where the Δ_i would measure the perceived difficulty of the whole test made of 30- questions each of 6 levels in our case, whereas τ_i measures theoretical differences between two successive levels at coordinate (i). Coming back to Table 1, we also observe that the threshold levels are nearly symmetrical regarding the zero point. This mirrors the consistency of the test used, but we do not expect otherwise, as long as the values are gathered from a very standardised FCI test. Bot adding to that, the symmetry of the observed parameter is seen herein as an indicator of the correspondence in the whole system of the compression and education.

A comparative view of results and further discussion

By comparing these results with recent measurements conducted across several student groups, as presented in Table 2, we observe a similar qualitative pattern for the threshold parameter at the lowest proficiency level. It is worth noting that Table 2 employs a 4-level scale, which does not differ substantially from the 6-level scale used in the current analysis

Table 13. Threshold parameter for various groups of students

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
0	0	0	0	0	0
-2.32	-2.23	-2.05	-1.59	-1.87	-1.87
-0.86	-0.87	-0.83	-0.58	-0.87	-0.947
0.51	0.31	0.85	0.37	0.49	0.58

It appears that the first difficulty level is the most challenging, as reflected by the high amplitude of its corresponding threshold parameter. The findings presented in Table 2 are interpreted according to the standard definition of threshold difficulty in the Rasch model (see www.realstatistics.org for illustrative applications). Furthermore, the measured

difficulty thresholds can be related to the practical effort required in teaching. Since the Rasch scale is standardized and linearized, the differences between successive thresholds serve as indicators of the relative effort needed to advance students by one conceptual level. By convention, these differences estimate the effort required to improve knowledge from one level to the next. They primarily reflect the relative effort for enhancing students' conceptual understanding, while also providing insight into the efficiency of instructional investment.

So, when we acknowledge our absolute effort to change the knowledge situation, that is the vector $\tau = [\tau_i]$, we might compare those values to choose the opportune scenarios. Notice that we cannot exaggerate with the homogeneity and linearity of the difference parameters, but locally, we believe that the comparison is logically supported. Therefore, the new estimator $\Delta\tau_i = \tau_{i+1} - \tau_i$ measure the differences between the effort for improving the knowledge by one unit, that is, the opportunity effort for working for improving level (i). A similar analysis of this parameter has been conducted by Prenga (2024). In general terms, it provides a measure of the homogeneity of the effort required to improve students' knowledge levels. While this metric is related to the concept of "effort," we have emphasized its interpretation for clarity. Table 3 presents the values of this new relative parameter. As indicated by the measured differences in threshold difficulties, the relative effort required to improve students at the lowest proficiency level is higher than that needed for students at intermediate levels. This finding highlights the unequal distribution of pedagogical effort across different knowledge levels and suggests that targeted strategies may be more efficient when focused on students with medium-level knowledge.

Table 14. Relative absolute differences between threshold levels

Sample: 100 random records from 280 responders (2022)	Whole interviews, 280 Students	Sample: 100 random records 100 from a set of 360 responses, (2023)	Sample: 180 students (2023)	Sample: 53 students (2024)	Sample: 100 random records from 280 responders, (2022)
2.3753	3.7471	3.0522	2.5901	2.7405	2.90104
0.733	0.4317	0.1203	0.6582	0.5847	0.50558
0.5574	1.53	1.4429	0.9692	1.6345	1.2268
2.7265	1.5504	0.0742	1.9681	0.6409	1.39202

These results suggest that, from a pragmatic perspective of cost-efficiency in pedagogy, it may be more effective to focus educational strategies on students at intermediate proficiency levels, rather than attempting to elevate students with very low conceptual knowledge. Similarly, efforts to advance students at the highest proficiency levels require disproportionately more resources compared to interventions aimed at students in the medium range .

While acknowledging certain statistical limitations that prevent definitive generalizations, these findings should be considered indicative. We recommend that researchers and educators apply this approach, along with other IRT-based techniques, to inform evidence-based instructional practices and educational research.

Conclusions

Original research

The use of Concept Inventory (CI) tests can be successfully extended to assess specific features of conceptual knowledge. While the development of ad hoc CI instruments remains a viable and appealing approach, particular attention must be paid to the formal definition and consistency of scoring schemes. The present study indicates that combining dichotomous and polytomous scoring algorithms can help overcome several practical limitations, while maintaining students' engagement over extended testing sessions. In this context, carefully designed CI-based polytomous questionnaires may provide valuable tools for educational research.

Through direct measurement of latent traits associated with conceptual knowledge, we estimated threshold difficulty levels within a six-level scoring framework. The results show that these difficulty thresholds are not uniformly distributed, suggesting that larger-scale measurements would further strengthen the robustness of the findings. Our analysis highlights the Rasch threshold parameters as meaningful descriptors of the underlying difficulty structure of conceptual knowledge.

Accordingly, the pedagogical effort required to promote learning gains is more efficiently directed toward students at intermediate proficiency levels, as the lowest and highest threshold transitions require substantially greater effort than transitions within the medium range. Given certain limitations in the statistical prerequisites for broad generalization, we encourage researchers to apply this approach, together with other Item Response Theory–based methods, for quantitative assessment and evidence-based inference in didactic and pedagogical research.

References

1. *Andrich D. The Rasch model explained. In: Applied Rasch measurement: A book of exemplars: Papers in honour of John P. Keeves. 2005. p. 27–59.*
2. *Crooks NM, Alibali MW. Defining and measuring conceptual knowledge in mathematics. Dev Rev. 2014;34(4):344–377.*
3. *Ding L, Chabay R, Sherwood B, Beichner R. Evaluating an electricity and magnetism assessment tool: Brief Electricity and Magnetism Assessment. Phys Rev ST Phys Educ Res. 2006;2(1):010105.*
4. *Hamolli L, Prenga D. Force concept inventory analysis by using indexes and the Rasch model. J Nat Sci. 2024;36:210–230.*
5. *Hestenes D, Wells M, Swackhamer G. Force concept inventory. Phys Teach. 1992;30(3):141–158.*
6. *Jones M, Jones G. Biologjia 10–11. Tirana: Pegi; 2016.*
7. *Kim JS, Sunderman GL. Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. Educ Res. 2005;34(8):3–13.*
8. *Klymkowsky MW, Garvin-Doxas K. Concept inventories: Design, application, uses, limitations, and next steps. In: Active learning in college science: The case for evidence-based practice. 2020. p. 775–790.*
9. *Kumar P, et al. Using empirical science education in schools to improve climate change literacy. Renew Sustain Energy Rev. 2023;178:113232.*
10. *Kushta E, Prenga D. Using tests' indexes to improve the assessment of the conceptual knowledge: A case study. Int J Educ Learn Syst. 2023;8.*
11. *Planinic M, Ivanjek L, Susac A. Rasch model-based analysis of the Force Concept Inventory. Phys Rev ST Phys Educ Res. 2010;6(1).*
12. *Planinic M, Boone WJ, Susac A, Ivanjek L. Rasch analysis in physics education research: Why measurement matters. Phys Rev Phys Educ Res. 2019;15(2):020111.*
13. *Pople S. Fizika 10–11. Tirana: Erik Botime; 2022.*
14. *Prenga D, Kushta E, Musli F. Enhancing concept inventory analysis by using indexes, optimal histogram idea, and Likert analysis. J Hum Earth Future. 2023;4(1):103–120.*

15. *Prenga D. A thematic review on the combination of statistical tools and measuring instruments for analyzing knowledge and students' achievement in science. Eur Mod Stud J. 2024;8(3):687–706.*
16. *Rasch G. On general laws and the meaning of measurement in psychology. In: Proc 4th Berkeley Symp Math Stat Probab. Berkeley: University of California Press; 1961. p. 321–333.*
17. *Savinainen A, Scott P. The Force Concept Inventory: A tool for monitoring student learning. Phys Educ. 2002;37(1):45–51.*
18. *Vilia PN, Candeias AA, Neto AS, Franco MDGS, Melo M. Academic achievement in physics-chemistry: The predictive effect of attitudes and reasoning abilities. Front Psychol. 2017;8:1064.*
19. *Wright BD. Solving measurement problems with the Rasch model. J Educ Meas. 1977;14:97–116.*
20. *Wright BD, Masters GN. Rating scale analysis. Chicago: MESA Press; 1982.*