ISSN 2522-9400 **European Modern Studies Journal** Vol 8 No 3
available at www.journal-ems.com
DOI: 10.59573/emsj.8(3).2024.42

A Thematic Review on the Combination of Statistical Tools and Measuring Instruments for Analyzing Knowledge and Students' Achievement in Science

Dode Prenga
Department of Physics, Faculty of Natural Sciences,
University of Tirana, Albania

Abstract. To overcome difficulties encountered in the analyses of some factors related to the students' knowledge and for avoiding expensive or difficult to realize tests, we have combined herein particular features of instruments used for measuring the knowledge and for controlling the quality of the testing, with some statistical tools. Initially we propose to extend the metrics of the standard instruments used for measuring knowledge and testing reliability, as the Rashc model and indexes theory. In this framework, the features of indexes of a certified Concept Inventory test are recognized as responses to specific factors, including latent ones, which affect the overall knowledge state. Specifically, by such a straightforward analysis we estimate the quality of the teaching efficacy in physics, which is not a directly measurable quantity by standard tools. Similarly, some results of the Rash analysis for those certified tests, such as the misfitted occurrences and guessing behavior, are treated as auxiliary indicators of knowledge state and are used for analyzing the cause factors which affect it. Also, the threshold parameter appearing on the polytomous Rasch procedure is considered for evaluating the effort needed to improve the tests' difficulty perceived by students, and next as a measure of the possible academic achievements and proficiency that can be attained through an appropriate improvement of the learning conditions. This idea is advanced by employing the features of the histograms and distributions of students' abilities calculated by the Rasch technique. We used for those purposes several certified CI tests in certain groups and circumstances to mimic different initial condition of cause factors, and analyzed similarities and dissimilarities of the outcomes of the Rasch analysis' seen as the system's responses. By comparing their results, we achieved a better evidencing of problems on the efficiency of teaching and learning fundamental sciences. Also, the combination of different tools is seen useful to improve the resolution of standard instruments of knowledge measurement. Even though the illustrations of those ideas consist of some particular case -studies, the technique proposed herein is believed to be of a more general nature and can be used for analyses in similar circumstances.

Keywords: concept inventory, physics knowledge, the Rasch model, Likert scale

Introduction

The assessment of students' conceptual knowledge and their progress in understanding sciences is realized in practice by using measurement instruments known as Concept Inventory tests, which are based on FCI methodology introduced by (Hestenes, 1992) and developed further on in various categories, (Sands, 2018; Handhika, 2017), etc. Those tests are adjusted by employing calibrating techniques like Rasch analysis introduced in (Rasch, 1964) and can be tuned by using special methods like indexes analysis, presented in (Ding et al, 2006). Other elements of knowledges regrading students' ability to resolve problems and for employing scientific knowledge in practice and applications can be evaluated by using purposed procedural tests and by processing the information gathered through official data and exams. On the other side, for evidencing and analyzing the factors which contribute on, or affect and conditionate students' knowledge and their academic progress, each of that measurement instrument is not a priory sufficient and alternative solution and complementary tools are needed. Another problem regarding knowledge evaluation is related to the measurement procedures that cannot be conducted in full compliance with statistical requirement, to measuring several knowledge features that needs for repeated or paired tests, or for conducting tests in specific circumstances which are difficult to realize in practice. Before presenting some practical suggestions and illustration, let's consider an overview on the complexity of factors that affect the knowledge, and some elementary introduction for the knowledge's measurement instruments or techniques. In present time, and specifically for Albania where we have conducted some measurements and analyses, the students 'studying focus and their learning methodology varies among groups and schools' profiles (Prenga et al 2022), which make very difficult a thorough factorial analysis for explaining knowledge outcomes. Also, the interest of students on fundamental sciences such as physics and mathematics has been modified for a wide set of factors and reasons, complicating further the attempt for investigating factors that affect students advances in sciences. Evidently, this complexity of the influences penalizes the capability the official exams for measuring their effects, and also reduce the resolution of the measurement itself. As an alternative way-out for evidencing problems on education system, students' progress in understanding sciences etc., one can use the outcomes of independent testing on conceptual knowledge, which are known for their accuracy of measurement. The measurement of the conceptual knowledge through concept inventory tests (CI) is employed routinely for studying or analyzing education problems, (Crooks & Alibali, 2014; Planinic et al, 2010, 2019). A large data set of CI prototypes and analytic methods regarding them can be consulted in the dedicated pages like (www.physport.org), or on literatures cited throughout this work. Those tests become very important nowadays, when the tendency toward algorithmic solution of the problems is gaining terrain on the present education methodologies in the high schools, mostly dictated by the pragmatism related to the next stage of studies. Dislike the typical prevalence of the achieved scores as the measure of knowledge in the procedural test, the concept inventory test is oriented also in evidencing shortcomings, knowledge's failure or common errors in general understanding of the science etc., which are closer to concerns of education's analysis for research purposes. Therefore, it worth mentioning some aspects of CI analysis herein. A CI test is basically a multiple-choice test, containing a set of basic and well-defined concepts' check through several items which are formulated correctly and aims at revealing specific features of the conceptual knowledge. Basically, an item of a CI test like the Force Concept Inventory test aims on evidencing student's knowledge as per their fundaments, and not a simple solution for a particular problem. It looks like the whole is asked in a single question. In this regard, every item plays the role of a projection of knowledge for the whole discipline, therefore, when students fail to give a correct answer, it means that this failure characterizes the knowledge for the whole matter. Those specifics convince the CI test as an instrument for measuring several knowledge features, but additional efforts are aimed for guaranteeing its resolution power. So, we will consider herein some opportunities for extending the use of CI test in the study of latent factors which contribute on knowledge formation. Remember that the quantitative analysis of the CI tests outcomes and especially its calibration is achieved in practice by using a special sociometric model proposed by Danish scientist G. Rasch, (Rasch, 1964), so we will introduce some elements of this model below. Also, an indirect calibration procedure can be made through the index's analysis, see (Ding et al, 2006, Gliem, 2003) and references therein. The analysis of the tests' indexes is very important for investigation of the reliability and usefulness of the uncertified tests as for example the ad-hoc CI test and the common exams which we prepare routinely in our pedagogic activities. But a very intriguing application could be based on a retrograde view: the verified in-appropriateness of the indexes for a certified CI test (FCI for example) would mirror knowledge failures among students, so we can explore for the latent factors behind the appearing heterogeneities immediately after evidencing them through indexes. Note also that measuring academic proficiency and achievement by CI tests is debatable, because those features of the knowledge encompass critical thinking, practical skills and performance etc., see (Kim et al, 2005) for details of this category, which are not among CI test outcomes. Alternatively, the procedural test, which measures the ability of a student to solve problems step by step, his (her) fluency to employ instructed algorithms etc., can be helpful for measuring academic proficiency and achievement, despite some natural limitation. It followed that an adequate instrument for measuring academic progress must include both types of the tests, but not limited to. This is because of the multidimensionality of students' ability to use achieved knowledge and skills to solve real problems, getting a clear understanding of basic concepts of the syllabus, etc., as by definition of the academic proficiency. It can be evaluated through the well-known PISA test or by employing dedicated test as the VALUE test -Valid Assessment of Learning in Undergraduate Education (Rhodes, 2008), but with the price of extensive costs and means needed. For more details and discussion for those on-purpose testing, see also (York et al, 2019) etc. However, the dedicated tests like PISA, might fail to be an accurate instrument for measuring academic achievement and proficiency, if participating students do not know its proper structure. We made an explanatory pre-testing for this element, and interestingly, we observed that students use to calculate things that were provided by the platform itself, and were to be selected among alternatives only! More than 80% of students which participated in the pre-tests lost their time dealing with unnecessary things and being defocused from the main questions, and 65% of them declared that they do not own a PC or laptop in their house. We believe that that those factors have been decisive for the very low level of Albanian students on the PISA test in 2023. Therefore, when considering those standard tests for measuring academic advances, an additional precaution and preliminary analysis should be made for avoiding uncertainties of the measurements. Moreover, and strictly related to our focus in this review, a classical assessment for students' achievement based on the applicative competencies, seems un-appropriate for high schools' students for a couple of reasons. On the other hand, gathering information about this knowledge feature is a very important from pedagogical and research perspectives. So, students' academic achievement in the high school plays a crucial role in the next stage of their studies, and which the most important, it affects directly their preferences for engineering branches, for education in physics and mathematics, at least for our system, (Prenga et al, 2023). However, the surveys and tests based on voluntary participation are important and indispensable for a thorough knowledge of education problems. Those tools must be improved and developed to deal with weighting factors that behave as latent and remain partially hidden from the standard optics of the investigative tools or below the resolution capabilities of the standard instruments. Those features that disqualify a single instrument for measuring specific features of the knowledge, but firmly accept the importance of a specific measurement, suggest the use of combined instruments, extension of their metrics and use of additional statistical tools. So, the idea for harmonizing of the features provided by VALUE test structure, regarding the combination of the quizzes, exams, reports, research projects, course evaluation, surveys etc., is intriguing for this purpose, and we propose to employ this philosophy for measuring intermediate or hidden variables in our system. Firstly, we proposed to start by using certified CI tests (like FCI) and the corresponding analysis, which are easy to handle. Next, several auxiliary tools can be adopted to improve the outcomes. Following this idea, we have conducted several low scales or sampled-based CI tests on high school students (Kushta et al, 2023; Prenga et al, 2022), whom indicated that, aside of the pragmatic decreasing of the interest on studying physics alike in other developing countries, see (Vilia et al, 2017), several inadequacies on teaching physics programs on high schools are present in the system. Those defects are exhibited as a lack of the preferences for studying physics, for example, (Prenga et al, 2023). Also, by analyzing our recent findings on CI analysis, we confirmed a tendency of our students to perform better in procedural exams than in conceptual tests. It is likely triggered by the dominancy of the procedural nature on our common exams, but from the behavioral point of view, it has the potential to drive the learning process toward the algorithmic problem solving, affecting so the conceptual knowledge, academic proficiency, and critical thinking in a large community of students. It makes the assessment of academic achievement and a thorough knowledge analysis more complicated, which inspire the combination of the methods proposed in this work, for analyzing more generally the efficacy of the knowledge transfer. Finally, for conducting typical analyses of this category, we need for pre and post-test surveys, large sample size, and full randomness of the sampling, which are expensive and time consuming. To overcome those problems and for increasing the capacities of common investigative tools for evidencing hidden factors that influence the knowledge, we have re-considered the CI outcomes, indexes' analysis and combination of common statistical tools of descriptive analysis. In the following we will describe those ideas by illustrated examples and applications.

An Indirect Analysis of the Knowledge by Using Indexes of the Certified CI Tests

The use of CI testing is a normal routine for studying conceptual knowledge in physics. In the first stage of the analysis, test's results must be checked for the integrity, reliability, significance and validity. For this purpose, the most utilized estimators are the test's indexes namely *difficulty index*, *discrimination index*, *reliability index*, *self-consistency*, *discriminatory power index etc.* (Ding et al, 2006; Aubrecht et al, 1982), see Table 1. We proposed using them also as indicator of student's concentration on the answering the test and overall knowledge, providing that the sample fulfill statistical prerequisites and the CI used is a standardized and a certified exemplar, (Prenga et al, 2022; Kushta et al, 2023). We suggest to use indexes also for identifying hidden factors that influence the knowledge transfer and its advances, by comparing results of different tests. Based on the description of the indexes in (Ding et al, 2006; Glim, 2003), we emphasize that the observed differences between current tests' indexes and their plausible values mirror the influences of factors which act by the intermediacy of the realization process of the testing, as the clearness and understanding of an item, face validity issues, the coverage of all subjects included in the test, contextual knowledge problem etc. Now, by using a certified CI test like FCI for example, we exclude the potential of test's editing problems, so we can compare directly the outcomes which would mirror the differences and specifics of the learning circumstances and features for the group of students being tested. So, by hypothesis, if teaching of mechanics is realized according to the syllabus, every student is expected to perceive a moderate difficulty for a given FCI item. Alternatively, if this item is perceived very difficult by all students of the sample, it suggests a failure on transmission of the knowledge, indicating problems in teaching stage for this group. Using those indicators, we may hypothesize that 'the indexes' inadequacies for a certified CI tests are related to the teaching phases' which in turn, would be accepted or rejected through the analysis of the arguments provided by other investigative tools or measurement instruments. Similar arguments can be used for explaining other index's inadequacies regarding general agreements which are provided in the last row of Table 1. In this stage we suppose to evidence some signals for a failure in teaching process which is paired with learning stage and understanding, and is expressed on the abnormal value observed for this index of the item, or for the whole FCI test.

Table 1. A summary for tests' indexes. Legend: I indicate test item, N is the number of students, n_i are the number of correct answering of item (i), f are frequencies, σ are standard deviances, $\tilde{x}$ are the respective means

Indexes	Formulae	Desired Range	Description for out of range
Difficulty index	$\delta_i = \frac{n_{Correct}^{answer}}{N}$; $\delta_{test} = \frac{1}{n_i}\sum_{i=1}^{n_i} \delta_i$	0.3-0.9	Abnormally difficult or easy item/or test
Discrimination index	$d = \frac{(N_{highestScore}^{Succses} - N_{lowestScore}^{Succes})}{N}_{p\%}$ $D_{Test} = d_{avg}$	>0.3	Item(s) is not effective to discriminate well between students with high or low total scores.
Reliability index (Cronbach alpha)	$r_i = \frac{\tilde{x}_i - \tilde{x}}{\sigma(x)}\sqrt{\frac{\delta_i}{1-\delta_i}}$;$r_{avg} = \frac{1}{n_i}\sum_{i=1}^{n_i} r_i$	>0.2	Responses are not based on solid knowledge and uncertainties are influent. When repeating the test later probably different scores are obtained.
Discriminatory power	$\rho_{test} = \frac{n_i}{n_i - 1}(1 - \frac{\sum^{n_i} \sigma(x_i)}{\sigma^2(x)}$	>0.7	The test is not capable to discriminate between different abilities.
Self-consistency index	$\Delta = \frac{N^2 - \sum_i^{n_i} f_i^2}{N^2 - \frac{N^2}{n_i+1}}$	>0.7	Tells how broadly the total scores of a sample are distributed over the possible rang.

In the following, for illustrating the idea, we have presented some findings based on the comparison of the indexes for different certified CI tests. Those tests have been carried out on different groups of undergraduate students enrolled in the Faculty of Sciences of Natures, and on high school students in different districts during 2023-2024. Notice that there is not a distinguishable change between conceptual knowledge in mechanics in the end of high school and bingeing of the first year of the bachelor at university. By nature, the program of mechanics provided for all students being interviewed was the same, and as supposed to have been realized completely. In this regard, the differences in indexes observed for the tests would be related to specific education circumstances of each group. For an extended comparison, we have considered also an ad-hoc test on basic Climate Knowledges. This last has been made up of a set of standard questions edited carefully based on the programs of disciplines of sciences in high school, on illustrative examples and textbooks used by students etc. The significant difference between this test and CI tests in physics is highlighted: The climate knowledge on u high school is provided through a distributed system of lecturing and examples and basically unstructured one. The idea is that similarities on the indexes between indexes of the two types would suggest for non-functionality on the teaching of structured disciplines. In Table 2 are shown indexes calculated for several CI standard tests (FCI and BEMA) and the Climate test. Notice that we have adjusted further to this ad hoc test by filtering the items with high outfitted ratio r ($r = \frac{(x_{inital} - x_{estimate})^2}{var(x_{estimate}}$) calculated using Rasch analysis for the initial tests of 30 questions. Finally, the accustomed climate test contained 22 elementary questions, is considered as equally good as certified tests. It can be viewed also as similar to those certified, but carried out in different teaching circumstances. Notice that we cannot realize a FCI test for other education circumstance than those defined on our education system!

Table 2. A comparative view of the CI test indexes

Test Name	Sample Details	Difficulty index	50 to 50 Discrimination index	Reliability Index (Point Biserial)	Self-consistency index	Discrimination power
FCI	IMI + Physics, 2023, 43 students	0.56	0.28	0.65	0.67	0.83
BEMA	Physics, 13 students	0.53	0.58	0.66	0.77	0.88
FCI	**53, high school students, 2023**	0.43	0.68	0.25	0.67	0.83
Climate CI	**53, high school students, 2023**	0.53	0.8	0.55	0.78	0.93
PCA Pretest	Sampler: Technical Branches, 2023	0.57	0.73	**0.24**	0.76	0.91
PCA post-test	Sampler: Technical Branches, 2024	0.53	0.77	**0.65**	0.78	0.93
FCI-1-pretest	Students enrolled at Natural Sciences' Branches	0.58	0.65	**0.44**	0.66	0.87
FCI-2 - Post test	Students enrolled at Natural Sciences' Branches	0.53	0.8	**0.48**	0.87	0.89
Reference values		≥0.3	≥0.3	≥0.2	≥0.7	≥0.9

So far, we observed that the difficulty parameter for each of the CI tests presented in Table 2 falls in the desired range, and self-consistency index is also acceptable for every case. The test's scores are distributed similarly broadly, and those tests are perceived as of similar difficulty. Referring to those indexes, the tests were considered acceptable, and so are the results obtained by them. Regarding our interests, we highlight that there was no difference observed regarding difficulty index for CI tests in Physics, Calculus and Climate test, despite their different nature explained above (probably it affirms that our clearing procedure for climate test was adequate). Next, we observe that the alpha Cronbach or reliability measure for FCI and PCA test that were realized on technical branches, were significantly different. Those findings demonstrate lower homogeneity of knowledge in physics than in calculus, at least for students of those branches. Interestingly, the improvement rate of the reliability index after the course (obtained by comparing them for pre- and post-test) has resulted better for Calculus CI test compared with for FCI tests. Also, the PCA post-test reliability index is better than its initial value observed on pre-test, that indirectly indicates a good reaction of the students after the lectures on Calculus, which is not observed in FCI tests. Notice that conditions of the FCI and PCA differ in the fact that the learning in physics needs for the support of demonstration and laboratory works, therefore, FCI and PCA could be considered as the same learning test but carried out in different circumstances. Knowing by our precedent estimates the deficiencies in conceptual knowledge on high schools, we have explained this finding by arguing that the inherited conceptual knowledge problems in physics from the high school, were not repaired significantly through university courses. This interpreted is based on the assumption of additional latent factor in learning physics, which is not active when learning math and calculus. The important thing is that this difference is revealed by the proposed technique and the respective quantity can be measured straightforwardly. It should be analyzed further. Similar conclusions are obtained by commenting on the climate knowledge test which is specific because the knowledge (being tested) is not offered to student's dashboard through a structured syllabus or program. Moreover, this test was constructed in such a way as to mirror elementary academic proficiency, so we refer the finding regarding it as indicators for academic achievement features. So, we observe that the climate information which comes from non-structured syllabuses has a surprisingly higher reliability index than the FCI test. This is an indicator of the presence of *severe problems in learning and teaching physics*. On the other hand, the relatively good comportment of students on Climate test which was intended to check also student's capacities to interpret some facts etc., indicates that the group is capable for achieving progress in scientific knowledge. There are many more readable features that can be observed by comparing one by one every index, that can be used for hypothesizing about latent traits on the system of knowledge transfer. By comparing outcomes of different types of CI tests, we might judge for capacities for academic proficiency. This parameter is explored herein by analyzing indexes of BEMA and FCI on a small group of students from the physics branch. The BEMA test, is more comprehensive regarding academic achievement, see (Smaill, 2012) for more, and we will use it for further references too. We see that indexes of BEMA test were in the desired range, which is not the case for FCI test in the first year of physics, indicating that testing electromagnetic concepts in physics' students is more reliable, and the overall knowledge has improved. By employing figuratively, a kind of ergodicity idea, we might assume that students of second year of bachelor's in physics' branch are in higher level of the learning state due to the intensive and well elaborated courses in physics. On the other side, students of the first year represent lower level of knowledge in physics (the one inherited from high school). So, by comparing features of the CI tests on those groups we might learn about academic advances, regardless of the missing information about CI scores for first year physics group. This is a qualitative approach and debatable, but we hope it is helpful when we lack outcomes of the repeated test for measuring academic advances. Notice that by this approach we are not referring directly to the CI scores obtained on different test for different group, but we are comparing the global knowledge outcomes mirrored on the tests' reliabilities and adequacies through their indexes, which, indirectly would indicate the advances in learning and understanding physics according to the basic arguments presumed in this work. So, it resulted that the measurement of conceptual knowledge in the starting point (beginning of the university studies) is disputable and not reliable as evidenced by the indexes of the FCI test. After the intensive work which is embodied on physics students of the second year, the reliability and general performance of CI tests is improved as seen on the indexes of the BEMA test, Table 2. The goodness of BEMA test over the initial FCI test can be recognized as an indirect measure of academic advances and achievement, regardless of the restricted number of participants (13 students). Logically it cannot replace the repeated test for accurate measurement, but as indicator, it worked and can be considered as a measurable feature of the knowledge. Note also that the BEMA outcomes are directly related with the effect of physical laboratories in electricity and magnetism, that are part of the applicative works and academic advances during learning process. Following similar idea based on good indicators of BEMA test, and by employing a modeling approach similar with (Prenga et al, 2021), we have verified a better improvement on conceptual knowledge when adding one hour of laboratory work per week in the program, compared with adding one hour of seminars (exercises) or lectures. We will elaborate the detailed arguments and related assessments in another work, and herein we highlight that by extending the use of the indexes toward the measure of the global learning indicators, we might uncover the effect of latent variables, and in some extend we can overcome the problem of the real-world difficulties or incapacities for redoing CI test aiming on evaluating the advance and academic achievement in learning physics.

Description of the CI Metrics and Additional Tools for Measuring Specific Knowledge Features

The standard literature of conceptual knowledge tests and their analysis outlines that, among main goals of CI test, the most specific target is the evidencing of failures on conceptual knowledge (Hestenes, 1992; Rahmawati, 2018; Savinainen, 2008; Luangrath, 2011), etc. For the FCI case, it is concluded through analysis of the commonsense errors in mechanics, based on a rigor error taxonomy described thoroughly in (Hestenes, 1992), etc. so, through a combined analysis in (Prenga et al, 2022), we have discussed a direct relationship between common errors in procedural exams, and calculus and physics 'conceptual failures. This idea is elaborated further in this work. For the clarity of the reader regarding this extension, we are introducing briefly some elements of the Concept Inventory analysis and Rasch model. The Rasch technique which was improved later by B.D. Wright (Wright, 1997; Andrich, 1978, 2005), utilizes several principles and assumptions of the Item Response Theory, see (Van der Linden, 2005) etc., for more. Accordingly, one identifies the relationship between measured quantities or responses and latent traits or latent causes, which correspond also with our main interests in this work. The Rasch analysis is a well-known psychometric technique designated also for calibrating the instruments of knowledge' measurement, see (Anderson, 2010; Wim, 2005; Hambleton, 1985) for details. In the beginning of the analysis, the CI tests 'answers' are represented by a matrix $T(i,j) = (0,1)$ by assigning the binary value (1) for the correct answer and (0) for incorrect one. In a rude approach, the matrix' elements of the raw matrix T are recognized as the reference probabilities of successes, $T(i,j) = P(x_{ij} = 1)$., where x represents the answer of student (i) for the question (j). Shortly and naively, if $T(i,j) = 1$, one admits that student (i) has firmly understood the question (j), but it might have 51%:49% balance in deciding what to mark! The Rasch model suggests firstly that a better estimator of the probability that student (i) would solve question (j) should be taken $P_e(i,j) \equiv P(x_{ij} = 1|\beta_i, \delta_j) = \frac{exp\,(\beta_i-\delta_j)}{1+exp\,(\beta_i-\delta_j)}$, where $\beta_i = ln\frac{P_{correct}(i)}{1-P_{correct(i)}}$ and $\delta_j = ln\,\frac{1-P_{correct}(i)}{P_{correct}(i)}$ are the students' (i) ability and questions' (j) difficulty in logit units. The use of logit unites is based on a logistic response paradigm, see (Rasch, 1964), (Planinic, 2010), which guarantees the linearity of the scale for the measured quantities. Here $P_{correct}(i) = \frac{1}{N_I}\sum_{j=1}^{N_I} T(i,j)$; $P_{correct}(j) = \frac{1}{N_S}\sum_{i=1}^{N_S} T(i,j)$ are the raw probability that student (i) answered the test correctly and the probability that all students have answered correctly the corresponding item, and N_I and N_S indicate the number of test items and students respectively. In the calculation procedure of the Rasch model one starts from $P_e^0(i,j) \equiv T(i,j)$, and for improving their significance, matrixes $P_e^{n-1}(i,j)$ are replaced iteratively with $P_e^n(i,j)$, until a stopping criteria regarding the residuals of the quantities therein is meet. There are calculators (https://real-statistics.com/) and dedicated programs for performing Rasch analysis as (https://www.winsteps.com), but we have used herein our own routine written in MATLAB/Octave. Next, after performing the Rasch calibration procedure, elements which are characterized by the ratio of squared residuals to the variance out of a given range are classified as misfitted events regarding the statistics used. In a rude approximation those elements could be excluded from the final assessment of the Rasch model's quantities, but from our point of inserts, they contain precious information regarding students' perception of the CI test, and therefore, are considered herein as important indicators. Based on the estimate probability for answering an item $P(i,j) \sim exp\ (\beta_i - \delta_j)$, we can identify all pathological cases which consists of unexpected differences $P(ij) - T(ij)$. More specifically, if $\{\beta_i < \delta_j$ so $P(i,j) <$ ½, }and if the corresponding $T(i,j) = 1$, it means that this student has answered correctly one item that is more difficult than its measured ability to answer the test, hence, he concluded by guessing or by choosing the answer randomly. In this view we might trace all events when the expectations and observed probabilities are in different sides of the null values $P_{\beta_i=\delta_j} = \frac{1}{2}$. Statistically those evidence are flagged as pathological events (guessing, bad conduct, lack of interest to answer etc.), whereas we might use them as important indicator of the knowledge state if a standard CI test is used.

The Quantities Measured by the Rasch Analysis and Some Extension

The most important quantities which are measured through the Rasch analysis last are the calibrated CI scores, the final values for items' difficulties and students' abilities and the probability estimates, see Table 3. Notice that The CI test could be based also in multiple values responses like Likert values, or by polytomous responses which looks like the grades in classical assessment.

Table 3. Measurable quantities in the Rasch model

Quantity/observable	Formula	Description
Probability for success (correct answer for item)	$P\left(x_{ij} = 1 \middle\vert \beta_i, \delta_j\right) = \frac{exp\ (\beta_i - \delta_j)}{1 + exp\ (\beta_i - \delta_j)}$	Probability that student (i) would solve question (j) depend only on its ability and items' difficulties
Item difficulty in logit units	$\delta_j = ln\ \frac{1 - \frac{1}{N_s}\sum_{i=1}^{N_s} P^{final}(i,j)}{\frac{1}{N_s}\sum_{i=1}^{N_s} P^{final}(i,j)}$	A linear representation of the difficulty calculated based on the current results of the test held in the current group of students. It varies depending on the group being interviewed
Students' ability to solve the test in logit units	$\beta_i = ln\ \frac{\frac{1}{N_i}\sum_{j=1}^{N_i} P^{final}(i,j)}{1 - \frac{1}{N_i}\sum_{j=1}^{N_i} P^{final}(i,j)}$	A linear representation of the difficulty calculated based on the actual results of the student. It expected to pure characteristics of students' knowledge
Guessing and students' confusions	$(\delta_j - \beta_i > 0) \wedge (T(i,j) = 1)$ $(\delta_j - \beta_i < 0) \wedge (T(i,j) = 0)$	Student has answered a difficult item, but his ability is smaller, or students has failed though he could solve it (ability higher than difficulty)
Gain factor	$\frac{\%FCI_{scores} - \%FCI_{scores}}{100 - \%FCI_{scores}}$.	Improvement on conceptual knowledge

Rate of changing between answering levels	$(x_{ij} = h\|\beta_i, \delta_j, T) = \frac{exp\ [k\left(\beta_i^{(k)} - \delta_j^k\right) - \sum_{j=0}^{h} \tau_j]}{\sum_{k=0}^{m} exp\ [k\left(\beta_i^{(k)} - \delta_j^k\right) - \sum_k \tau_k]}$	This calibrated rate could estimate indirectly the academic proficiency under a quasi-ergodic approach * *= our own suggestion,

In the last row of the table, parameter τ is considered the 'threshold' difficulty parameter, and represents an intriguing quantity for measuring differences between polytomous levels of knowledge (https://www.rasch.org; Planinic, 2010). Aside classical quantities listed in Table 3, we considered also some derived statistical objects related to them, as the distributions of the students' abilities and classes of the perceived difficulties (Kushta, 2023; Prenga, 2022). Adding to that, we qualified as indicators of the knowledge the mismatching of the magnitude of student's ability with the item's difficulty $\chi = \{\delta_{max} - \delta_{min}\} - \{\beta_{max} - \beta_{min}\}$. Ideally, the magnitude of the student's ability and items' default should match. Now, when using a certified CI test for the measurement, we recognize it as an indicator of the student's shortcoming on the knowledge, hence we select is as another measurable quantity. Next, the guessing ratio measured by $g_{rating} = \frac{TotalGuessing}{TotalAnswers}$ is seen as the response to several factors, but again, we interlinked it with the failure of understanding the question asked by the item, if a standard CI test like FCI or BEMA is used. When conducting a multiple-choice test, we might be interested also in students' behavioral issues during the decision process. So, students might hesitate to give an answer, might be confused during selection between alternatives and so on. Obviously, this behavior is related to the knowledge level because he(r) had not to worry or hesitate if he(r) now the correct answer. Those secondary variables are very helpful for understanding the whole pattern of the learning process. The student might have understood the context of the question, but he does not know the correct answer. In this case he might decide *to skip answering*, to pick up by random an alternative, to choose more than one alternative etc. Additionally, students mightn't ever know what is asked by the question! Again, he might check by random the alternatives which modifies the final ratio of answering the test $r = \frac{N_{Correct}}{N_{questions}} \rightarrow r_{real} = \frac{1}{N_{question}*n_{alternatives}}$. Therefore, in the presence of the guessing events, the ratio of answering or the raw ability would result smaller than the one calculated directly. It highlights the importance of regarding *the guessing ratio* not only as misfit statistics' estimates, but also in the role of a *global disturbing factor for the testing* itself, which could and should be measured. Initially, we considered those behaviors as markers of the knowledge shortcoming: mostly representing the fact that the student has no idea what is asked on this item, and not a normal situation where the student simply does not know the correct answer for it. Related to this parameter, we might judge also the efficacy of the tests themselves, in the sense of their validity as a measuring instrument, despite the fact that we are using a certified CI test. Additionally, frequent occurrences of the guessing events might indicate that students were not willing to respond to the test, so they just participate in it and finish the job irresponsibly, because nothing would impact him. This another argument for extending the metric of the Rasch analysis deeper on the students' overall comportment and their behavior, qualifying the guessing ratio as an important parameter for the testing efficacy. To evidence the nature of frequent pathological events of the guessing type, we have realized a pilot FCI test in similar groups by modifying the secondary conditions of the testing protocol. So, for the first test, the records of students are taken anonymously e.g., assigning $Student1, Student2$ ..., instead of their names. The other test has been conducted like a normal exam", by assigning '$Name1, Surname1$, ... for students list. The idea is to check their relationship between *answering responsibility* and *observed guessing behavior*. Herewith, we observed a reduction of about 80% in guessing events, which confirms our doubt that common tests outcomes do not report real knowledge of students. So, our proposed indicator varies by changing some condition, indicating for another measurable quantity on CI analysis, which is *the guessing ratio*. Seeking for other measurable quantities on CI testing activity, in a pilot FCI testing carried out on a small group of students which participated in it voluntarily, we measured the time needed to finish the standard CI test. We instructed the participants to continue the exam after the deadline (fixed at 40 min according to recommendations on www.PhysPort.net), asking them to fill the reviewed answers in a separate sheet. By this experiment we observed that around 70% of students have reviewed at least one item. This was an indicator that during normal tests, they had problems of face validity, or the time allocated for covering specifical subject is likely to be insufficient. This example indicates that there are several measurable features that can be analyzed by the Rashc model, in addition to the standard ones included in Table 4. This element was not detailed further in this work, letting it for future reference.

Table 4. An outlook of misfitting ratio for several CI tests

Test Type	Branch, number	Outfits (1.5 level)	Infits (1.5 level)	Possible Guessing	Comments
FCI-1-1	Informatics, 55 students	14%	15%	18%	Questionable results
FCI-1-2	Informatics, 85 students	12%	6%	16%	Questionable results
FCI-1-3	IMI, 35 students	6%	6%	8%	Acceptable level *
BEMA*	Physics (13) IMI (17)	18%	12%	23%	Unsatisfactory statistics. But indicatory
FCI*	High school, 53	9%	11%	12%	Moderate to acceptable
Climate C*I	High school, 53	13%	11%	12%	Moderate to acceptable

Note: Marked by asterisk the test where the sample is equal to the whole classes (e.g., no sampling)

Here, the guessing ratio has dropped to around 5% for all students which do not finish the test on the scheduled time (40 min). By crossing those results, we were able to identify that the reason for guessing on those investigative attempts were not caused by a sort of *unfair conduct* in the exam, nor from *lacking the desire to resolve it*. It resulted mostly in students' decision to answer randomly because *they do not know what is asked by the question, or they do not know the answer*. In this case, this parameter consists partially as in the original definition by theory, and can help for understanding the knowledge in general, additional to the total CI scores. From one side, it can be used as preliminary step for qualifying the validity of the testing itself, and next, if it the validity is confirmed, we can use its value for evaluation knowledge features and specifically the efficiency of the knowledge transfer through education process. This last property is based on our argument, that if students provide an incorrect answer to a question, it means that he failed on the undertaking the concept being asked on that time, but if he guess too frequently, he does *not understand what is asked at all*, signaling a more general failure which is related to the *teaching process and education (non) efficiency*. Both those lasts are latent traits and behave like hidden variables that cannot be measured directly. Similar argument can be elaborated by comparing guessing in a standard and structured knowledge testing like FCI and unstructured knowledge embodied in the Climate test. So, from the fact that the same group of students (shown by two last rows on Table 4) exhibits the same level of guessing in FCI and Climate test, suggests that the process of knowledge transfer in physics might suffer important drawbacks, because its results are similar to those of a nonstructured discipline! Finally, we considered an additional analysis of the CI score gain $G(\%) = \frac{\%Scores_{AfterCourse} - \%Scores_{BeforeCourse}}{100 - \%Scores_{BeforeCourse}}$, (Coletta,et al, 2005), which originally is used for measuring the efficiency of the teaching process. Here we will present the observed relationship between this quantity and some other cause variables, by the same idea elaborated in (Prenga et al, 2022). This examination consists of a double test, before and after the course of physics. The example discussed herein aims an illustration of the use of CI test outcomes to facilitate a modeling attempt for learning progress, and also for measuring the effect of cause factors that influence it. This observation is realized by different surveys and stage, which are assembled and reorganized later for this analysis. Some of those observations are realized in a limited number of students, because it was not easy to gather the same sample of students for redoing a physics test! consistent with our worries presented in the beginning of this paper. In the case-study aimed on evidencing the relationship between CI score gain and the laboratory component in the program (Prenga et al, 2022) the conclusion was deduced statistically by comparing the results of the paired test on two different branches, Engineering in Informatics and Mathematics (IMI) and Chemistry at the Physics in the Faculty of Natural Science, University of Tirana. Other conclusions refereed in Table 5, have been found after filtering original tests based on the values of the variable of interest. Those value were declared by students in the beginning of the testing process, by filling (voluntary) the additional form attached to the main CI test. For this sub-case, we are aware of statistical inadequacies due to the small number of the records, un-matched students' samples etc., but we believe that those deficiencies do not hamper significantly the validity of the conclusions obtained. The results of this survey are shown in Table 5. Notice also that for the purpose of this analytic review, the gain coefficient is recognized here as an indicator of the academic progress, so we have used those terms interchangeably herein. So far, it is evidenced a correlation between the CI score gain g and several variables related to the process of the education, and indirectly, we can report the corresponding academic advance on the terms of Concept Inventory improvement. To be clear, we do not equalize the Conceptual Knowledge advance with academic progress and proficiency, nor we claimed a direct measurement using CI tests, but unless the purposed test would take place, those finding and the methodology for achieving them are considered valid. This feature is solidified by the fact that students who reported that they have finished regularly the homework (that by nature involve several competencies that indicate also academic proficiency and students' ability to solve more general problems), showed a better improvement of the g parameter. Table 5 can be viewed also as preliminary factorial analysis for modeling the progress of knowledge transfer and efficacity of learning objectives stated on the current syllabus.

Table 5. The CI test score gain

CI Gain	Frequentation of lectures	Initial CI score	HomeWorks finished	Other possible categories-Gender
10%	Less than 60%	Up to 30%	Occasional	Mixed
20%	Les than 60%	30-35%	Occasional	Mixed
30%	65%- 75%	40-43%	Mostly	Mixed
40%	80%-100%	43%-50%	Mostly	Girls
60%	100%	> 50%	Mostly	Mixed

It is interesting to evidence that initial knowledge (inherited from previous stage of studies) acts also as a cause variable for the overall academic progress, and therefore, it is possible to measure directly those interrelationships via modeling. For those application, a purposed testing and a carefully selection of the variables is necessary, however. In summary, a useful estimation and indicatory assessment of the academic advance and proficiency can be realized by using the gain on CI scores and some auxiliary indicators which are easier to handle compared to the dedicated tests. It completes the set of the additional parameters of the Rasch analyses that can be used for indirect measurement of academic proficiency, overall knowledge and teaching functionality in physics and sciences in general.

An Assessment of the Academic Proficiency on High School Students and First Years of Bachelor's by Using Polytomous Rasch Model

Estimating academic proficiency for high school students and first year of bachelor's is a challenging goal of testing and surveys, because the concepts of the syllabuses in those stage of the education are mostly basic, and few students achieve the ability for employing knowledge and skills on solving real problems. Another difficulty comes from the '*pragmatic learning'* practice imposed by the current methods of teaching and rules in force regarding the opportunities for university studies, which are based mostly on procedural learning, (Kushta et al, 2023). They drive the interests of students toward procedural learning instead of critical thinking and deepening into the conceptual understanding in physics and science. The evaluation of the knowledge advancements requires two testing procedures which must be scheduled and edited carefully. Sometimes, those purposed tests turn out be difficult for realizing, especially regarding statistical requirements, and their costs. Moreover, students' (non)willingness to conduct repeated tests is not favorable at all. Alternatively, if we know the distance between knowledge levels being measured in a general CI test, we can parallelize the advance in learning with the gap between knowledge levels. This is basically our proposal for measuring indirectly the progress of understanding during the learning process. More explicitly, we suggest to initially measure the knowledge' levels by using polytomous variable, and next, to extract specific features related to the academic achievement based on perceived tests' difficulty levels using the Rasch analysis. The idea is based on the assumption that every individual would advance at the same rate, if the cause factors that affect his knowledge would be the same. Under this presumption, the perceived difficulty would be the same. Also, following this idea, if students' abilities based on the CI scores would be partitioned into n-natural categories, the normalized probability for going from the next level $i \rightarrow i+1$ will report the rate of advance in learning. In (Prenga et al, 2022), we have proposed the estimation of the natural levels of the ability by using optimal histograms. However, in this reference we do not concluded on the evaluation of the transitive probabilities between levels, because the distributions of the abilities that we have obtained were highly non-stationary. On the other side, the calculation of the threshold parameter for the difficulties by the Rasch procedure, is straightforward and statistically based. So, as mentioned above, the estimated probability for answering the question at level h for a h-scales polytomous variable is given by the relation

$$P_e(i,j) \equiv P\left(x_{ij} = h | \beta_i, \delta_j, T\right) = \frac{exp\ [k\left(\beta_i^{(k)} - \delta_j^k\right) - \sum_{j=0}^{h} \tau_j]}{\sum_{k=0}^{m} exp\ [k\left(\beta_i^{(k)} - \delta_j^k\right) - \sum_k \tau_k]} \tag{2}$$

where τ is the global difficulty parameter, and it measure the gap between successive levels of the perceived difficulty. Resuming the polytomous Rash analysis, the threshold difficulty element can be seen also as perceived average effort of students' for improving their knowledge. Now, for elaborating our technique, which is based on using outcomes of the standard CI tests instead of expensive and time-consuming dedicated ones, we need for the polytomous variable. Let consider a common assessment of the knowledge by a dichotomous CI test. For matching our proposal with classical CI outcomes, we considered the set of several

CI test conducted in various sample or on virtual sampler being realized by random selection of (n) records like in bootstrapping techniques used elsewhere in statistics applications, (Muller, 2020; Brownlee, 2019). For realizing 'different FCI questionnaire' based on a unique FCI test, we choose randomly a set of 5 FCI items 'records, intended to mimic an item with a 5-scale-valuation inspired by the Likert psychometric scale. Next, for every compound of FCI items, we assigned the corresponding students score by simply summing up the binary values of the constitutive questions. For avoiding the effect of the error propagation through those manipulations, we have performed a preliminary Rasch analysis, after the which, all highly outfitted items have been filtered. Here we used the ad hoc criterion of a 1.8 level of the squared residual to the variance ratio, which actually is higher than the standard recommended one for misfitting analysis elsewhere at 1.3, (Planinic, 2010), (www.real-statistics.com), because the level of outfits and infits in our examples were relatively high, but we intend to illustrate the idea based on a qualitative evaluation, only. Finally, we populated the raw table with those new compound items to produce our virtual polytomous CI test. We used records gathered on various FCI tests reported above, and we produced 6 virtual tests, based on an extra control variable, for comparison purpose. For this purpose, we have selected '*the significance of the use of demonstrations and laboratory works for supporting physic lectures'*, which has been reported by students in separate sheet when conducting original test. We have selected this factor based on common pedagogical agreement and arguments elaborated in (York et al, 2019; Nyutu, 2021). The threshold parameter for this polytomous tests was calculated by employing the Rasch techniques and are shown in Table 6. Therein, the category (I) contains 3 groups which were made by redoing the procedure of producing polytomous tests, based on the original FCI records of students which reported 'i*nsufficient laboratory works and demonstration in their physics classes*'. The category (II) contains groups 4-6 that was built with the FCI records of students which report those supportive activities as '*sufficient*' or '*mostly*'. Based on the description in this section, the differences of the would report the effort needed to go from one level of the perceived difficulty to the other for each group, which report indirectly the thought-learning progress. The row of the table contains the threshold estimate for each level of polytomous variable. Their differences are shown separately on this table too.

Table 6. The threshold parameter and successive gaps for various group of students

Polytomous value (Scores)	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
1	0	0	0	0	0	0
2	-2.3398	-2.2292	-2.0454	-1.5896	-1.8743	-1.869
3	-0.857	-0.8732	-0.8302	-0.5781	-0.8671	-0.9404
4	0.5047	0.3073	0.8525	0.3716	0.4851	0.5807
5	2.6921	2.7951	2.0232	1.7961	2.2563	2.2286
Differences -gaps between difficulty levels						
$\tau_2 - \tau_1$	-2.3398	-2.2292	-2.0454	-1.5896	-1.8743	-1.869
$\tau_3 - \tau_2$	1.4829	1.356	1.2152	1.0115	1.0072	0.9286
$\tau_4 - \tau_3$	1.3617	1.1805	1.6827	0.9497	1.3522	1.5211
$\tau_5 - \tau_4$	2.1873	2.4878	1.1707	1.4245	1.7712	1.6479

From Table 6 we observe a significant gap $\tau_2 - \tau_1$ between levels (2) and (1), corresponding to the effort for going from the lowest knowledge level to the second one, for students of category I, which report *un-significant demonstration and laboratory work supporting one their physics lectures,* compared to the other groups. The same feature is observed for differences between very good knowledge and good knowledge. This finding supports the importance of physical demonstration and laboratories of lecturing physics, and also, indicates that the effect is more significant on two extremes of the knowledge represented by the perceived difficulty of the tests, etc. Also, from Table 2, we observe that according to our terminology, more effort is needed to brink students from lowest level to the next better ones, than the effort needed for increasing knowledge level starting from other points. The same result is observed for the effort needed to send them from level 4 to the highest one, for both cases. More interesting finding can be elaborated by comparing each element of the Table 6. At this point we might consider those findings as an affirmation of the possibility for measuring academic achievement indirectly by using a common and certified CI test as the FCI or other similar ones.

Using Miscellaneous Statistical Arguments for Analyzing Schooling Efficacy

The study of hidden knowledge features and the factorial analysis related to it is stirring because of the difficulties for conducting appropriate social experiment, or for realizing alternative measurement. Aiming on identifying some hidden elements on learning physics, in (Prenga et al, 2022) we have proposed a Likert-scale for identification of the most influent commonsense errors in FCI test. We assigned the error states which typify failures on the standard FCI test as follows {*no-errors, calculation failures, conceptual failures, equalized/both errors presence}* the corresponding eigenvalues values [0,1,2,3] which are based Likert scale values. Also, the calculation that we are not reproducing and commenting here but can be found in (Prenga et al, 2021), are based on the Likert analysis. Based on this idea, we have reanalyzed the pairing of the typical common senses' errors in Mechanics with basic drawbacks that characterize learning process in physics, namely problems with calculus and algebra, shortcomings in physics' conceptual knowledge or their mix. The results are shown in Table 7.

Table 7. Likert values and error states

		Typical Common sense				
Test	Sample size	Kinematics Confusions	Impetus	Active Force	Action/Reaction pairs	Other Influences in Motion
2020-2021	203	0.86	1.47	2.03	1.18	2.44
2022-2023	135	0.65	1.43	2.23	2.22	2.64
2023-2024	53	0.76	1.23	1.98	2.05	2.25
Error states	No errors	Calculation	Conceptual	Conceptual Calculations		
Likert intervals	0-0.75	0.76*-1.50	1.51-2.25	2.26-3.00		

So, by a repeated analysis for other test conducted in the following, after a straightforward assessment which can be found also in (Prenga et al, 2022), we have evidenced a pairing pattern of each error type occurring on a pilot procedural test intended to mimic common exams, with every element of the basic misconceptions on mechanics {*kinematics confusions, impetus, active force, action/reaction pairs, concatenation of influences, other influences in motion},* according to the taxonomy introduced and analyzed in (Hestenes, 1992). *T*he calculating errors were found dominant for items were *kinematics confusions* were mostly probable, and also they were likely the causes of the failure on solving problems in which the *action-reaction common-sense* is present; for problems where the most probable common-sense confusion is *active force, the* conceptual errors are more frequent; and for problems in mechanics where o*ther influence in motion* common-sense are most likely we obtained that '*both errors are present*'. This pattern is obtained for some group of students enrolled in first year of bachelor on engineering-branches at the Faculty of Natural Sciences, University of Tirana, but we believe that this conclusion would remain valid for a broader group of students. In this case, the method can be considered as a diagnostic tool for the whole education in physics in the high school. This finding is interesting because it evidences a permanent problematic issue regarding teaching physics also at bachelor level, which must be addressed. But again, the most important conclusion that we want to highlight herein is the capability of the combination of statistical tools with the outcomes of the CI analyses for investigating and even for measuring the effect of the latent variables or non-evident cause factors that influence academic progress and knowledge transfer in physics.

Another interesting application of the proposed methodology entails the evaluation of the classes of students' abilities by using optimal histogram idea. In this case, after performing the Rasch evaluation for logit abilities, we used those output for assessing natural classes of knowledge expressed by the histograms of the abilities, (Prenga et al, 2022). Also, the largeness of the classes which is identified by the bin size, is an important indicator of the distribution of the abilities to solve the CI test. So, for the group of the high school students (being tested immediately after COVID, in (Prenga et al, 2022) we obtained that optimal number of histograms was 6, so we assumed that students' abilities can be categorized in 6 classes. Based on this finding we concluded that, methodically, if we want to avoid penalizing low-ability students in a procedural test, we might include problems with 6 levels of difficulties. This is an important conclusion if we want to measure academic proficiency by using a general-purpose test based on problem-solving abilities. Later on, for a group of 135 students of FNS tested during 2022-2023, we obtained 7 levels, and in these cases, results are of a more general nature because its indexes were all in the validity range. For 156 students which were interviewed on academic year 2023-2024 we obtained 5 levels of the knowledge. Despite of the small number of records for concluding in an adequate histogram, those findings suggest that in general, the echelons of the difficulties of the items in common tests must be projected to be scaled on 5-7 levels. Regarding the analysis based on the distribution of the abilities, we have considered also the changes of the distributions' asymmetry as an indicator of the dissimilarities between knowledge patterns on different groups, which can be scrutinized further for a more comprehensive analysis.

Table 8. Features of the student's abilities distributions for FCI tests

Academic Year	Nr. Participants	Test type	Indexes validity	Specifics of the filed	Ability Distribution. Type	Item difficulty Distribution	Optimal classes
2020-2021	203	FCI	No	Regular syllabus	Skewed, apparently TlocationScale	Slight skewed	6
2022-2023	135	FCI	Yes	Regular syllabus	Skewed, TlocationScale	Slight skewed	7
2023-2024	53	FCI	Yes	Regular syllabus	Skewed, not identified	Moderately skewed	6
2023-2024	53	Climate	Yes	Distributed source	Apparently Symmetrical	Slight skewed	5

For evidencing similarities between the distributions of abilities for different tests, we can use standard statistical tests (Kolmogorov Smirnov, Shapira, Anderson). Also, we might read information by comparing perceived difficulties of FCI items by different groups of students. In this case, one can consider the rank-sum test (Wilkinson) to deal with small number of points. Detailed conclusions related to this section are not in scope of this paper, and are in the process of publishing elsewhere, but we might illustrate some finings in Table 8. Specifically, we obtained and commented the apparent dissimilarities between distributions of abilities for different groups and difficulties perceived by them, and similarly for different tests types. By the ways we cannot realize the measurement in different learning and teaching condition, but again, for mimicking cause-responses relationships we propose comparing of the outcomes of the CI test for two subjects that differ remarkably by channels of knowledge transferring that they use. In short and commenting the findings in Table 8, the composite mechanism of the knowledge transferring, and students' prior affinity to the taught matter, has been found of low efficiency for the physic discipline in high school, as shown by the unstable distribution of the abilities (tLocationScale) and its skewness. Also, this feature has reflected the change in teaching process which is evident for the test that was carried out during online learning (held in 2020-2021), and the one conducted on regular-learning period (2022-2023). So, for the second test, the distribution is smoother and identifiable, indicating that the underlying process is well defined for this period. Just for mentioning, we can measure the level of the un-stationarity on those distribution by using q-Gaussians, which are used for unequilibrated processes and states, based on generalized Tsallis' statistics. Next the symmetrical distribution of students' abilities to answer our ad hoc climate test shown in Table 8, mirrors homogeneous student's affinity to learn about climate, because the other component, the teaching efficiency of those matters cannot be evidenced for a system of the knowledge transfer via a non-organized syllabus. Again, we highlight the possibility for using the parameters of the distributions of the CI test' outcomes as measurable quantities of similarities or dissimilarities, which can be analyzed and discussed further by employing pedagogical, sociometric and other scientific arguments of education disciplines.

Conclusions

Several problems encountered on studying students' understanding in science regarding realization of the adequate and reliable tests, analyzing latent factors and estimating indirect effects, can be overcome by combining standard tools of knowledge measurement and extension of standard metrics, with statistical techniques and tools. By these techniques we have analyzed some semi-latent variables that affect knowledge and its advancement on high schools, which are presented as illustration in this work. So, the ratio of misfits' items and guessing events resulting after Rasch analysis for certified Concept Inventory tests, and un-pleasant indexes' value obtained for them, are considered as measure of overall knowledge drawbacks, learning in-competencies and failure of knowledge transmission during teaching process. Through a comparison of the indexes for CI tests carried out in various conditions and pedagogical circumstances, or the CI scores' gain, we evidenced indirectly the efficacy of the teaching of physics in high school. Similarly, we considered the comparative approach between different tests aimed to mimic different circumstance, to overcome the unpracticality of the repeated test or for conducting needed experiments for measuring academic progress and proficiency. So, we used the outcomes of the Rasch analysis for BEMA and FCI conducted on specific group of students assigned purposely, for estimated the academic achievement and progress through the studies. In another application we combined the outcomes of the CI tests in structured disciplines with tests of knowledges gathered by a dispersed learning system as the Climate Knowledge, for analyzing the efficacy of teaching the first, especially in physics. Also, by using Likert-like measurement, we investigated the paring of the commonsense errors in physics' exams with calculus failure and inherited conceptual drawbacks. By using similar idea, we considered the histograms of the calculated students' abilities and the features of corresponding distributions' features for exploring the structure of the ability and its relationship with the efficacy of the knowledge transfer in teaching. The techniques used herein offer an alternative and helpful solution to overcome difficulties encountered during testing stage, and overall students' knowledge analysis.

References

Anderson, C. J., Verkuilen, J., & Peyton, B. L. (2010). Modelling Polytomous Item Responses Using Simultaneously Estimated Multinomial Logistic Regression Models. *Journal of Educational and Behavioral Statistics*, *35*(4), 422–452.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–73.

Aubrecht, G. J., & Aubrecht, J. D. (1983). Constructing objective tests. *American Journal of Physics*, *51*(7), 613-620. https://doi.org/10.1119/1.13186.

Brownlee. J. (2019). A Gentle Introduction to the Bootstrap Method. by on August 8, 2019 in Statistics. https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/

Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, *73*(12), 1172-1182. https://doi.org/10.1119/1.2117109

Crooks, N., & Alibali, M. (2014). Defining and measuring conceptual knowledge in mathematics. *Developmental Review*, *34*, 344–377.

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics—Physics Education Research*, *2*(1), 010105.

Gliem, J. A., & Gliem, R. R. (2003, October). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. In *Midwest research-to-practice conference in adult, continuing, and community education* (Vol. 1, pp. 82-87).

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

Handhika, J., Huriawati, F., & Fitriani, N. (2017). Force concept inventory (FCI) representation of high school students (SMA & MA). *Journal of Physics: Theories and Applications*, *1*(1), 29-34.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The physics teacher*, *30*(3), 141-158.

Kim, J. S., & Sunderman, G. L. (2005). Measuring Academic Proficiency under the No Child Left behind Act: Implications for Educational Equity. *Educational Researcher*, *34*(8), 3–13. http://www.jstor.org/stable/3699961

Kushta. E, Prenga, D., Miço, S., & Dhoqina, P. (2022). Assessment of the Effects of Compulsory Online Learning during Pandemic Time on Conceptual Knowledge Physics. *Mathematical Statistician and Engineering Applications, 71*(4), 6382–6391.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, *140*, 1–27.

Müller, M. (2020). Item fit statistics for Rasch analysis: can we trust them?. *Journal of Statistical Distributions and Applications*, *7*, 1-12. https://doi.org/10.1186/s40488-020-00108-7

Nyutu, E. N., Cobern, W. W., & Pleasants, B. A-S. (2021). Correlational study of student perceptions of their undergraduate laboratory environment with respect to gender and major. *International Journal of Education in Mathematics, Science, and Technology (IJEMST)*, *9*(1), 83-102. https://doi.org/10.46328/ijemst.1182

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2), 020111.

Planinic, M., Ivanjek, L., & Susac, A. (2010). Rasch model based analysis of the Force Concept Inventory. *Physical Review Special Topics—Physics Education Research*, *6*(1), 010103.

Prenga, D., Kushta, E., & Musli, F. (2023). Enhancing Concept Inventory Analysis by Using Indexes, Optimal Histogram Idea, and the Likert Analysis. *Journal of Human, Earth, and Future*, *4*(1), 103-120.

Prenga, D., Kushta, M., Peqini, K., Osmani, R., & Hyseni, M. (2022). Analyzing influential factors on physics knowledge weakness in high school students due to the pandemic-imposed online learning and a discussion for enhancing strategies. IC-MSQUARE, 5-7 September, Budapest.

Prenga, D., Peqini, K., & Osmani, R. (2021, November). The analysis of the dynamics of the electorate system by using q-distribution-a case study. In *Journal of Physics: Conference Series* (Vol. 2090, No. 1, p. 012073). IOP Publishing.

Rahmawati, R., Rustaman, N., Hamidah, I., & Rusdiana, D. (2018). The Development and Validation of Conceptual Knowledge Test to Evaluate Conceptual Knowledge of Physics Prospective Teachers on Electricity and Magnetism Topic. *Journal Pendidikan IPA Indonesia*, *7*(4), 283-490.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* IV (pp. 321–334). Berkeley, California: University of California Press.

Rhodes, T. L. (2008). VALUE: Valid assessment of learning in undergraduate education. *New Directions for Institutional Research*, *2008*(S1), 59-70. https://doi.org/10.1002/ir.262

Sands, D., Parker, M., Hedgeland, H., Jordan, S., & Galloway, R. (2018). Using concept inventories to measure understanding. *Higher Education Pedagogies*, *3*(1), 173-182. https://doi.org/10.1080/23752696.2018.1433546

Savinainen, A. & Viiri, J. (2008). The Force Concept Inventory as a Measure of Students_ Conceptual Coherence. *International Journal of Science and Mathematics Education, 6*, 719Y740.

Smaill, C. & Rowe, G. (2012). Electromagnetics misconceptions: how common are these amongst first- and second-year electrical engineering students? American Society for Engineering Education. AC 2012-4051.

Van der Linden, W. J. (2005). Item Response Theory. *Encyclopedia of Social Measurement* (pp. 379-387). Elsevier.

Vilia, P. N., Candeias, A. A., Neto, A. S., Franco, M. D. G. S., & Melo, M. (2017). Academic achievement in physics-chemistry: the predictive effect of attitudes and reasoning abilities. *Frontiers in psychology*, *8*, 1064.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97-116.

York, T. T., Gibson, C., & Rankin, S. (2019). Defining and measuring academic success. *Practical Assessment, Research, and Evaluation*, *20*(1), 5. https://doi.org/10.7275/hz5x-tx03

https://real-statistics.com/reliability/item-response-theory/building-rasch-model/

https://www.mathworks.com/products/matlab-online.html

https://www.rasch.org/rmt/rmt213d.htm

https://www.winsteps.com/winman/misfitdiagnosis.htm