Using tests' indexes to improve the assessment of the conceptual knowledge: a case study.

ELMIRA KUSHTA, DODE PRENGA,

[1)]Department of Mathematics, Faculty of Technical Sciences, University of Vlora, ALBANIA
[2)]Department of Physics, Faculty of Natural Sciences, University of Tirana, ALBANIA

Abstract: The use of statistical indexes for a standard concept inventory test is a standard tools for evaluation of the test reliability and discrimination power, whereas the Rasch analysis is a well-known calibration technique for these tests. Herein we harmonize both tools in a step-by-step procedure for a better assessment of the student's conceptual knowledge in physics. Initially we conducted a standard Force Concept Inventory test in a sample of students and evaluated the reliability and discrimination indexes of the items. Next we requested that all items' indexes measurement in a given test must fulfill reliability and discrimination prerequisites, before using them for further calculation. After adding data from the supplementary FCI test we realized the fulfillment of this ad-hoc criterion, and next those data are used for a better assessment of the CI in physics knowledge and evaluation of major factors that affect students' knowledge.

Keywords: physics concepts, education study, the Rasch technique, factorial analysis
Received: July 9, 2022. Revised: February 3, 2023. Accepted: March 11, 2023. Published: April 11, 2023.

1. Introduction

From the education perspective, the conceptual knowledge plays a vital role in science understanding and for guarantying the students' capabilities of using scientific knowledge in engineering and other application. This highlights the importance of conceptual knowledge measurement and its improvement. In practice, student's conceptual knowledge is assessed by conducting concept inventory test following the Force Concept inventory introduced in [1] and discussed in large physics education and methodical literature, see [2-9]. In general, the scientific knowledge in each education stage is conditioned from conceptual knowledge in previous stage, which reinforce the necessity of concept knowledge analysis. When discussing physics teaching and physics science, the first impression that is shared commonly among students and professors is that physics knowledge is indispensable for sciences of nature and engineering, and that physics itself is a very attractive science. However, in modern time this advantageous position of physics might not be proportional with the number of students following it, at least for some countries. So, we observe that in our country for example, the number of students enrolled in physics branch has decreased constantly toward survival levels as of today. It is likely that in many countries and especially for those in development, the pragmatic interest to join physics branch has reduced as result of many socioeconomic factors. The idea behind this reality is how to react for enhancing the teaching of physics as to make it attractive enough for students. In this framework and in a coherent relationship with determination of the factors that affect scientific understanding from learning process, the improvement of knowledge measuring instruments become a pivot pillar for activities aiming the development of efficient teaching methodology. We have observed recently that due to COVID closure and full-scale online learning, the level of conceptual knowledge has been impaired seriously [10-13]. So, for a relatively representative sample, we obtained a CI score for Force Concept Inventory at the compromised understanding level in [12]. Therein, typical factors implicating knowledge's failure have been identified by contemplating a formal logistic model based on the FCI test results. By other observation for our system for the post covid period, it resulted that the abnormal conceptual knowledge failure has (wrongly) figures out physics or mathematics as very difficult, significantly affecting the students' unreadiness to follow the studies in physics, mathematics etc., aside other important factors affecting students study branch choices.

In a purposed survey held in 2021-2022 on a total 318 students for this perception we have evidenced that students mostly think that physics is very difficult. Also, a very significant part of them thinks that physics is not necessary for university study in the branches they hope to attend. In this survey we asked participants (high school students in four districts) to rank the difficulty of physics by Likert scales 1-5 as fallow: (5) for the most difficult and unpleasant; (4), very difficult and unpleasant; (3) difficult but attractive; (2) normal, (1), easy and very pleasant. Students were asked to strikethrough the qualifier e.g., encircle (5) but strikethrough 'unpleasant'. It resulted that very few of them have eliminated the adjective, keeping their ranking as in original test. The result was interesting, the mean attitude was obtained at $l_{avg} = 4.21 \pm 0.43$ belonging to the class "most difficult" even in the edge (4.2-5 correspond to the most difficult). Considering some issues of the statistical significance, we concluded that students thinks that physics is seen as very difficult to the most difficult course. The most surprising finding was the attitude of the student's sample for usefulness of physics knowledge for their future university studies. Here we asked students to provide their preferences and only students that were interested in natural science branches, engineering, and medicine. Here the highest value (5) was appointed for indispensably useful, (4) for very useful, (3) for medium usefulness (2) for useful and (1) for unusefulness. The Likert average for 83 students belonging to the target category explained herein was $l_{avg} = 2.43 \pm 0.79$! Those students do think that the role of physics knowledge for their future studies is estimated as only 'useful' which correspond to the interval [1.8-2.6]'. Under those limitation physics appeared as very difficult subject and unpleasant (!) which merits a strong 'why?'. Aware of some statistical significance issues, both results of the descriptive survey mentioned herein advocate the claim that knowledge in physics might suffer significant problems whatsoever. However, this claim must be verified through measurement. Note that student's knowledge can be recognized from result of the official or standard exams, which are mostly procedural. For a bundle of other reasons, conceptual knowledge would be better recognized and assessed by direct tests of the type of Concept Inventory. Remind that conceptual knowledge on physics is focused on the understanding of concepts and relationships between physics variables and models and procedural knowledge considers the ability to solve step by step problems, see [1-3] etc. Based on the Force Concept Inventory test introduced in [1] many other CI tests have been introduced and refined by methodologist and education scientists for physics, mathematics and for science in general. The CI-scores obtained as the first outcome, is analyzed thoroughly by the Rasch. However, by nature, the CI tests aims mostly on analyzing errors and commonsense beliefs which affect scientific conceptual knowledge, rather in measuring direct knowledge, see [1-5], [17-18]. In our recent address on assessment of the conceptual knowledge on high school students affected from the COVID-closure, we measured the CI scores and other related quantities for FCI and Simplified -FCI tests [13]. Nevertheless, therein we didn't consider thoroughly the test as seen from the student's perspective, despite that we evidence that contextual misunderstanding were the most frequent causes for conceptual shortage. In [14] it is argued that faculty members don't think like a typical introductory physics student and for reliable results, the questions in a good concept inventory test should be based on research into the way students think and are often designed to elicit common student ideas that are surprising to faculty. Also, in this reference some interesting statistical indexes have been used to diagnose the BEMA CI test. Following those idea and the preliminary finding mentioned herein, we have considered an improvement of the CI testing stage by exploring reliability and validity issue before evaluation the test outcomes. Next, by contemplating standard test CI tests which are fully validated from experts, we propose to use it to analyze the student conceptual knowledge perspective, that is how the item has been understood as read from students. By limiting our goal in a concrete measurement of the knowledge issues and the efficiency on assessing it, we will present the result of a small group of students which might not be statistically roughly representative for all students on the country, but at the same time it can reveal problematic aspects that worth to discus in this framework. Within this limitation, the assessment and analysis herein are developed from a quantitative perspective.

2. The statistical indicators for standard and common knowledge tests

The validation of the test is important step before using it for assessment purposes. Like the Rasch calibration procedure. A detailed discussion about validation steps is presented in [1]. For standardized tests, this validation can be considered as unnecessary, while it is typically important for common tests. However, for specific condition of teaching and learning the procedure of validation become important and can be usefully employed for evidencing problems in student's understanding of the test itself rather than analyzing the test. For example, "face validity" [1] can be determined by a surface level, common sense reading of an instrument, and therefore, if a standard test like FCI, BEMA etc., would results as lacking face validity in some extend, it indicates that some concepts being tested are perceived form the students as not related to the subject matter. The other validity elements e.g., "content validity" "construct validity and "criterion-related validity", see [101] for f=definition and use, are not considered here because they do not represent a strong conjecture with student understanding on physics concepts which we are interest to analysis for the post-closure period. For the face validity a set of indexes can be very useful. We are listening to them shortly herein.

The Item difficulty index $P_i = \frac{N_{correct}^{Aswnres}}{N_{students}}$ measure the perceived difficulty of the item. It ranges [0,1]. The test difficulty is

$$P = \frac{1}{N_{items}} \sum_{i=1}^{n_{student}} P_i \tag{1}$$

evaluates the total difficulty. A common value is accepted to be in the range [0.3-0.9].

Item discrimination index measure the capability of the test to recognize the differences on students' knowledge. To calculate the item discrimination index (d), the whole sample of tested students is divided into two groups of equal size by referencing their individual total score according to the median total score of the group. For a given item, one counts the number of correct responses in those groups: namely, n_h , n_l. The discrimination index D of this item can be calculated as $d = (n_h - n_l)/\left(\frac{n}{2}\right)$ or more generally in the form $d_{x\%} = (n_{h,top-x\%} - n_{l,bottom-x\%})/\left(\frac{n}{\frac{1}{x\%}}\right)$ where x is the top percentile. If x=50, we have the first formula, if x=25 it means that we should consider first and fourth quarks in the sorted list and so on. It ranges in [-1,+1]. An item provides good discrimination if $d > 0.3$ [101]. Again, the average discrimination index of the test is

$$D_{avg} = \frac{1}{n_{items}} \sum_{i=1}^{N_{student}} d_i \tag{2}$$

The reliability index for each item is a measure of consistency of a single test item with the whole test. It is calculated as the correlation coefficient between the item scores (a dichotomous variable, 0,1) and the total scores (continuous variable) by following formula.

$$R = \frac{\widetilde{x_1} - \tilde{x}}{\sigma(x)} \sqrt{\frac{p}{1-p}} \tag{3}$$

where $\widetilde{x_1}$ is the average total score for those students who answered the item correctly, $\tilde{x}$ is the average total score of all the sample, p is the difficulty index and $\sigma(x)$ is the standard deviation score of the sample. A widely adopted criterion for this parameter is $r > 0.2$ [101]. Again $R_{avg} = \frac{1}{n_{items}} \sum_{i=1}^{N_{student}} R_i$ refer the average reliability index.

The self-consistency of the test (known as Kuder-Richardson index) is defined following the idea: If a test is administered twice at different times, one expects a highly correlation between the two test scores, assuming that the students' performance is stable and that the test environmental conditions are the same on each occasion. This parameter is given by formulae.

$$R_{test} = \frac{n_{items}}{n_{items}-1}(1 - \frac{\sum_{i=1}^{N} \sigma^2(x_i)}{\sigma^2} = \frac{n_{items}}{n_{items}-1}(1 - \frac{\sum_{i=1}^{N} P(1-P)}{\sigma^2}(4)$$

where $\sigma(x_i)$ and σ(x) are is the standard deviation of scores for item (i) and for the whole test. The final form is obtained by contemplating the fact that the variable is dichotomous (0 or 1). It is admitted [101] that typically r_{test} values above 0.7 are acceptable for physics test in the sense of reliability issues.

The discriminatory power known as Ferguson's delta. It measures how broadly the total scores of a sample are distributed over the possible range. By nature, the test should be designed to discriminate between students, so a good test should result in broad distribution of the scores. It is calculated by the following formula

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

Here the frequencies f_i is the number of occurrences of cases at each score. The values above 0.9 are considered good. Fallowing above indicators we can have a very adequate information on the testing features regarding students' capacities and abilities to answer it. We can use the indexes in two main applications. Firstly, by exploring the understanding and the perceived clearness of standard tests we get information about specific shortcomings on the teaching or learning process. When analyzing the test's responses, if indexes signal or evidence the disproportional difficulties as perceived from the sample, or if the values obtained fall outside typical intervals values, we qualify the sample as atypical, so more participating students are needed for a conclusive analysis. In another application, the assessment of the indexes

3. Analyzing test's indexes for standard FCI

We started the analysis of the conceptual knowledge by exploring the test's understanding clearness which herein is proposed to be evidenced indirectly from the index's values. Note that we have observed recently that contextual shortcomings were found typically significant among students that reported lack of or rare experimental and laboratory support in learning process due to the Covid-closure consequences in education process. Students' perception means knowledge, so a backward view of statistical indicators would indicate the specifics of the physics understanding for the group inquired or for the population they statistically represent. In this sense we can declassify as statistically non-representative the cases with abnormal indexes. We note that randomize procedure are not worthy enough because students that accept to conduct a physics test voluntary are not random if they will do it.

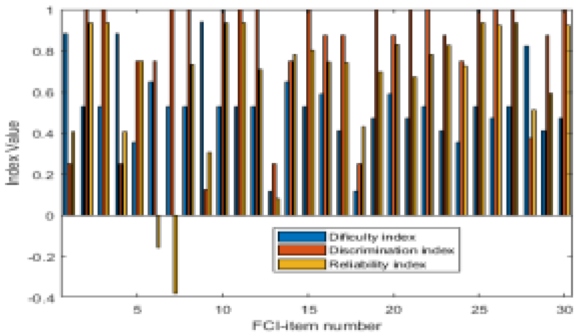

Figure 1. Item's indexes for the FCI-1 cases.

The first test named here FCI-1 has been conducted in a mixed group of bachelors first year students and high school students from four districts of the country. The test for the group of bachelor students has been held at the beginning of the academic year 2022, hence the knowledge were just the one gained form high school studies. Both have had the high school learning by the online system due to COVID restriction. It resulted that like in [10-12] the survey aimed in the evidences scientific knowledge issues due to the full on line restriction. However, it is likely probable that those results could reflect more general situation. After gathering the results of the test as recorded in the literature we assigned values 1 for correct answer and 0 for incorrect one. Next indexes are calculated by directly using formulas (1-5) above. It is seen immediately that indexes of FCI items' test showed high variability as seen from students' perspective. Knowing that FCI test is considered as standard instrument for knowledge measurement, the significant variability observed between different items indicates influential knowledge shortcomings. It resulted that the perceived difficulty index for majority of the FCI items is found in the normal range [0.3-0.9] but for question 13 and 16 the difficulty index is obtained abnormally low whereas for question number 9 the difficulty is high, above 0.9. By specifically analyzing those questions, it resulted there is no remarkable difference on conceptual or calculus issue, therefore we consider it as indicator of sample heterogeneity. Also, the reliability index for question (6) and (7) have been quite low. We argue that in our sample, 4 items that showed high deviance in at least one index might indicate the presence of conceptual confusion to the students. Next, we observe that, FCI test is not satisfactory self-consistent, that indicates high uncertainty in answering similar question, which obscure the representative power of the result itself (CI). It is also probable that when answering the test, students may have been guessing or in some extend, they may 'suffer' from contextual knowledge issues. Following above arguments, we classify this case as indicator of a relative conceptual shortcoming's possibility due to the high heterogeneity and for appropriate measurement Concept Inventory we should consider larger sample. Note that standard calculation of the sample size can not be used here because we were not able to guarantee a random sampling at all. Form a general point of view, we can use those findings herein to conclude that a CI scores and other conclusion based on this sampler should be considered with precaution, as long as a standard test resulted in inappropriate item' feature that affect directly the integrity of the measurement.

The second test named herein FCI-2 has been conducted on a mixed group of 109 students pursuing engineering and education branches at the Faculty of Natural Sciences, University of Tirana, for the academic year 2021-2022. The test has been held voluntary, and the number of participating students is considerable compared to the total students following those branches. In this case we observe that reliability index for 3 question (13,21,26) is negative reflecting the fact that for those question the answer has been found quite different (anti correlated) with the whole test scores. Also, two other items have a very low correlation with total scores. The item's discrimination index in this case is lower than 0.3 for five items but according to arguments in [101] this is not a bed value at all. However, those findings suggest again that the standard test has not been perceived uniformly by students' sample, indicating that significative heterogeneity might be present in the sample or in individual physics conceptual knowledge.

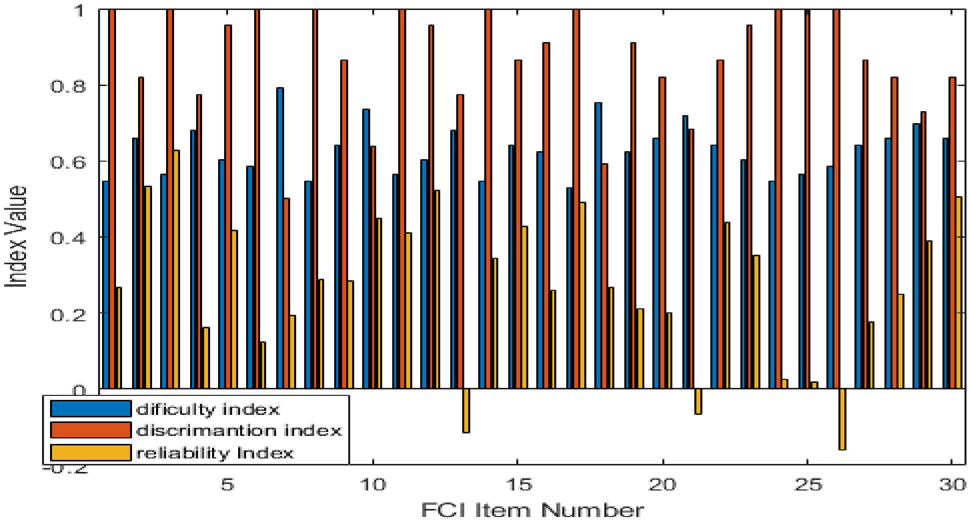

Figure 2. Item indexes for FCI 2 conducted on students form engineering branches.

This analysis suggests a comparative view on the whole test statistics and the significance of the indexes for knowledge measurement by the FCI test for the sample considered herein.

Table 1

Test Type	Details	Difficulty index	50 to 50 Discrimination index	Reliability Index	Self-consistency index	Discrimination power
FCI-1	Sampler from high school students	0.58	0.65	0.44	0.66	0.87
FCI-2	Sampler from students enrolled in natural sciences branches	0.53	0.8	0.65	0.87	0.89
Reference values		≥ 0.3	≥0.3	≥0.2	≥0.7	≥0.9

Also, FCI where not perceived as sufficiently discriminating among students' knowledge's level. We assumed that mixing students of engineering, technical branches and education branches has produced such a de-validation of typical standard test. Under these conditions, any result of CI measurement cannot be considered as representative for whole students in the country, despite our effort to improve the randomizing. Following specifics of university branches whom students come form, a better sampling procedure should have considered students from the top selected branches as the finances, justice, or medicine. We tried to conduct the test for those branches but unsuccessfully a very low number of students participated it. Also, a considerable number of high-level students follow the studies abroad, so we cannot get any information for this important group. However, if we want to identify problems and shortcomings, both those categories wound tell much, because they have hade high scores in their courses and therefore failures, shortcoming and knowledge defects are not characteristics therein. In the next paragraph we will discus about a third test with smaller number of students 55, that has been mechanically mixed with the FCI-2 resulting in better indexes. Note that we do not mix the data with FCI-1 because this test consisted on students that have had their studies in full online system due to pandemic closure.

4. The improved knowledge's level assessment by the Rash model

Following above discussion, we propose to review hereby the preliminary phase on the calibration of the knowledge measurement instruments by the Rash method introduced by Danish mathematician G. Rasch. Let have a short look in this well-known sociometric technique. Detailed arguments can be found in documents in [17], references [18-19] and a large psychometric literature. So, after recording results of the CI test by dichotomous or polytomous variables in a table $T(\#Students, \#Items)$, the matrix element $T\ (i, j)$ are considered initially as the probability that student (i) can solve the item (j). For multiple scale matrix elements, the probability is obtained similarly for each category, hence we should consider n matrices of the type $T(i,j)$, so for methodical purposes and reader chariness we will consider the binary case in the following. Now consider that it might happen that student (i) does not know the correct answer for item (j) and decide to encircle it by random, or the difficulties of the items differs remarkably whereas the scores awarded to them are the same etc., so the test must be calibrated for a correct measurement. Firstly, from the table T, one calculates two initial probabilities for solving the whole test by student (i) and for answering the item (j) by all students following the table:

	Probabilities	Response variables
Items	$P_{correct}(j) = \frac{1}{\#Students} \sum_{i=1}^{\#Students} T(i,j)$	$\beta_i = \ln \frac{P_{correct}(i)}{1 - P_{correct(i)}}$
Student	$P_{correct}(i) = \frac{1}{\#Items} \sum_{j=1}^{\#Items} T(i,j)$	$\delta_j = \ln \frac{1 - P_{correct}(i)}{P_{correct}(i)}$

Based on the average values δ_j, β_i the table of the probabilities estimate is generated according to the formula

$$P_e(i,j) \equiv P\left(x_{i,j} = 1 \middle| \beta, \delta\right) = \frac{\exp(\beta_j - \delta_i)}{1 + \exp(\beta_j - \delta_i)} \qquad (2)$$

Next, one replaces $T = P$ iteratively until a threshold criterion is met. In each step the variance is calculated straightforwardly and matrix elements with higher variance than a threshold called outfit and infit are identified. Those items present deviance from model assumption and should be analyzed in a separate step. So far, we know the representative vectors $\{\beta, \delta\}$ and final estimate probability $P_{estimate}(i,j)$. By using this last matrix in (1), we will get the calibrated instrument for measurement. The calculation procedure has been simplified in the lectures on [21].

As discussed above, by assessing the CI score and performing a dedicated analysis we can shed light in the knowledge shortage, conceptual shortcomings, and failures etc. Particularly the results of the FCI evaluation can be analyzed in the framework of the dimensions of errors in mechanics understanding as analyzed in [1-3] etc., going step by step toward the factors that affect conceptual understanding among students. However, in the process of measurement of the CI, the reliability, face validity of the test can impair the correctness of the analysis and conclusions. We may avoid this effect by omitting mechanically all items that presents an elevated difficulty from the student's point of view (roughly, the sampler we interviewed) based on the arguments that conceptual knowledge control in mechanics is still covered sufficiently by the reduced test (here 25 items). The second alternative will be the reduction of the heterogeneity which probably caused the observed values by adding more records. For this second alternative we have conducted a new FCI test at the beginning of the academic year 2022-2023 in the same branches and same year. The total number of students participating the test was 55, mostly from chemistry branch and engineering in mathematics and informatics. By doing so, we obtained an improved index picture as seen in the figure 3. We observe that for this sampler all items have their indexes in desired zone, meaning that student's perception and next their capability to solve them has nothing abnormal. Also, average parameters fall in the desired range. Following the above arguments, we conclude that this sample could be used for methodical measurement and analysis.

Table 2

Test Type	Details	Difficulty index	50 to 50 Discrimination index	Reliability Index	Self-consistency index	Discrimination power
FCI-Mix	Mixed sampler from students enrolled in natural sciences branches at 2022 (109) and 2023 (55)	0.49	0.91	0.39	0.81	0.94

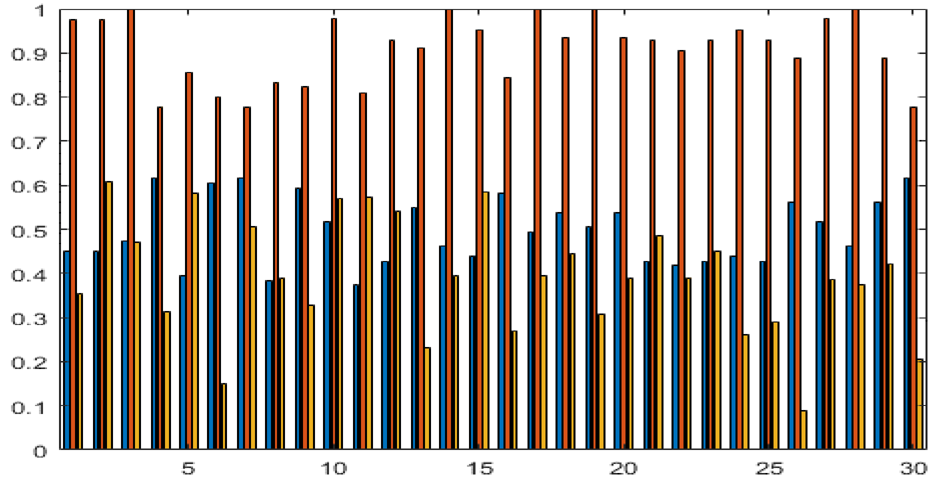

Figure 3. Indexes in the mixed group of the sampler in 2022 and in 2023

In this sampler we obtained $CI_{intial} = 15.22 \pm 0.5$ which correspond to "low understanding Force Concept" level according to the definition in [2]. However, this is a rude result as read form the initial bale of results. After performing the Rasch procedure, we have identified outfits and infits which must be excluded from the count because they impair the statistical meaning of the calculated values in this framework. Usually, the limit is set to 1.3 variances and in this approximation the average scores in the FCI for this sample is $CI_{statisticallyCorrect} = 14.37 \pm 0.19$ that correspond to 47.90 % that is considerably smaller than the limit of 60%. We may generalize the finding herein, it resulted that the physics knowledge level of students attending studies ate engineering and faculty of natural sciences is significantly problematic. Knowing that in our survey were missing students form medicine branches, finance and econometrics or justice. The result cannot be generalized for all students in the country in quantitative sense, but a broad estimation of the problems are possible however. We will consider it in the factor analysis part of this work.

5. Empirical factor analysis

In [12] we have analyzed the influence of some factors in the FCI-test outcome by focusing our attention on relevant causes related with the full online learning regime imposed by pandemic closure. Here we considered the set of categorical factors: *X= {weekly mathematics lectures, weekly physics classes, location of the school Laboratory Support, gender, category of the school, school profile, branch affiliation, textbook issues, teaching performance}.* In the model the response variable is logit(score) or the student ability for solving FCI test as calculated by the Rasch model hence we regressed this linear form $log\left(\frac{average_{score}}{1-average_{score}}\right) = \mathrm{AX}$ after performing the regression, we calculated back the estimated probabilities and next the expected average scores for the FCI test. According to the infrastructure condition we classified the location variables 1-3 for urban, suburban, and rural, the laboratory support on physics lectures is usually (1), rarely (2) and never (3); the school category consist in public (1) and nonpublic (2) , school profile is based on the orientation natural sciences (1), general (2) and social sciences orientation (3); the branch affiliation is set (1,2) according to the students preferences for physics, the textbook variable is set 1-3 respectively how students estimate the comprehensibility of the test and teaching performance also is classified in 3 level. We performed standard linear and logistic regression to evaluate the importance of each factor. By considering average scores as probability that students could solve the whole FCI test, we considered the extended sample as discussed above e.g., we worked out in the indexes' compliance first, and next we performed the Rasch model calculation. Finally, we used the student's calibrated ability as calculated in this step.

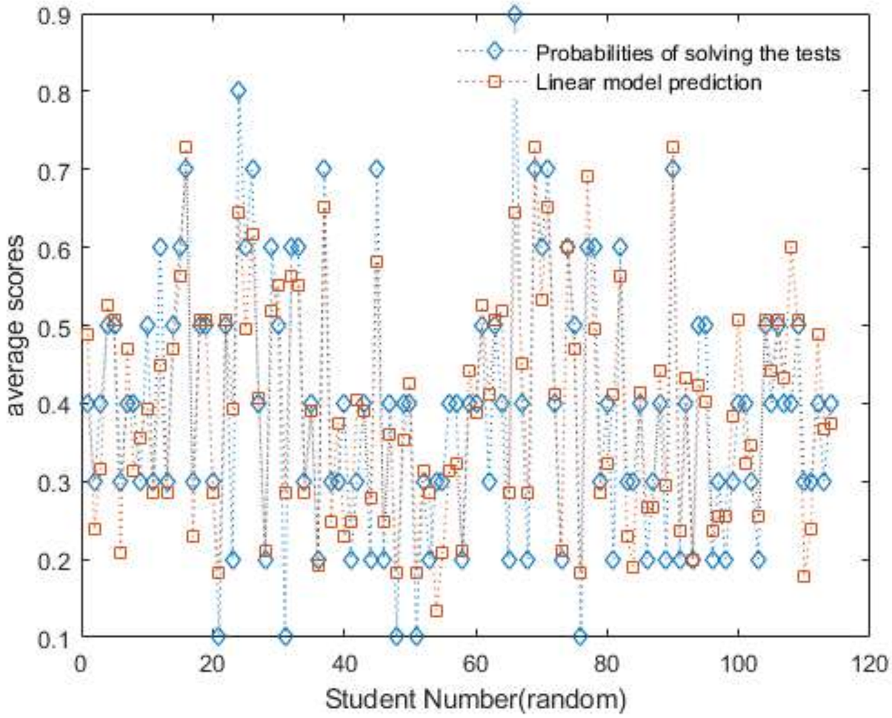

Figure 4. A logit linear regression for relationship between factors and $log(\frac{average_{score}}{1-average_{score}})$

We observed that the most influencing factor in the ability of the students to solve the FCI test for the students group interviewed herein are "*Laboratory Support", 'The Textbook Quality" and "Teaching Performance".* The variable '*weekly mathematics lectures*, *weekly physics classes, location of the school' have* similar influence and acceptable statistics, whereas the variable "*gender, category of the school", school profile, branch affiliation*" showed disputable statistics of the fit and for the sample herein they were filtered by the stepwise procedure ate the level 5%. In figure 4 it is represented the graph of expected average scores by using the model proposed Due to the randomizing issue, we consider those results as an reference estimation of the set of factors considered herein. However, we believe that after the step by step procedure presented in this work has provided a reliable and conclusive findings. It highlight the relevance of failure in laboratory work and demonstration for a successful FCI test, and also highlight that textbook quality and teaching performance are among most influential factors which entail conceptual knowledge failure that are evidenced by standard concept inventory test. For a full quantitative assessment of the factor weight, we should consider larger sample which remain for our future works.

6. Conclusions

Before employing a Standard Concept Inventory tests to measurement student's knowledge in physics for a group of students that have had their basic physics courses during pandemic closure or immediately after it, we used its reliability and discriminatory indexes to confirm this testing representative legitimacy in the sense that for further use of the outcomes of this test, the standard and certified test should have all times in the validity range. After assuring that the all items of standard FCI test have resulted with indexes in the desired range, the Concept Inventory has been evaluated. It resulted that for this group the level of knowledge in physics for high school students has resulted with remarkable impairment. By using again this re-certified test, we have identified major factors that influence the ability of the students to solve the FCI test. In this case the lack of laboratory work, textbook issues and teaching performance appear to be among very influential causes for the low CI level observed. Those results are strictly related with the immediate post pandemic time and students who attend the study in engineering and education branches.

References

[1] David Hestenes, Malcolm Wells, and Gregg Swackhamer. Phys. Teach. 30, 141 (1992); doi: 10.1119/1.2343497

[2] Noelle M. Crooks, Martha W. Alibali. Defining and measuring conceptual knowledge in mathematics. Developmental Review 34 (2014) 344–377

[3] Antti Savinainen and Philip Scott. The Force Concept Inventory: a tool for monitoring student learning. 2002 Phys. Educ. 37 45.

[4] 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. Jurnal Pendidikan IPA Indonesia, 7(4), 283-490. doi:https://doi.org/10.15294/jpii.v7i4.13490

[5] Heidelberg, 1996), 89-114.Antti Savinainen and Jouni Viiri, The Force Concept Inventory As A Measure Of Students_ Conceptual Coherence. International Journal of Science and Mathematics Education (2008) 6: 719Y740

[6] Ann O’Shea, Sinéad Breen and Barbara Jaworski. The Development of a Function Concept Inventory. Int. J. Res. Undergrad. Math. Ed. (2016) 2:279–296

[7] Assessing the Simplified Force Concept Inventory as Adaptation for English Language Learners Daniel Doucette International School of Latvia, Pinki, Latvia

[8] Phimpho Luangrath, Sune Pettersson, and Sylvia Benckert. On the Use of Two Versions of the Force Concept Inventory to Test Conceptual Understanding of Mechanics in Lao PDR. Eurasia Journal of Mathematics, Science & Technology Education, 2011, 7(2), 103-114

[9] Farida Huriawati , Nur Fitriani. Jeffry Handhika, Force concept inventory (FCI) representation of high school students (SMA & MA), J. Phys.: Theor. Appl. Vol. 1 No. 1 (2017) 29-34

[10] Dode Prenga, Silvana Miço,Polikron Dhoqina, Elmira Kushta. *An estimate of the effect of the online learning limitations during pandemic period on the physics conceptual knowledge-a case study*. International Conference on Advance Education. Istanbul, 28.07.2022

[11] Dode Prenga, Klaudio Peqini, Rudina Osmani, Elmira Kushta. Analyzing influential factors on physics knowledge shortcomings for high school student’s during covid closure and estimation of the opportune strategies to improve it. AIP (in pres)

[12] Elmira Kushta, Dode Prenga, Silvana Miço, Polikron Dhoqina,. (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.

[13] Lin Ding, Ruth Chabay, Bruce Sherwood, and Robert Beichner.Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment PHYS. REV. ST PHYS. EDUC. RES. 2, 010105 2006.

[14] R.K. Thornton, “Using large-scale classroom research to study student conceptual learning in mechanics and to develop new approaches to learning,” in *Microcomputer-Based Labs: Educational Research and Standards*, edited by R.F. Tinker, *Series F, Computer and Systems Sciences* **156** (Springer Verlag, Berlin,

[15] Planinic, M. (2006). Assessment of difficulties of some conceptual areas from electricity and magnetism using the conceptual survey of electricity and magnetism. American Journal of Physics, 74(12), 1143–1148.

[16] www.rashc.org

[17] Rasch Model Based Analysis of the Force Concept Inventory Planinic, Maja; Ivanjek, Lana; Susac, Ana Physical Review Special Topics - Physics Education Research, v6 n1 p010103-1--010103-11 Jan-Jun 2010

[18] Boone W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How?. CBE life sciences education, 15(4), rm4

[18] Camparo, J., & Camparo, L. B. (2013). The Analysis of Likert Scales Using State Multipoles: An Application of Quantum Methods to Behavioral Sciences Data. Journal of Educational and Behavioral Statistics, 38(1), 81–101.

[19] https://www.real-statistics.com/reliability/item-response-theory/building-rasch-model/

[20] Best Practices for Administering Concept Inventories Adrian Madsen, Sam McKagan American Association of Physics Teachers, College Park, MD Eleanor C. Sayre Department of Physics, Kansas State University, Manhattan, KS

[21] Robert E. Furrow1*† and Jeremy L. Hsu2†. Concept inventories as a resource for teaching evolution CURRICULUM AND EDUCATION Furrow and Hsu Evo Edu Outreach (2019) 12:2 https://doi.org/10.1186/s12052-018-0092-8

[22] Anderson, C. J., Verkuilen, J., & Peyton, B. L. (2010). Modeling Polytomous Item Responses Using Simultaneously Estimated Multinomial Logistic Regression Models. Journal of Educational and Behavioral Statistics, 35(4), 422–452. http://www.jstor.org/stable/40864755

[23] G. Kuder and M. Richardson, “The theory of the estimation of psychometrika test reliability,” Psychometrika **2**, 151 _1937_.

[24] J. Bruning and B. Kintz, *Computational Handbook of Statistics*, 3rd ed. _Scott, London, 1987_.