Available online at www.HEFJournal.org

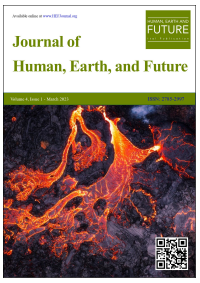
HUMAN, EARTH AND FUTURE Ital Publication

ISSN: 2785-2997

Review Article

Journal of Human, Earth, and Future

Vol. 4, No. 1, March, 2023



Enhancing Concept Inventory Analysis by Using Indexes, Optimal Histogram Idea, and the Likert Analysis

Dode Prenga [1*], Elmira Kushta [2], Fatjon Musli [2]

[1] *Department of Physics, Faculty of Natural Sciences, University of Tirana, Albania.*

[2] *Department of Mathematics, Faculty of Technical Sciences, University of Vlora, Albania.*

[3] *Department of Physics, Faculty of Technical Sciences, University of Vlora, Albania.*

Received 17 December 2022; Revised 05 February 2023; Accepted 17 February 2023; Published 01 March 2023

Abstract

Since the introduction of the Force Concept Inventory (FCI) in 1992, the CI tests have been widely used for measuring conceptual knowledge and for studying teaching issues in almost all disciplines and levels of study. A standard concept inventory analysis includes the design of a qualitative test, adequate realization of testing, calibration procedure, and comprehensive analysis of its findings. Usually, the CI test calibration is carried out through the Rasch sociometric technique, which is also used for calculating crucial indicators of knowledge such as item difficulties, students' abilities, and many more. Whereas the quality of the tests' design can be guaranteed by using certified and professional CI tests, the statistical adequacy of the testing merits critical attention before going on to the final step of the analysis. Also, the analysis of CI outcomes can be advanced by contemplating auxiliary tools and complementary techniques. In this framework, we propose to enforce the test index validity requirement for qualifying the CI outcomes as local or global. Specifically, the conclusions of CI analysis are acceptable for the whole population from which the sample has been extracted if the test's indexes comply with the validity requirements provided by the index theory. In the case when test indexes are out of validity range and re-conducting them is impractical for some objective circumstances or research specifics, we suggest injecting some new records into the existing one or mixing the data gathered from different samples until the new indexes are in the desired range. Using this methodology, we have reviewed our previous FCI tests, which were initially intended to demonstrate the impairment of learning in the physics discipline triggered by online learning during the pandemic closure. Through this renormalization procedure, we obtained a credible assessment of the understanding of mechanics and electromagnetism in high school students who followed online lectures during the pandemic closure. Also, by using indexes' validity as an auxiliary tool, we identified that for measuring the knowledge of electromagnetism in students enrolled in branches where physics is a basic discipline, a shortened version of the BEMA test was a better instrument than the corresponding shortened EMCI test. Next, we used the optimal histogram idea borrowed from distribution fitting procedures to identify the natural levels of students' abilities for solving a certain CI test. Another intriguing proposal presented in this work consists of combining an ad-hoc Likert scale assignment for usual errors in physics exams with the FCI designation of the basic commonsense confusion in mechanics for identifying their pairing features in common exams. We believe that the methods proposed herein can improve CI analysis in more general senses.

Keywords: Concept Inventory; Physics Knowledge, The Rasch Model; Likert Scale.

* Corresponding author: dode.prenga@fshn.edu.al

http://dx.doi.org/10. 28991/HEF-2023-04-01-08

1. Introduction

The assessment of students' acquaintance with the sciences is realized through procedural and conceptual knowledge tests. A procedural knowledge test intends to evaluate the ability of a student to solve problems step by step, his or her fluency to employ instructed algorithms, etc., so they are commonly used in everyday pedagogic activities. A conceptual knowledge test aims to assess students' knowledge of fundamental relationships between variables and features of the system under study. Comprehensive discussion about those tests, their specific features, and their application is presented in dedicated literature such as [1–6], etc. The success of concept inventory analysis in physics disciplines, which began with the introduction of force concept inventory in 1992 [7], has promoted the use of similar exemplars in chemistry, biology, astronomy, material science, mathematics, etc. [8]. Concept inventory analysis is based on the data of the respective test, which serves as a measuring instrument of knowledge. A CI test consists of multiple-choice items, and therefore, the choice of the alternative is also a decision-making process. Consequently, a clearer picture of student understanding is achieved when the CI analyses are combined with item response theory (IRT) arguments. In this regard, when designing an efficient CI test for a certain discipline, the researcher must also consult the dedicated IRT literature [9–12].

Many scholars prefer to avoid improving test drafting by utilizing certified ones, which facilitate focusing on the students' group features. Also, based on sociometric arguments, a CI test must be calibrated. This procedure is accomplished in practice through the Rasch technique [13], whose outputs are also important quantities that characterize the system, such as perceived difficulties, students' abilities, etc. The Rasch calculation and its outputs involve averaging across participants' answers and across items, so for ensuring a credible CI conclusion, the statistical features of the sampling process should be considered with precaution. Most concerns regarding tests as social measurements can be addressed by utilizing standard guides or practical solutions proposed for specific cases provided by the dedicated literature [14–16]. However, their realization could be expensive, time-consuming, or even impossible in practice, which reveals the idea for alternative solutions in several applications. Notice that by nature, the study of knowledge issues is a dynamic and complex process, so the concept inventories have been expanded and improved impressively regarding their uses, components, and analysis. For example, in Smaill & Rowe [17], Raduta [18], and McColgan et al. [19], interesting developments of the CI analysis for several parts of electromagnetism are presented; in Laverty and Caballero [20], it is underlined the need for aligning standardized assessments with modern learning goals such as engagement in scientific practices; in Klymkowsky & Garvin-Doxas [14], it is highlighted the use of interactive assessments in which students are called upon to identify and justify their assumptions; in Sands et al. [8], it is stated that concept inventories are not perfect tools, suggesting improvement from the measurement perspective, etc. In this view, there is little room for methodical novelties. But on the other side, we believe that there are still opportunities for improving the statistical significance of CI outcomes, for assisting standard CI analyses with other techniques for increasing their explanatory ability about hidden or mixed factors' effects, etc.

Remember that a CI begins with the testing procedure, in which the fulfillment of statistical requirements is hard. For example, when conducting the test with voluntary participation, randomness is not respected. Regarding this problem, we have proposed an ad-hoc solution for repairing the descriptiveness feature of our recent CI test, which could be applied elsewhere. Specifically, in Prenga et al. [21], Kushta et al. [22], and Prenga et al. [23], we have remarked that due to statistical inadequacies in the sampling process, the conclusions of the FCI analysis therein should be taken as indicators, not representative of the whole students' population initially considered in those studies (e.g., those who followed their lectures online during pandemic closure). When reviewing them, we wanted to assess the level of knowledge in physics for all students that had their classes online during 2020–2021 and estimate the impairment in knowledge inflicted by mandatory online learning. In this attempt, redoing the test was not possible for a couple of reasons. However, the 203 test responses administered by the initial tests [21] contain valuable information for the system under scrutiny, which we didn't intend to lose. Based on the standard CI analysis performed therein, we didn't identify credible evidence to disqualify or generalize their outcomes. Frankly speaking, misfit occurrences evidenced through the Rasch model calculations can be used for discussing statistical incompatibilities observed. But those events represent overlapped effects of statistical nature and CI test perception issues, which cannot be separated in the final stage of the CI analysis. For resolving those issues and similar others, we proposed to use indexes' validity range definitions discussed in the literature [24] as auxiliary tools for the CI analysis. Notice that the test validity issue introduced in Aubrecht and Aubrecht [25] is a comprehensive and delicate concept, but since we employed a certified CI test in that work, we have restricted this notion to having index values in the validity range according to [24]. The replacement of statistical sampling adequacy with indexes' validity is debatable, but we have used it herein as an opportunistic solution. It is worth mentioning that indexes are statistical indicators of the testing, including readability issues, reliability, discriminatory power, and several other features [24], which are much more descriptive than misfit occurrences as evidenced through Rash calculation. Therefore, we have considered them an appropriate tool for identifying and resolving statistical issues in our example and propose to use them for CI analysis improvement, too.

In the second example, we have considered a common concern regarding knowledge assessment: which type of test is more efficient for measuring and studying students' acquaintances in a certain field of physics. In this case, we have

re-examined our recent EMCI, which consisted of a shortened version of the Electromagnetism Concept Inventory, discussed in Notaros [26], Hansen and Stewart [27], and McColgan et al. [19]. This test has been conducted with a group of students who have had at least one school year in the online system and, at the time, preferred to follow their studies in branches where physics is a basic discipline. We observed that the level of knowledge in electromagnetism measured by our EMCI test was notably lower than our estimations based on official exams' scores. In this case, we hypothesized that a test based on proper conceptual knowledge questions might not be the right instrument for measuring knowledge in our students for a couple of reasons, which we will discuss below. Therefore, based on the ideas proposed in this work, we have analyzed the indexes of this test, which resulted out of validity range. We next guessed that the compatibility of the EMCI test with our students' custom preparedness and learning stereotypes was questionable, or that the current formulation of the test was not adequate, etc. Disregarding the causes, we have considered them inappropriate tools for measurement in our system. Alternatively, after performing a simplified version of the BEMA test, which includes some calculations, Hansen & Stewart [27] observed that indexes were in the validity range. Despite the fact that absolute scores were again unsatisfactory, we have admitted that for measuring knowledge in the current student's group, the BEMA test works better, but the most relevant conclusion is that the use of indexes' validity has enhanced the knowledge measurement.

Another concerning didactical question regarding students' abilities measured by CI tests is the assessment of their knowledge levels or classes. Such information can be accessed naively by arranging the calculated individual abilities in the empirical classes based on a fixed grade system. This classification does not accurately reflect the features of the current group of students, so we proposed to employ the optimal histogram idea borrowed from the distribution fitting procedures. We argued that the bin size of the abilities obtained through this procedure has better described the groups' units of the abilities, and next, the solution is proposed for a larger application.

In our last proposal for advancing CI analysis with the assistance of other techniques, we employed the Likert scale idea to perform a thorough investigation of dominant errors occurring in common procedural exams. In pedagogical practice, we face ambiguities in identifying the most blamable error for common exam failures. Also, the weightiness of the calculus and conceptual errors in physics exams is a debatable issue. Kushta et al. [22] devoted a concrete investigation and observed that failures on physics exams were dominated by physics conceptual knowledge shortcomings, and the calculus incapacities weren't found to be principal causes for them. Based on this indicatory finding, we have proposed to use a Likert scale idea to assist the CI analysis in identifying the pairing between six commonsense concepts in mechanics with conceptual deficiencies or calculus errors [21, 23, 28]. To advance the investigation, we initially decomposed the error state into four elements by their dominance, and then their significance was identified by performing the Likert analysis. In the first application herein, we utilized the FCI test capability for evidencing common-sense errors in mechanics. In the second application, the same idea is used to analyze the error configuration and its weightiness in a more general physics exam. After this ad-hoc resolution of our current concern, we proposed to utilize it for more general purposes. Notice the quantitative results provided through the working examples and evidence features of the system analyzed herein, but our focus and intention are to demonstrate the effectiveness of our proposals for CI analysis enhancement. The calculations presented for illustration through this research were performed using our ad-hoc algorithms written in MATLAB, but interested readers can consider dedicated commercial software [29] or statistics guides [30].

2. Methodology and Data Elaboration

This work has been initiated as an attentive approach to rectifying statistical inadequacies that we encountered in the organization of the CI tests that we developed with high school students in Albania, with the aim of identifying the knowledge impairment inflicted by online learning during the pandemic closure. Next, we have advanced this idea with more proposals for improving CI analysis in a broader sense. So, let's briefly introduce the problem that we wanted to resolve. The first investigation in this framework has been the measurement of knowledge in classical mechanics using the FCI test. This test was carried out with the voluntary participation of 213 students from four secondary schools, selected in four cities in the country [21]. After performing the CI analysis according to the standard methodology described in Anderson et al. [31], Planinic [32], Planinic et al. [33], etc., the conclusions found are considered indicative only, being aware of the statistical problems related to the testing [21, 23]. Regarding the initial goal of the test, we have considered the outcome of the test important, informative, and potentially useful for a generalized analysis due to the satisfactory number of participants and the fact that it was carried out at the right time. The research question was, to what extent can we generalize those results, and how can we attain a credible estimate for the knowledge of all students who attended high school during 2020-2021. Notice that full-scale testing encountered objective obstacles. Firstly, the target group had to "inherit loyally" the consequences of the closure. Considering that this review was initiated in 2022, only students enrolled in university during 2021–2023 fulfilled this condition. From them, those following the 2nd year at university had advanced their knowledge in physics during their study, so they were found not appropriate.

Skipping other details of the current solution, which we will describe in the corresponding paragraph in this work, the encountered problem raises a more general question regarding CI analyses: how to avoid unfounded generalizations

of the conclusions and how to validate the outcomes of an existing test if redoing it is not opportune. Notice that statistical inconsistencies in the realization of the tests are quite common. First, voluntary participation on the test is not in accordance with the requirement for random selection of participants in the survey, which is fundamental according to the statistics literature [34, 35]. In this case, all individuals who do not wish to participate in the survey carry important information regarding the entire population under investigation, inflicting sample-biased measurement according to McCombes [15] and Martínez-Mesa et al. [36]. Referring to our example, students enrolled in the branches of economics, social sciences, medicine, etc. were practically not reachable by us, even because they showed no willingness to participate in such tests at all. Also, students' knowledge depends on the location or categories of the school, etc., but similar circumstances are commonly present in such surveys, so the problem is quite general.

Regarding redoing the CI test, we should not exclude the costs of the realization. But when we study phenomena related to specific periods, redoing the test becomes impossible. So, for resolving statistical issues in our CI analysis, we propose to use indexes as auxiliary tools. We argue that this technique can be fruitful when using certified CI tests, like FCI [7], EMCI [26], BEMA [24], Chabay [37], McColgan et al. [19], etc.; otherwise, problems arising from test features and defects would obstruct the correctness of this logic. Notice that indexes' theory aims to analyze the tests from a statistical perspective (see below); hence, the suggestion to use them to resolve our problem seems reasonable. In short, if indexes lay in the valid range according to Ding et al. [24], the CI findings can be considered acceptable and representative for the whole population under investigation. If not, we propose to add more records gathered by an adequate supplementary test, attempting to ensure that indexes of the composite data would be in the desired range. In general, we believe that performing an indexes' check-up procedure before Rasch calculation and CI assessment would improve the CI analysis itself. Notice that a CI analysis comes together with the Rasch calibration routine [30]. All occurrences with high deviances between current (raw) values and estimated ones by the Rash techniques are marked as outfits and are analyzed specifically [30, 33], as described below. In this regard, anticipating CI analysis with index verification will prevent high-rate misfit occurrences in final outcomes.

Bringing all those arguments together, we have improved our CI analyses by using indexes' validity verification. Procedures. In our first application, we found that the indexes of our recent FCI test were out of validity range. To obtain valid indexes, original data have been supplemented with new records from a low-scale and appropriately conducted new test, which is described below. Similarly, the findings of an EMCI test have been qualified as local because their indexes were out of the validity range. Next, indexes of the BEMA test conducted in the same category have resulted close to the validity range. We used those results to qualify BEMA as a more convenient tool to measure the knowledge of electromagnetism for our students. It is worth saying that, for portraying this idea, we are also based on a synthetical analysis of the comments and arguments about Rasch technique features provided in Liu [38] and Coletta and Phillips [39]. Before presenting our working examples, we will briefly introduce a summary of index definitions and the Rasch technique.

2.1. Indexes of the Concept Inventory test

The test's indexes consist of statistical estimators of testing integrity, significance, validity, discriminatory power, difficulty measure, etc. [24]. In this section, each index is explained briefly; for more information and discussion, see references [24, 25]. Here are their definitions and calculation formulas.

The perceived ***difficulty index*** (or easiness index). The item difficulty index and the test difficulty as their average, are given by Equation 1:

$$p_{item} = \frac{N_{CorrectAnswers}}{N_{students}} \tag{1}$$

$$P_{test} = \frac{1}{N_{items}} \sum^{n_{items}} P_{item}$$

According to Ding et al. [24], acceptable values for the difficulty indexes (1) are in the range [0.3-0.9].

The ***item discrimination index*** measures the capability of the item to recognize the differences between students' knowledge. The average discrimination index indicates the same feature for the entire test. They are given by Equation:

$$d = \frac{f(N^{x\%}_{highestScore} - N^{x\%}_{lowestScore})}{N_{students}} \tag{2}$$

$$D = \frac{1}{n_{items}} \sum^{items} d_{item}$$

The ***item's reliability index*** (known as the point biserial coefficient) is a measure of the correlation (the consistency) of a single item with the whole test. The test reliability index is the average of the items' reliability indexes. Reliability indexes are calculated by the equation:

$$r_{item} = \frac{\widetilde{x_i}-\tilde{x}}{\sigma(x)}\sqrt{\frac{p}{1-p}}; \quad (3)$$

$$r_{avg} = \frac{1}{n_{items}}\sum_{i=1}^{n_{items}} \quad r_{item}$$

Here $\widetilde{x_1}$ is the average score for those students who answered the item correctly, $\tilde{x}$ is the average of total score for the sample, p is the item difficulty index and $\sigma(x)$ is the standard deviation of the total score [24]. Every item in a test should be correlated with the total score, but for a test with large number of times it is admitted that $r_{item} > 0.2$ is acceptable. So, desired, or valid values of the test reliability index are $r > 0.2$ for both terms in Equation 3

The ***self-consistency of the test*** (Kuder-Richardson index) is calculated by Equation 4:

$$\rho_{test} = \frac{n_{items}}{n_{items}-1}(1 - \frac{\sum^{n_{items}} \sigma(x_i)}{\sigma^2(x)} = \frac{n_{items}}{n_{items}-1}(1 \frac{\sum^{n_{items}} P_{item}(1-P_{item})}{\sigma^2} \quad (4)$$

where $\sigma(x_i)$ and σ(x) are the standard deviation of scores (x) for the item (i) and for the whole test. The values $\rho_{test} > 0.7$ are acceptable for the in-group measurement. The reliability values higher than 0.8 are acceptable for individual measurement. Notice also, that different tests have different criteria for the reliability index, according to their purposes. If a given test has a reliability index greater than 0.7, we have an indication that the group is sizeable enough according to the test data, and therefore the test is reliable.

The ***discriminatory power of the entire test*** (Ferguson's delta) measures how broadly the total scores of a sample are distributed over the possible range. If a test is designed and employed to discriminate among the students, one would like to see a broad distribution of total scores [24]. The discriminatory power is given by the relationship

$$\Delta = \frac{N^2-\sum_i^{n_{items}} f_i^2}{N^2-\frac{N^2}{n_{items}+1}} \quad (5)$$

where the frequency f_i is the number of occurrences of each score obtained, and N is the number of students participating in the test. The values $\Delta > 0.9$ indicate good discriminatory power.

In principle, if the values of indexes resulted outside the desired range, the test must be re-edited. The out-of-range values for indexes characterizing a certain CI test would indicate teaching or learning problems. In this regard, the indexes of a certified CI test would be in the validity range, provided that all its topics are lectured sufficiently. Herein, we propose to use this feature in the following application. If indexes of the standard or certified CI test result outside the validity range, we argue that the disproportional difficulties, inconsistencies, discriminatory, and reliability issues related to those values are consequences of an inappropriate sample. Under those circumstances, we propose to pursuit a "sample enlargement" attempt to achieve the validity of the testing pair *{sample, test}.* After accomplishing this step, we can follow the core and standard CI analysis by using the Rach technique. In the following paragraph, we are also briefing on the Rasch techniques. It is used for calibrating the CI test as an instrument and for calculating CI test outcomes.

2.2. Introductory Elements of the Rash Analysis

Rasch analysis is a psychometric technique that was developed to improve the precision with which researchers construct instruments, monitor instrument quality, and compute respondents' performances [40, 41]. Notably, it provides calibrated assessment of the concept inventory outcomes: the student ability to solve the test, the items' difficulties, the estimated probability for a student to solve an item, pathological behaviors as guessing, etc, [31, 32, 42]. Here is the description of its core calculation procedure. Initially, the answers of the CI test are recorded in a matrix $T(i,j) = (0,1)$ by assigning (1) for correct answer and (0) for incorrect one, [30, 32]. Unanswered questions are left blank. One calculates student's average scores obtained for the test $P_{correct}(i)$ and the average scores that all students realized for the item $P_{correct}(j)$

$$P_{correct}(i) = \frac{1}{NumberItems}\sum_{j=1}^{NumberItems} T(i,j)) \quad (6)$$

$$P_{correct}(j) = \frac{1}{NunmberStudents}\sum_{i=1}^{NunmberStudents} T(i,j)$$

Next, the student's ability to solve the test β_i and the item's difficulty perceived by all students δ_j are calculated by Equation:

$$\beta_i = ln\frac{P_{correct}(i)}{1-P_{correct(i)}} \quad (7)$$

$$\delta_j = ln\frac{1-P_{correct}(i)}{P_{correct}(i)}$$

Quantities in Equation 7, are measured in logit units, which are linear and homogeneous [32, 33, 41]. Using them, the probability that student (i) having the ability β_j could solve the item (j) whose difficulty is perceived δ_i is

$$P_e(i,j) \equiv P(\beta,\delta) = \frac{exp\,(\beta_i-\delta_j)}{1+exp\,(\beta_i-\delta_j)} \quad (8)$$

Similarly, all above parameters can be evaluated for polytomous variables or Likert-scale assessment for the answer [31-33], but we won't use them in this work. The estimated value $P_e(i,j)$ represents better the probability of success than the initial (and naively evaluated) $T(i,j)$, [32, 41]. So, the old matrixes are replaced iteratively by the improved (estimated) ones, until the sum of squared residuals between current values and original ones is reduced below a threshold value [29, 30, 43]. The final values for difficulties, abilities and probability estimates obtained by this procedure represent better the students' responses for the CI test. According to Planinic et al. [33] and literature provided therein, this model has item and person invariance properties, unlike the commonly used "percentage correct" statistics which are strongly sample dependent and test dependent. The model provides also linear units for measuring the quantities of interest [32, 40, 42], which is plausible especially for their interpretation. For example, the difference between the abilities of the students (n) and (m) with $\beta_m = 3\ logit$ and $\beta_n = 2\ logit$ respectively is the same as the distinction between student (k) and student (l) when $\beta_k = 1\ logit\ \beta_k = 2\ logit$ whereas the knowledge gap between students A and B which won 90/100 and 70/100 scores, and C and D which won 30/100 and 10/100 respectively is obviously different [33, 41, 42], for detailed arguments.

On the other hand, the differences between the final and original values contain interesting information and can be viewed as another estimator of CI testing. Elevated (squared) deviances called misfits are classified into outlier-sensitive fit (outfit) and information-weighted sensitive (infit) occurrences. Outfits identifies unexpected students' answering on items that are relatively very easy or very hard for them (and vice-versa). The large outfit value of an item indicates that people who are far in ability from the difficulty of the item have responded in an unexpected way. For example, if the outfit of a hard item is large, it means that several students of low ability have answered it correctly, and a large outfit value for an easy item means that unexpectedly, some students of high ability have failed to solve it correctly. A large item' infit value indicates that some people of the ability that is close to the difficulty of the item have not responded in a way consistent with the model [42]. There are other interesting and important indicators and estimators related to the Rasch analysis [29, 43], etc., which we are skipping for now. For our purpose, we remark that the analysis of the misfit occurrences in a CI test includes various points of view. So, the presence of the infits and outfits would urge the reformulation of the corresponding item and might signal teaching and learning issues, insufficient coverage of the subjects, statistical inadequacies, etc. Also, the Rasch procedure identifies the "pathological individual" cases, which correspond to significant differences between original $T(i,j)$ and estimated probabilities $P_e(ij)$, which are named as guessing events. Guessing behaviour could be random, but potentially could be related also with the unfair conduct (the case $P_e(ij) \gg T(i,j)$). As we depicted before, outcomes of the Rasch model are based on several statistical assumption and calculation, therefore, some misfit occurrences are direct consequences of the inappropriate sampling or small size of the group interrogated. In this regard, performing some precautionary statistical steps like indexes validation before the core CI analysis, is expected to improve its results. Considering those elements, we conclude this descriptive paragraph by proposing the use of the test' indexes validation as preliminary step on the CI analysis, to improve the statistical significance and reliability of its outcomes and conclusions.

3. Improving the Concept Inventory Analysis by Employing the Test Indexes

As a first working example, we reconsidered the results of our recent CI test in mechanics, which was mentioned above. In Planinic et al. [33], it has been urged that the construction of a measurement instrument with the Rasch model is a systematic process and not a routine application. Liu [38], arguing that Concept Inventory items used in Rasch analysis should be constructed purposefully according to a theory and empirically tested through Rasch models to produce a set of items that define a linear measurement scale. Following those arguments, after conducting a standard FCI test, we have organized another version in simplified form, named SFCI, which is based on the version introduced in Umarov et al. [44], but further simplifications have been made according to the principal gaols of our studies [21–23]. The original Force Concept Inventory test was conducted on 213 students from four major districts of the country. By implementing the Rasch technique for this CI analysis, we have obtained a level of understanding of mechanics that was around 35% in terms of the original FCI analysis [7]. It is projected as atypically low according to our general knowledge and teaching experience, despite the negative effect of the unscheduled shift to online learning because of the pandemic closure. In the work presented by Prenga et al. [21], we have noticed that such findings should be considered indicators, and further analysis and testing are needed. By assuming that the online lecturing might have significantly reduced the laboratory support and demonstration capability when teaching, thereby imposing additional contextual confusion among students, we tried an easier and contextual-confusion free test [21]. It has been drafted in collaboration with some high school teachers and based on the simplified FCI called SFCI, introduced by Popp and Jackson [45] and elaborated further on in Stoen et al. [46]. We observed that the level of knowledge measured by this second instrument improved to the new value of 48%. It resulted in a 22% gain. Calculated Equation 9:

$$g = \frac{\%SFCI_{scores} - \%FCI_{scores}}{100 - \%FCI_{scores}} \quad (9)$$

which we have adopted from the scores' gain discussed in Planinic [32], Planinic et al. [42], Coletta and Phillips [39], etc. In Kushta et al. [22], we argued that the low score level obtained in the original FCI test was a consequence of the lack of demonstration and laboratory support during the online lecturing, etc. However, being aware of some questionable issues in a statistical sense, those findings were qualified characteristics for the group that participated in the tests and potential indicators for a larger-scale phenomenon. As noted above, the voluntary participation in the test denied us to gather information from students who weren't enthusiastic to participate in a physics test with 30 not-trivial questions! Also, several students who participated in the test reported that they had missed some lectures due to an internet connection issue, etc. Later, we gathered more information about the coverage of the physics program during online learning, which merited more attention. However, the number of participants has been satisfactory (213), and for practical reasons, we intended to retain these CI test outcomes as important. Note that in principle, the Rasch analysis can be conducted with small datasets [33], but randomizing the sample should be fulfilled correctly, which in turn is not easy. Under such circumstances, we performed the indexes' analysis according to the proposal of this work. So, we observed that the self-consistency index for the original FCI test was 0.58, which is lower than the limit of its validity (0.7), according to Ding et al. [24]. On the other side, the SFCI test had its self-consistency index at 0.68, very close to the validity limit. The other indexes for both FCI and SFCI were found inside the validity internal. Therefore, we considered the results of the SFCI to be statistically more representative than those of the FCI. By the way, the original FCI test findings are very important for the group that we interviewed, but because one of its indexes does not comply with the validity prerequisite, we cannot generalize them for the entire population of high school students. Consequently, the SFCI score is considered a better indicator of the students' knowledge in physics for the period of the measurement (in the year 2020).

Considering the simplifications that we made to obtain our ad-hoc SFCI, we cannot consider its outcome as definitive and quantitatively accurate, but it can be recognized as the best value of the knowledge indicators. So, we concluded that students' knowledge grade in mechanics for the period under scrutiny should be taken in the interval [9%–48%]. Surely this assessment has debatable uncertainty, but we believe that it better represents reality. Nevertheless, it is low according to the general agreement in Hestenes et al. [7], and therefore it confidently mirrors the negative effects of the compulsory and unscheduled online education system during pandemic times. It reaffirms the qualitative conclusions reported in Kushta et al. [22] and Prenga et al. [21], but it is more confident from the statistical point of view. In the following, we will demonstrate a direct and efficient use of the index idea to generalize the FCI test outcomes. We considered herein a FCI measurement on high school students who have expressed the preference for studying at the university branches where physics constitutes a basic part of the curricula. Again, the analysis aimed at evidencing the consequences of online education in physics understanding, so we started the discussion based on our previews of the FCI test. So, from the FCI test mentioned above, we selected 101 records of the students who belong to our interested group (the FCI-1 on Table 1). Before proceeding with Rasch analysis, we obtained that the discriminatory index for this sub-test was 0.66 and the self-consistency index was 0.87, both lower than the corresponding validity values of 0.7 and 0.9, respectively (Table 1).

According to the indexes' theory, this test is considered inappropriate for general conclusions, and since FCI is a certified test, this deficiency is triggered by sampling issues. Notice that a fresh FCI test on our group of interest was not conceivable for objective reasons. Then, based on the idea elaborated above, we conducted an additional test, which reached 55 students enrolled at the Faculty of Natural Sciences, University of Tirana, in the academic year 2022-2023. We did not intend to retrieve comprehensive information by using this test separately, nonetheless. The sample belongs to a single faculty, and its size is small. The idea is to use it as a complementary test for the old one, according to the proposal of this study. So, both data were merged into a compound set, attempting to determine the validity of the indexes. The mixed data were named the FCI-2 test for reference (Table 1). It resulted in the reliability index being improved from 0.66 to 0.87 and the discriminatory index of the mixed test becoming close to the validity limit of 0.89~0.9. Considering that other indexes are in the valid range and that the self-consistency is well above the limit (0.7), we qualified the composite test as suitable for further analysis. Particularly, the level of knowledge in mechanics measured in the FCI-2 test resulted in ~48%, which corresponds to the value found by the SFCI test above. Specifically, the 35% knowledge level claimed by the old FCI test is classified as the characteristic for the group of students that partaken it. Hence, we assume that a better estimation for the level of understanding in classical mechanics of high school students for the period of observation would be taken in the vicinity of 48%, and this is nearly true for those students that preferred to follow their studies in the branches where physics is a basic subject. From a general didactical point of view, knowledge in physics has been impaired because of the pandemic closure, but not destroyed. Although there were no previews of the FCI test measurement, in supporting this conclusion, we considered our long teaching experience and informal consultation with high school professors of physics, which favor the average knowledge in physics at around 55% in terms of definitions in Hestenes et al. [7]. Note also that the goal of this work is to provide a practical method to improve CI measurement rather than go deeply into its concrete outcome. Also, we have an additional argument that the SFCI outcome is a good estimator of knowledge level, and therefore, the level of 48% is

achievable, starting from the bottom of only 35% level of knowledge, so the 22% gain obtained by a mechanical maneuver in (9) is also an estimator of a general gain. In this sense, we believe that the gain in conceptual knowledge in mechanics would improve by around 22% if the strategies and methods of teaching physics in high school could cure contextual unclearness issues.

Table 1. The indexes for FCI-1 and FCI-2 tests, the self-consistency in FCI-1 was out of desired range. By adding Validating some records it is improved to the validity range. Other indexes have remained in desired range

Test Name	Details	Difficulty index	50 to 50 Discrimination index	Reliability Index (Point Biserial)	Self-consistency index	Discrimination power
FCI-1	Sampler from high school students	0.58	0.65	0.44	0.66	0.87
FCI-2	Mixed Sampler: FCI-1+students enrolled at 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

3.1. A Discussion about Using EMCI Test for Measuring Knowledge in Electromagnetism

In this paragraph, we will discuss the enhancement of an CI analysis and measurement of knowledge in electromagnetism by involving additional statistical tools. Here, the original study has been designed for spotting difficulties in the main topics of electromagnetism and for evidencing the after-course gain in physics knowledge. For this purpose, a shortened and simplified version of the EMCI test containing 20 items was initially conducted on the first-year students of the General Chemistry and Engineering on Mathematics and Informatics (EMI) branches, Faculty of Natural Sciences, throughout the years 2021 and 2022. The EMCI test used herein has been drafted based on the literature of Ding et al. [24] and Notaros [26], and additional arguments are provided in [17–19]. The physics course takes place for 60 lessons in one semester for the EMI branch, and for chemistry, there are 120 classes in two semesters, including 30 classes for laboratory work. However, electromagnetism's topics and lessons' classes matched. For a factorial analysis, we should consider the differences between their knowledge inherited from their previous education, etc., but the investigation herein was not focused on such details.

The EMCI test was organized for the courses in 2021 and 2022. After performing Rasch calculations for every exemplar, 3-5 items resulted in outfits 3-4 infits. We observed that misfitted items belong to the magnetic flux and field and to the relationship between the electric field vector and electric equipotential lines or surfaces. After we got such information from the standard CI analysis, we considered the tests' indexes again to enhance the analysis and interpretation of the results. First, we observed that the difficulty index for the misfitted items was hound lower than the validity limit, $P_{items}^{2021,2022} \sim [0.1 - 0.2] < 0.3$. Under such circumstances, all the findings are classified as local, sample-based outcomes. Also, we underlined the fact that misfitted items (by the calibration procedure) were perceived very hard on an absolute scale. Up here, we are not clear on the nature of the factors that imply misfitting statistics for certain subjects because the interpretation of Rasch analysis outcomes contemplates both tests and sample influences. We considered the index findings instead. After consulting teachers of physics from several high schools, we learned that some of the topics that appear to be misfitted above should have been instructed with laboratory activities, necessarily as part of the program, but this was not realized in many schools during the online learning period. This information helped us to qualify the corresponding items in the ECMI test as lacking "face validity" case. Such subjects are elicited as responsible for obstructing the Rasch analysis and the sources of the high misfitting ratio observed. Following this argument, accordingly, the three hardest items have been excluded from further CI calculation based on the procedure suggested in Zaiontz [30]. Nonetheless, the items related to the magnetic field and electric field concepts persisted to be difficult, despite the filtering adopted beforehand, and indexes remained out of the validity range. Considering also that we applied simplifications to the original EMCI test to get our draft used in this investigation, we realized that the EMCI test was not the right tool for measuring electromagnetism's understanding of the group of students considered in this example. Again, by using indexes to assist the concept inventory analysis, we were able to distinguish more details in this last. Interestingly, for the same group of students, we obtained better results when the test's items contained a few calculation elements. Based on this preliminary observation, for investigating knowledge in electromagnetism, we proposed to use the BEMA test, which harmonizes simple calculations with conceptual knowledge questions.

3.2. Analyzing Knowledge in Electromagnetism by Using Shortened BEMA Test

In this example, we have used a shortened version of the BEMA test, which contains 20 items based on conceptual questions and elementary calculations. It has been drafted based on literature [24, 27, 28] by excluding some chapters from the original BEMA version and keeping similar subjects with EMCI used previously. Also, we reduced the number of alternatives to 5, from 10 in the original version discussed in Hansen & Stewart [27] and McColgan et al. [19]. From our point of view, BEMA construction mimics common exams in physics better. Like ECMI (also called CSEM in literature), it is considered not an essay exam, so we are not surprised to obtain low results. There are claims that results obtained for the original BEMA are lower than those obtained for the original EMCI, but since we are using simplified

and shortened versions of them, the absolute comparison of corresponding scores is not considered herein. Instead, we are focused on the students' perceptions and responses about their construction.

The initial goal of this test is to examine the efficiency of teaching and learning electromagnetism with the current structure and syllabus. This test is realized in 2022 and 2023, before and after the physics course. It consists of low-scale testing, targeting only the students of the branches mentioned at the beginning of this paragraph. The number of students who participated in the test each year and from each branch was small for objective reasons, so we mixed the data from the same category, assigning the composite set in the column 'Cumulative Number', and all subcategories by rows in Table 2. Juts from mentioning, after first evaluation of the result, like in EMCI, the results were not satisfactory but a little bit better than those obtained by the EMCI test. We observed that the category 'before course' did not fulfill the validity requirement for self-consistency and discriminatory index for none of the branches. Next, the category EMI branch, after the course, has its indexes close to the valid zone. By composing a mixed Chemistry and EMI "after course" group, we get indexes in the valid zone for all components (Table 2). Therefore, we qualified this category only as suitable for a general concept inventory conclusion. Regarding our target group, we obtained that the electromagnetism knowledge at the beginning of the course level was at a low level, 40% and 42% for each branch, respectively. It is likely to be a consequence of the online learning system (2020–2021), when students had their physics courses in high schools. After the course, it goes up to 52% and 55% for EMI and the Chemistry branch, respectively. The learning efficiency is evaluated by the knowledge's gain value:

$$g = \frac{\%ScoreAfterCourse - \%ScoreBeforeCourse}{100 - \%ScoreAfterCourse} \tag{10}$$

Table 2. BEMA tests indexes

Branch	Cumulative number	Difficulty index	50% to 50% Discrimination index	Reliability Index (Point Biserial)	Self-consistency index	Discrimination power (Ferguson delta)
Chemistry Before Course	73	0.58	0.55	0.40	0.62	0.81
Chemistry After the course	49	0.63	0.58	0.55	0.67	0.84
EMI Before Course	127	0.45	0.58	0.54	0.65	0.87
EMI After the course	127	0.55	0.69	0.65	0.68	0.89
Mixed EMI + Chemistry After the course	**200**	**0.63**	**0.66**	**0.59**	**0.71**	**0.91**
Valid values		≥ 0.3	≥0.3	≥0.2	≥0.7	≥0.9)

For the chemistry branch, we obtained $g_{Chemistry} = 33.3\%$ and for EMI $g_{EMI} = 20.1\%$. We relate the higher gain for chemistry students to the fact that they have laboratory activities in their physics program. Nevertheless, the level of knowledge at the end of the course has remained at a low level (52%–55%), which reveals the persistence of knowledge shortcomings inherited from online learning during 2020–2021. Again, because all indexes of the 'before courses" tests were in the validity range, we acknowledge those findings as characteristics for the group of students that were interviewed. More conclusions can be drawn subsequently, but according to the goal of this paper, we are not going into details hereto. However, we are highlighting one more time the filtering ability of the indexes' analysis to prevent subjective generalization of the present CI test findings. Next, we distinguish that the mixed group's data shown in the last row of Table 2 satisfies index validity prerequisites, providing statistical credentials for generalization, according to our proposal. Since the BEMA test is a certified exemplar, we believe that our simplification has not damaged this feature; therefore, the outcomes of our test mirror the trustworthiness of the students' understanding of the subjects of the electromagnetism included in it.

3.3. Discussing the Features of Item Difficulties on BEMA Test

The test's item difficulties are basic outcomes of the Concept Inventory analysis and of the Rasch model. They mirrored averaged students' responses for the teaching performance, structure of the syllabus, and their implantation in certain subjects and chapters. By nature, those parameters vary from one test to another, facilitating a comparative view of the weight of the factors mentioned herewith. In this framework, we are showing a few findings from this category for our working example. Firstly, we evidenced that calibrated difficulties differ slightly from their starting original values for all groups of categories 'after the course', indicating that the BEMA test has been understood correctly. In Figure 1, initial items' difficulties are denoted by marks, and final ones are shown by histograms. Those differences are found to be higher for the groups of "before courses". This affirms that the "face validity" issues inherited from the incomplete learning in some parts of the physics syllabus during high school studies have been recovered during university courses. Note that this statement is valid for our current students' group only because not all indexes of the test are in the valid range. Second, the small differences between the original and calibrated items' difficulties obtained for the test corresponding to the last row in Table 2 reaffirm the conclusion that the BEMA test is an acceptable instrument for the measurement of knowledge for our group of students. Knowing that the indexes for this case are in the valid range, this feature is admitted as general for the whole category of the students under investigation, that is, for

all of them who preferred to study in branches like those mentioned in this paragraph (and have their high school classes during 2020–2021).

Another conclusion emerges from the comparative analysis of the outcomes of BEMA tests conducted in different groups. We observed that items' difficulties measured on 'the'mixed group, after the course," whose indexes are very close to the validity range, differ significantly from those of the 'EMI after course' group, which is part of it, and have some of the indexes out of the valid range. From this fact, we noticed that even a mechanical mixing of CI test data can neutralize the local nature of the CI outcome. Regarding our main goal of the study in this working example, we reaffirm that, for analyzing students' perceived difficulties in electromagnetism and teaching efficiency related to this course, we must consider last row's data on Table 2. So, bearing in mind that in the Rasch model terminology, the negative values denote easier items and the positives signify harder ones, questions 1, 3, 5, 7, 9, and consequently the corresponding part of the course are classified harder locally, that is, for the EMI students. Those are shown by significantly positive brown histograms in Figure 1. On the other side, items 1, 5, and 9 in Figure 1 are likely to be harder for all students, which are presumably represented by the sample interviewed (e.g., students enrolled in branches where physics is a basic discipline). In this regard, we have identified a possible teaching shortcoming that might be related to the insufficient lessons on the corresponding subjects, etc. Regarding the very easy item number 2, we obtained that it corresponded to an outfit, so further analysis based on the Rash model is needed according to Planinic [32] and Zaiontz [30], etc. We observed next that the magnitude between extremal difficulties is higher for the mixed group, but it is likely indicating the heterogeneity injected by the mechanical mixing procedure. However, those conclusions, which were facilitated by harmonizing Rash analysis with indexes 'validity assistance, remain debatable if we consider the unmatching ability and difficulty interval. For the last row, the ability lay in the interval [-2.3, 2.3] and the items' difficulties in [-1.2, 0.8].

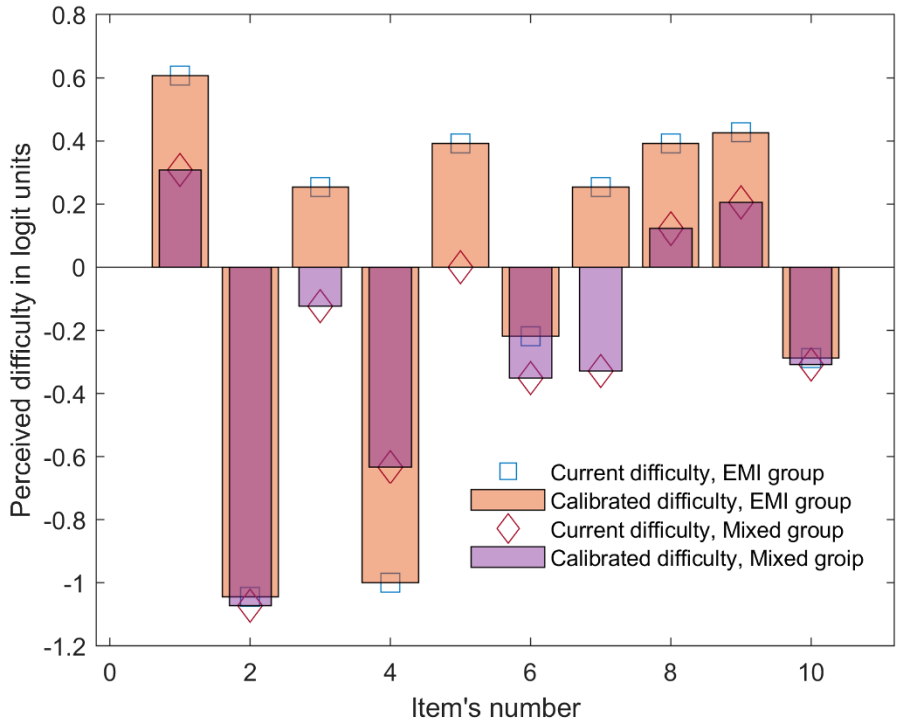


Figure 1. The perceived difficulties on the BEMA test

Based on Crooks and Alibali [1], Hestenes et al. [7], Planinic et al. [42], etc., for a perfect CI test, abilities' intervals must match difficulty intervals, so our pair {the sample, BEMA test} do not fulfill this requirement, despite the condition of the indexes' validity having been achieved. Therefore, current findings should be considered for a deeper analysis, and a fresh BEMA must be conducted for a thorough understanding of pedagogical and learning issues for the system under scrutiny. We are skipping them for now because those are not in the scope of this present work. Up here, we confirm that BEMA is a suitable instrument for measuring understanding and knowledge in electromagnetism for the students that follow studies in our university branches where physics is a basic discipline, but the most relevant remark consists in the effectiveness of employing indexes' theory to assist and advance the BEMA test analysis herein.

3.4. Identifying Levels of Conceptual Knowledge by Using the Distribution of the Abilities

In this part of our work, we will consider an ad-hoc auxiliary step for realizing a suitable partition of the abilities into levels. The idea is to improve the gradation or scoring system for deeper pedagogical analysis. As a Rasch model output,

abilities reflect the quality of the CI test, its compatibility with the students who participate in it, and the students understanding of the subjects of the tests. Each value on the ability array also provides an individual student's knowledge measure, according to Planinic et al. [42], where it is stated that Rasch analysis "is not just for instrument development but also for computing person measures". We may extend this idea by assuming that individual abilities belonging to a given interval represent an averaged entity, a certain level of knowledge. This new unit depends on the current test and therefore is a local feature relative to the group of respondents and the test used, but it is important from the pedagogical point of view. If the test used is certified, such as BEMA, and if the indexes of this test are in the validity range, then this clustering of the abilities would result in more general characteristics of the knowledge. Notice that this argument is not genuine. We use similar reasoning in grading practice. So, we assign the grade 7 if the scores obtained in the present exam are between 65% and 75% of the total. However, the difference stands in the method of this gradation. We will propose a more natural graduation compared with the fixed one. As an example, let's consider the abilities measured on the test corresponding to the last row of Table 2, which is qualified above according to our indexes filter. For comparative purposes, we will also consider initial abilities, which are calculated by Equation 7.

Again, the validity of the indexes for the last row implicates that even the raw values of these tests are credible and representative. The analysis in this paragraph is based on simple distribution arguments, including primary identification of the shape of the distribution (by operating the *ksdensity* function in MATLAB [47]) and final histogram optimization (*histcounts* and *histogram* functions). In those preliminary examinations, we observe that abilities measured for the BEMA test conducted on the mixed group seem to be drawn from an identifiable distribution function. We have concluded this finding by performing density function exploration based on the *epanechnikov* kernel, but because the number of points (200 records) is not large enough for quantitative analysis regarding distribution features, we are not going into details about this argument. This feature indicates that the group of interviewed students behaved smoothly toward the solution of the entire test. Therefore, we can use the term 'distribution' for students' capabilities to solve the BEMA test herein. Based on those initial arguments, it makes sense to group students that have "similar or close ability", or "level of knowledge" for this test, which coincides with our initial idea presented at the beginning of this paragraph. Also, from a didactical viewpoint, measured knowledge is supposed to exist at natural levels. Herein, we propose to employ an optimized histogram of abilities for identifying them. This proposal can contribute to another enhancement of CI analysis (Figure 2).

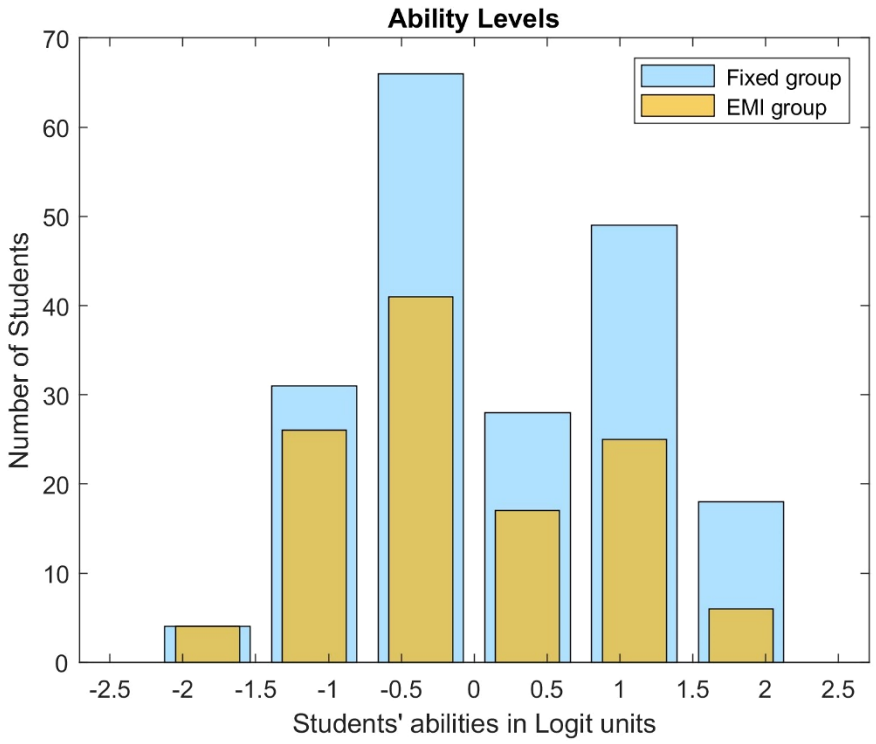


Figure 2. Observed ability for BEMA test. One observes six 'natural' levels of the abilities corresponding to six levels of knowledge. Abilities' levels have different populations

The idea is borrowed from basic distribution-fitting arguments [48–50]. During the preliminary kernel density exploration, we observed that the profile of the empiric distribution exhibits two undulating shapes, indicating that the distribution is not unique and hence not a stationary function. Therefore, the estimation of the bin size by standard Scott rule [49] which used mostly, $h(data) = \frac{3.49sqrt(data)}{n(data)^{-3}}$ is not applicable. Remember that optimal bin size searching refers

to the procedure of discretising (grouping) continuous data in $N_{optimal}$ bins to produce the appropriate histogram, [41-42]. So, we used the Freedman-Diaconis rules introduced in Popp et al. [45]. Note that this step can be performed automatically in the Matlab command 'histogram' by setting the 'FD' value or the'method' option, but it is important for our scope of work to clarify the ability classes idea, so the calculation formula for the optimal bin size is:

$$h_{optimal}(\beta) = 2 \times \frac{IQR(\beta)}{(n(\beta))^{\frac{1}{3}}} \tag{11}$$

Here, IQR is the interquartile of the data and $n(\beta)$ is the number of students. Please note that we have also examined the stationarity issue for the optimal distribution by employing an intermediate step based on the q-Gaussian distribution fit. This approach has been demonstrated to be an efficient method for measuring such features [44, 51]. The analysis unambiguously confirmed that the distribution is clearly non-stationary, which informs the choice of the bin size for the FD method. Now, the optimal number of classes, corresponding to the number of optimal histograms, is:

$$n_{classes}(\beta) = \frac{\max(\beta) - min(\beta)}{h_{optimal}(\beta)} \tag{12}$$

and the optimized classes of the observed ability are as follow:

$$\{[\beta_{min}, \beta_{min} + h_{optimal}], \beta_{min} + 2 * h_{optimal} + (n_{classes}(\beta) - 1)h_{optimal}, \beta_{max}\}. \tag{13}$$

The optimal histogram obtained consists of a better classification of the current abilities compared with fixed grades because it represents confidently the structure of the current students' knowledge. The unity of the ability in Equation 11 corresponds to the "natural units of grades" for this exam. Here we obtained that for the initial group of 127 participants belonging to EMI-after the physics course, which has best indexes except the mixed group, there are six ability levels centred on the values $\beta_{centered}^{classes,local} = [-1.8312 - 1.0987 - 0.3663\ 0.3662\ 1.0987\ 1.8312]$ and the classes largeness is 0.72 logits. The superscript 'local' indicates that this parameter characterizes strictly the group interviewed. It means that in this sample, the students having abilities closer than 0.36 logits should be considered always at the same level regarding their ability to solve the BEMA test. For the mixed group, which has best indexes and practically in the validity range, we obtained again six classes of the ability values as follow

$$\beta_{centered}^{classes,global} = [-1.8309 - 1.0985 - 0.3662\ 0.3662\ 1.0986\ 1.8309] \tag{14}$$

and the largeness of the classes is 0.631 logits. So, it follows that students whose abilities are closer than 0.315 logits belong always the same level of ability; students whose abilities differ less than 0.631 logits should be considered to have the same level of knowledge, etc. Notice that by convenience, all abilities whose difference is smaller than 'the natural unit' of 0.631 logits belong to the same level, but when we group them, we should agree on the starting and ending value for each level (histogram edges), so in a routine classification, students might be on the same natural level but in different ability histograms. Next, by considering the statistical credibility of the test corresponding to the last row in Table 2 and counting the fact that BEMA was qualified in the previous paragraph as a good instrument for measurement, we admit that the structure of the students' understanding of electromagnetism for our population under investigation is characterized by six levels, and students whose ability difference is less than 0.631 logits have the same level of knowledge. This statement holds for all students in the main category considered at the beginning of this paragraph.

4. Enhancing CI Analysis for Normal Exams by Using Likert Analysis for Dominant Errors

The assessment of a student's knowledge through routine exams is realized through a procedural knowledge test, which evaluates the student's capability to follow the pre-instructed solution method, calculation correctness, fluency in interpretation of the results, etc. Students might fail a common exam in physics due to conceptual deficiencies related to the physics subjects included in the question, due to the calculus difficulties, or because those shortcomings condition each other, resulting in an overlapped error appearance. In Prenga et al. [21], we have analyzed the nature of the errors in the mechanics exams in a certain medium and specific circumstances. The error states are identified by their dominance using the notation $[S_0, S_1, S_2, S_3]$ to ascribe each of the errors' occurrence in the exam respectively, say, $\{no - errors, calculation\ failures,\ conceptual\ failures,\ equalized/both\ errors\ presence\}$. Basically, our error space has three eigenstates, because the fourth one, '*equalized errors presence*' can be viewed as the mixture of the two first but bearing in mind the logic of the 'error dominance', one can acknowledge it as an eigenstate too. Their eigenvalues are given by the correspondence $[S_0, S_1, S_2, S_3] \leftrightarrow [0,1,2,3]$. This eigenvalue spectrum is not simply an ordinal or categorical assignment which are used in the standard Likert scale notation, however. Indeed, we may assume them as indicators of the weight or hardiness of the errors that they represent, which enables a little algebra too. So, the superposition of the states $S_1 + S_2$ has the eigenvalues 1+2=3, it exists the null element, etc. This property suggests adopting and extend Likert scale for analysing our system. In, Prenga et al. [21, 23], we have proposed to combine Likert scale with CI analysis to study the paring of the errors in classical mechanic's exams. Now, we want to advance these

analyses in a more general view, aiming to attain thorough conclusions for the causes of failures in a physics exam in physics, and for accrediting this technique as a fresh enhancement of CI itself. Notably, we advance herein with the Likert scale by adopting the quasi-numerical values for indicating error states and by employing test's reliability requirements for achieving a thoughtful error diagnosis. To clarify those elements, let's have a short insight on Likert scale analysis. It is based on a rating scale which used to measure opinions, attitudes, or behaviours. Basically, it provides a range of responses to a statement or attitude representing n-states, [46-49]. The most used values for the range are $n = 3, 5, 7, or\ 10$. A typical Likert scale assigns a categorical or ordinal value to each Likert state, for example, (1) for the full disagreement, (2) for neutral attitude, and (3) for the full agreement upon a given statement, but other conventions are employed too. Depending on the survey circumstances and specifics, the analysis of the data can be performed by using the frequencies of the Likert state occurrences or interval evaluations. In this work we have considered the analysis based on the interval evaluation, similar as in Nyutu et al. [52], because of the quasi-numerical nature of the Likert values. So, the attitude (L) of the N-members group is indicated by the average value $< L >= \frac{1}{N}\sum_{i=1}^{N} L_i$, hence, it is represented by a continuous variable. The Likert states L_j (j=1, 2…n) for the entire group are identified by intervals of the type $[j, j + \frac{n-1}{n}]$, whose average attitude $< L_i >$ belongs to. For an unambiguous assignment of the population's Likert states one consider self-report measures given by several testing indicators, [48*-50] etc. Due to the introductory nature of this work, we have skipped reporting about those elements.

A common approach for identifying the state of the population could be based on the reasonable rule that the edges values $L_{j\pm} \equiv < L_i > \pm\sigma < L_i >$ must not exceed an entire category. Regarding to the test self-constituency parameter which is important for the credibility of the Likert measurement, we have considered minimalist requirements for its estimator ($\alpha_{Cronbach} \geq 0.7$), [53], because the data used for these analyses have been gathered from existing homework and exams, hence, of a limited size by their nature. Fortunately, this condition has been met, so we were able to discuss conclusions quantitatively. Notice also that our system of four states/values is not a proper Likert scale, but according to Louangrath & Sutanapong [54], the non-Likert scales of the type [0, 1, 2, 3] (the number of even states is even) are more efficient for quantitative evaluation. Now, let's display our case study, which was purposed initially on the analysis of the pairing of errors in common physics exams and on evidencing influential factors related to them. For answering this question, we have proposed the combination of the Likert scale with CI analysis, which is more important for our view of research interest than its concrete results. So, let's try to identify which type of the above basic errors is paired with one of the basic misconceptions in mechanics: {*kinematics confusions, impetus, active force, action/reaction pairs, concatenation of influences, other influences in motion}* [7]. The raw data used for this research consisted of real midterm exams and homework, which were accomplished by students of the above-mentioned branches during the period 2020–2022. Notice that some items of those tests have been drafted intentionally to expose them to one of the six common senses in mechanics.

Table 3. The Likert-like values corresponding to the four exams' error states

State's notation	Error Type by the dominance	Individual Likert value	Interval values for L_{group}
S_0	No errors	0	0-0.75
S_1	Calculation issues dominate	1	0.76*-1.50
S_2	Conceptual errors dominate	2	1.51-2.25
S_3	Conceptual Errors and Calculations are alike present	3	2.26-3.00

Initially, 102 exemplars from this set have been selected for analysis. Two professors were asked to investigate them independently and to assess the errors on each test by the quasi-Likert scale assignment according to Table 3. After this procedure, 91 records whose both assessments had matched were considered for further analysis. This procedure is performed initially to guarantee the purity of the data, but it also works to some extent as the test-retest evaluation step in standard Likert analysis. We observed that our self-consistency requirement has been fulfilled. We are aware that our test contains problems regarding the six commonnesses' presence. For example, when solving the problem "find the distance between the train with M=100 T and its wagon with m=10T, 3 sec after the separation, if the force of the string was 1000 N", the phrasing of which was intended to provoke an Action/Reaction common-sense confusion, students might have faced the additional difficulty of the concept of the referential, which belongs to the kinematic common-sense confusion. Therefore, the outcomes of the analysis would also reflect the test's clarity, but it does not limit the validity of the method that we are implementing. Rather, it suggests an interpretation of the findings according to this insight. After getting the data ready for use, we have performed a Rasch calculation for the quasi-polytomous variable, evidencing misfit occurrences. It resulted that the problems that were exposed to the "*concatenation of influences"* confusions were outfitted, and the dominant error as per independent professors' estimations for those problems has been "both errors present". A second examination of this finding, which is a routine step suggested in Louangrath and Sutanapong [54], has provided that conceptual and calculus errors have been mixed mostly because of the formulation of the problems. Therefore, we have excluded this category from the test, reducing it to five items in Table 4. For the remaining set, we calculated the average (or group) Likert values and corresponding standard errors. We argued that if standard deviation of the average values is larger than the width (w) of the category ($\sigma(L) > w \equiv 0.75$), the assignment

for the Likert group' is fuzzy. Unambiguous conclusions would require that $\sigma(L) < \frac{w}{2} \equiv 0.325$ which guarantee that $L_j \pm \sigma(L_j)$ do not exceed the entire neighbour classes for any values of the L_j.

Table 4. The Likert-like measure of typical

Common-sense Confusions / Observables	Kinematics Confusions	Impetus	Active Force	Action/Reaction pairs	Other Influences in Motion
Likert Value $< L >$	0.86	1.47	2.03	1.18	2.44
Standard Deviation σ	0.22	0.31	0.63	0.43	0.35
Abbreviation of the Confusions	K	I	A-F	A-R	IM

In the first column of Table 4 which belongs to the *cinematic confusion common-sense error*, the Likert value is $L_K = 0.86 \in [0.75,1.50]$, so according to Table 3, the calculating errors (type S_1 *in Table* 2) dominate. Here, $\sigma(L_K) = 0.22 < 0.375$ so this statement is conclusive and convincing. It resulted that the *calculation error* ($type\ S_1$) dominates the failure on solving problems in which the *action-reaction common-sense* is present, as seen from the Likert value $L_{A-F} = 1.18 \in [0.75,1.50]$. The standard deviation $\sigma(L_{A-R}) = 0.43$, is higher than the lower limit 0.375, but since it does not exceed the upper limit of 0.75, this finding is considered mostly valid. To avoid ambiguities, more records are needed. Next, for problems where the most probable common-sense confusion is *active force*, we see that $L_{A-F} \in [1.51 \div 2.25]$ and based on Table 3, conceptual errors are more frequent. Note that $\sigma(L_{A-R}) = 0.63$ is considerable high, but remain below rejectable limit 0.75, so this conclusion is again valid, but it should be taken with little precaution. Probably both errors are present. Regarding the lower limit of the Likert value in this vase (2.03-0.63=1.4), we may first suppose that calculation shortcomings are dominant in several cases. However, if conceptual failures dominate according to the centered value, students are confused on the calculation step because of them, which favors the selection of conceptual confusions as the most dominant issue. Note Prenga et al. [21] found that conceptual knowledge shortcomings prevailed in calculation failures on our students' physics exams, which supports the reasoning above too. Finally, for the problems in mechanics where o*ther influence in motion* common-sense are most likely we observe that Likert value correspond to the error state '*both errors are present*', $L_{IM} = 2.44 \in 2.26 \div 3.00$, and $\sigma(L_{IM}) = 0.35$. Therefore, for this case we conclude that our students have confused their concepts and have obstructed the calculations at the same time or interchangeably

It is interesting to compare the results of successive assessments by using this method, too. We do not realize the need for double tests, but for elaborating on the idea, we have used the old test conducted at the beginning of the academic year referred to in Prenga et al. [21]. The data used in this study were gathered from students' activities during their academic year. Evidently, those examinations have been conducted in different groups, but qualitatively, we can use them for comparative purposes. Considering the remarks of this research, if the indexes of both tests would have been in the validity range, the analysis is considered legitimate. Despite this not being the case herein, we have shown the result of these proposals below (Figure 3).

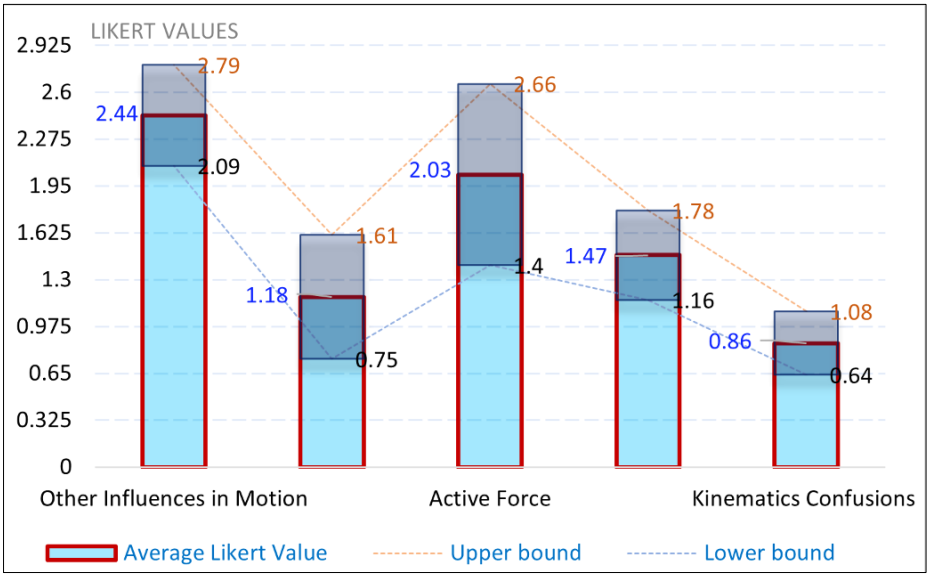


Figure 3. Diagram presentation of errors' Likert states. Shadowed bars represent the interval of the Likert value corresponding to the error states. It is assumed that groups' Likert states have their proper values at the midpoint of intervals given in Table 3. Problems with kinematic confusions or Action/Reaction pair issues seems to be dominated by a single error type.

So, we observe that for problems suffering from the kinematics confusions in Prenga et al. [21] $L_K \in 2.27 \pm 0.7$ and for the corresponding ones in the current test it is $L_K \in 0.86 \pm 0.22$. By assuming the test in Prenga et al. [21] as *"before the course"* measurement and the findings presented in Table 3 as respective *"after the course*" measurement of knowledge, we can read the improvement in conceptual knowledge for kinematics concepts after physics course, because the error type ascendency has shifted from '*both error presence'* to the '*conceptual error*" dominance for the selected category of problems. Note that there are a couple of assumptions in this last illustrating discussion, but we believe that this method can be productive in this approach. In another tentative analysis we use this method for exanimating errors in electromagnetism exams. The error states where chosen: $\{S_1 \rightarrow$ *calculus failure*; $S_2 \rightarrow$ *mechanics conceptual issues*, $S_3 \rightarrow$ *electromagnetism conceptual issues*, $S_4 \rightarrow$ *both conceptual failures are present}* and $S_0 \rightarrow$ *none of them are present, and* corresponding eigenvalues were assigned [0, 1, 2, 3, and 4]. The data have been selected from exams-papers realized by 86 students on the branches considered above, in academic year 2021-2022-2023. This analysis aimed a qualitative view, so we are skipping the rigorous treatment. We obtained $L = 3.23 \in [3.2 \div 4]$, and $\sigma(L) = 1.16$. According to the discussion in this paragraph, those findings suggested that the failures on electromagnetism exams were probably caused by conceptual failures in mechanics and electromagnetism. Interestingly, calculus issues were not found to be responsible for students' failures on exams analyzed. Despite the apparently enthusiastic findings herein, we are aware of the limitations of this approach, but if adopted carefully, it can be used for a thorough analysis of the difficulties that students faced in solving problems in electromagnetism. Also notice that the use of the mean and its deviance for identification of the Likert state for the group merits more attention and requires more supportive arguments. All those issues will be addressed in the forthcoming work. However, we believe that this procedure can be used successfully in various similar circumstances to those analyzed in this section.

5. Conclusion

Combining the Rasch techniques with indexes' analysis, Likert scale measurement, and histogram optimization presented in this work has demonstrated effectiveness in advancing CI analysis and improving physics knowledge's measurements. The test's index validity is initially used as a preliminary stage for filtering subjective conclusions and unjustified generalizations of Concept Inventory test outcomes. In this regard, for resolving indexes' validity disputes if retesting was costly, unpractical, or not realizable, adding a few records to the existing CI test data, conducting a partial or compensatory test, and mixing the data from similar tests have proven fruitful in our working examples. By using this approach, we determined that the online learning imposed during the pandemic closure has impaired the physics conceptual knowledge of high school students. The level of knowledge in mechanics for this period is evaluated at around 48%, which lies below the basic knowledge in Newtonian mechanics. Also, the conceptual knowledge gained after the physics courses at the university for students who preferred the branches where physics is a basic discipline is estimated at around 22%–33%.

It resulted in tests that are drafted, like procedural knowledge tests, being better instruments for measuring understanding in electromagnetism for the category of students considered. We believe that this feature mirrors limited capabilities for instructing high school students with sufficient conceptual knowledge due to the limitations of online teaching. Next, a better assessment of the knowledge levels is achieved by using histogram optimization borrowed from distribution fitting practices. In this case, the optimal bin size provides the natural ability's unit for the current test. Finally, by using the Likert scales idea to assign error states, we investigated the pairing occurrences between the dominant basic error and typical common sense in mechanics. For example, we found that the burden of conceptual errors when solving problems in which kinematics confusion is present has been reduced after the physics course at the university. For problems in which the common-sense error of the type of *other influences in motions* is present, calculus failures and conceptual shortcoming persisted at the same weight even after the physics course, etc. We concluded that the procedures proposed in this study are neither system-dependent nor limited to the physics CI test. They can be used as the role of auxiliary tool for enhancing concept inventory assessment in general.

6. Declarations

6.1. Author Contributions

Conceptualization, D.P. and E.K.; methodology, D.P.; software, D.P.; validation, E.K., D.P., and F.M.; formal analysis, D.P.; investigation, F.M.; resources, F.M.; data curation, D.P. and E.K.; writing—original draft preparation, D.P.; writing—review and editing, D.P.; visualization, E.K.; supervision, D.P.; project administration, D.P. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

No new data were created or analyzed in this study. Data sharing is not applicable to this article.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Declaration of Competing Interest

The authors declare that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7. References

[1] Crooks, N. M., & Alibali, M. W. (2014). Defining and measuring conceptual knowledge in mathematics. Developmental Review, 34(4), 344–377. doi:10.1016/j.dr.2014.10.001.

[2] Savinainen, A., & Scott, P. (2002). The force concept inventory: A tool for monitoring student learning. Physics Education, 37(1), 45–52. doi:10.1088/0031-9120/37/1/306.

[3] Rahmawati, Rustaman, N. Y., 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), 483–490. doi:10.15294/jpii.v7i4.13490.

[4] Savinainen, A., & Viiri, J. (2008). The force concept inventory as a measure of student's conceptual coherence. International Journal of Science and Mathematics Education, 6(4), 719–740. doi:10.1007/s10763-007-9103-x.

[5] O'Shea, A., Breen, S., & Jaworski, B. (2016). The Development of a Function Concept Inventory. International Journal of Research in Undergraduate Mathematics Education, 2(3), 279–296. doi:10.1007/s40753-016-0030-5.

[6] Luangrath, P., Pettersson, S., & Benckert, S. (2011). 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 and Technology Education, 7(2), 103–114. doi:10.12973/ejmste/75184.

[7] Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. The Physics Teacher, 30(3), 141–158. doi:10.1119/1.2343497.

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

[9] Khairandy, R., Barkatullah, A. H., Huda, M. K., & Amir, A. Y. (2022). Exploring Social Contracts: Enhancing Cooperation and Collaboration between Businesses and Communities. Journal of Human, Earth, and Future, 3(4), 452-460. doi:10.28991/HEF-2022-03-04-005.

[10] Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory. Springer, Dordrecht, Netherlands. doi:10.1007/978-94-017-1988-9.

[11] Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press, London, United Kingdom. doi:10.4324/9781410605269.

[12] van der Linden, W. J., & Hambleton, R. K. (1997). Handbook of Modern Item Response Theory. Springer, New York, United States. doi:10.1007/978-1-4757-2691-6.

[13] Rasch, G. (1961, January). On general laws and the meaning of measurement in psychology. Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, 1 January, 1961, Berkeley, United States.

[14] Klymkowsky, M. W., & Garvin-Doxas, K. (2020). Concept Inventories: Design, Application, Uses, Limitations, and Next Steps. Active Learning in College Science. Springer, Cham, Switzerland. doi:10.1007/978-3-030-33600-4_48.

[15] McCombes, S. (2023). Sampling Methods | Types, Techniques & Examples. Scribbr, Amsterdam, Netherlands. Available online: https://www.scribbr.com/methodology/sampling-methods/ (accessed on January 2023).

[16] 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. doi:10.20961/jphystheor-appl.v1i1.4706.

[17] Smaill, C., & Rowe, G. (2012). Electromagnetics Misconceptions: How Common Are These Amongst First- and Second-year Electrical Engineering Students? 2012 ASEE Annual Conference & Exposition Proceedings. doi:10.18260/1-2--21268.

[18] Raduta, C. (2005). General students' misconceptions related to Electricity and Magnetism. arXiv preprint, physics/0503132. doi:10.48550/arXiv.physics/0503132.

[19] McColgan, M. W., Finn, R. A., Broder, D. L., & Hassel, G. E. (2017). Assessing students' conceptual knowledge of electricity and magnetism. Physical Review Physics Education Research, 13(2), 20121. doi:10.1103/PhysRevPhysEducRes.13.020121.

[20] Laverty, J. T., & Caballero, M. D. (2018). Analysis of the most common concept inventories in physics: What are we assessing? Physical Review Physics Education Research, 14(1), 10123. doi:10.1103/PhysRevPhysEducRes.14.010123.

[21] Prenga, D., Kushta, E., Peqini, K., Osmani, R., & Hysenlli, M. (2023). Analyzing influential factors on physics knowledge weakness in high school students due to the pandemic-imposed online learning and a discussion for enhancing strategies. AIP Conference Proceedings, vol. 2872, no. 1. doi:10.1063/5.0162933.

[22] Kushta, E., Dode Prenga, S. M., & 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. doi:10.17762/msea.v71i4.1228.

[23] Pattanasing, K., Aujirapongpan, S., Dowpiset, K., Chanthawong, A., Jiraphanumes, K., & Hareebin, Y. (2022). Dynamic Knowledge Management Capabilities: An Approach to High-Performance Organization. HighTech and Innovation Journal, 3(3), 243-251. doi:10.28991/HIJ-2022-03-03-01.

[24] 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), 10105. doi:10.1103/PhysRevSTPER.2.010105.

[25] Aubrecht, G. J., & Aubrecht, J. D. (1983). Constructing objective tests. American Journal of Physics, 51(7), 613–620. doi:10.1119/1.13186.

[26] Notaros, B. M. (2002). Concept inventory assessment instruments for electromagnetics education. IEEE Antennas and Propagation Society International Symposium (IEEE Cat. No.02CH37313). doi:10.1109/aps.2002.1016436.

[27] Hansen, J., & Stewart, J. (2021). Multidimensional item response theory and the Brief Electricity and Magnetism Assessment. Physical Review Physics Education Research, 17(2), 20139. doi:10.1103/PhysRevPhysEducRes.17.020139.

[28] Kožuchová, M., Barnová, S., & Stebila, J. (2022). Inquiry as a part of educational reality in technical education. Emerging Science Journal, 6 (Special issue), 225-240. doi:10.28991/ESJ-2022-SIED-016.

[29] Linacre, J. M. (2020). Fit diagnosis: Infit outfit mean-square standardized, Winsteps. Available online: https://www.winsteps.com/winman/misfitdiagnosis.htm (accessed on February 2023).

[30] Zaiontz, C. (2023). Building a Rasch Model. Real Statistics Using Excel. Available online: https://real-statistics.com/reliability/item-response-theory/building-rasch-model/ (accessed on February 2023).

[31] 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. doi:10.3102/1076998609353117.

[32] 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. doi:10.1119/1.2366733.

[33] 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). doi:10.1103/PhysRevPhysEducRes.15.020111.

[34] Bevans, R. (2022, December 05). Choosing the Right Statistical Test | Types & Examples. Scribbr, Amsterdam, Netherlands. Available online: https://www.scribbr.com/statistics/statistical-tests/ (accessed on February 2023).

[35] Bruning, J. L., & Kintz, B. L. (1987). Computational handbook of statistics (3rd Ed.). Foresman and Company, Northbrook, United States.

[36] Martínez-Mesa, J., González-Chica, D. A., Duquia, R. P., Bonamigo, R. R., & Bastos, J. L. (2016). Sampling: How to select participants in my research study? Anais Brasileiros de Dermatologia, 91(3), 326–330. doi:10.1590/abd1806-4841.20165254.

[37] Chabay, R. (1997). Qualitative Understanding and Retention. AAPT Announcer, 27(2), 96.

[38] Liu, X. (2010). Using and developing measurement instruments in science education: A Rasch modeling approach. Information Age Pub, Charlotte, United States.

[39] 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. doi:10.1119/1.2117109

[40] Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? CBE Life Sciences Education, 15(4). doi:10.1187/cbe.16-04-0148.

[41] Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 140, 44–53.

[42] 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), 1-11. doi:10.1103/physrevstper.6.010103.

[43] Granger, C. (2008). Rasch analysis is important to understand and use for measurement. Rasch Measurement Transactions, 21(3), 1122-1123. Available online: https://www.rasch.org/rmt/rmt213d.htm (accessed on January 2023).

[44] Umarov, S., Tsallis, C., & Steinberg, S. (2008). On a q-central limit theorem consistent with nonextensive statistical mechanics. Milan Journal of Mathematics, 76(1), 307–328. doi:10.1007/s00032-008-0087-y.

[45] Popp, S. E. O., & Jackson, J. C. (2009). Can assessment of student conceptions of force be enhanced through linguistic simplification? A Rasch model common person equating of the FCI and the SFCI. Annual Meeting of the American Educational Research Association, April, 2009, San Diego, United States.

[46] 39-Stoen, S. M., McDaniel, M. A., Frey, R. F., Hynes, K. M., & Cahill, M. J. (2020). Force concept inventory: More than just conceptual understanding. Physical Review Physics Education Research, 16(1), 10105. doi:10.1103/PhysRevPhysEducRes.16.010105.

[47] MathWorks (2023). MATLAB Online. Available online: https://www.mathworks.com/products/matlab-online.html (accessed on February 2023).

[48] Knuth, K. H. (2019). Optimal data-based binning for histograms and histogram-based probability density models. Digital Signal Processing, 95, 102581. doi:10.1016/j.dsp.2019.102581.

[49] Scott, D. W. (1979). On optimal and data-based histograms. Biometrika, 66(3), 605–610. doi:10.2307/2335182.

[50] Scott, D. W. (2015). Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons, Hoboken, United States. doi:10.1002/9781118575574.

[51] Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L2 theory. Zeitschrift Für Wahrscheinlichkeitstheorie Und Verwandte Gebiete, 57(4), 453–476. doi:10.1007/BF01025868.

[52] 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, 9(1), 83–102. doi:10.46328/ijemst.1182.

[53] Gliem, J. a, & Gliem, R. R. (2003). Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales. 2003 Midwest Research to Practice Conference in Adult, Continuing, and Community Education, 1992, 82–88. doi:10.1109/PROC.1975.9792.

[54] 52-Louangrath, P. I., & Sutanapong, C. (2018). Validity and reliability of survey scales. International Journal of Research & Methodology in Social Science, 4(3), 99-114.