ISSN 2522-9400 **European Modern Studies Journal** *Vol 8 No 6*
available at www.journal-ems.com

DOI: 10.59573/emsj.8(6).2024.6

A Comparative Analysis of Mathematics' CI Testing on High School and Engineering Bachelor's Students in Albania

Sandër Kovaçi[1]*, Dodë Prenga[2]
[1]Department of Mathematical Engineering, Faculty of Mathematics and Physics Engineering, Polytechnic University of Tirana, Albania
[2]Department of Physics, Faculty of Natural Sciences, University of Tirana, Albania

Abstract. The use of Concept Inventory tests for analysing students' achievement and performance on understanding sciences has progressed last years in our academic and didactic mediums. Interesting and significant findings through those analysis have revealed the importance of deeper CI analysis and improvement of knowledge measuring instruments on our education communities. Based on this idea in this work we have discussed perceived features and characteristics of general Calculus concept inventory, and their significance regarding understanding physics and learning progress on this discipline. By measuring and analysing standard CI test indexes we can diagnose issues and shortcomings in conceptual understanding in physics. We evidenced an undesirable level of physics understanding not only in raw scores of the conceptual inventories, but also in the ability of students to correctly understand the physical context of problem. First evidence regarding these issues settles that students inherit shortcomings on conceptual knowledge from their undergraduate studies, and the gain on this indicator are relative to the initial conceptual knowledge. Consequently, we conclude that much effort should be made on improving conceptual knowledge during high school studies. To explore the cause factors of this indication we have proposed a factorial analysis, and preliminary findings support this assessment.
Keywords: tests indexes, concept inventory, mathematics knowledge, the Rasch model

Introduction

Conceptual and procedural knowledge tests are used routinely for assessment of student's acquaintance in science and for analysing and investigation education issues and features. Initially, CI tests have been developed and organised for studying physics' conceptual knowledge, see (Hestenes, 1992), but later, those interesting and important techniques have been employed successfully on student knowledge in natural science and Calculus. The editing procedure and organising elements regarding CI tests in physics and mathematics can be found in (O' Shea et al, 2016), (Carlson et al, 2010), (Crooks et al, 2014), etc., so we are skipping details regarding them. We will consider some specific aspects throughout this work, however. Basically, the test and exams in calculus and mathematical subjects used in routine academic activities are procedural. Procedural tests are focused on evaluating students' the ability to solve problems step by step, his (her) fluency to employ instructed methods, and this is practically important for professors of mathematics that are interested on the evaluation of practical skills of their students in resolving algorithmic problems in mathematical subjects, (Prenga et al, 2024). On the other side, conceptual test aims on measuring student's knowledge of fundamental relationships between variables and features of the physical system. Relative to those specifications, those tests are used for different commitments, and specifically, the CI tests are used mostly for research and investigative purposes. In this regard, several CI in physics have been organised for measuring students' knowledge in physics in Albania through the works of (Prenga et al, 2022, Prenga et al, 2023), and several CI tests in mathematics

* Corresponding Author

subjects have been employed too, but not as intensively as those in physics. However, ani increasing interest among our education researchers has been reported last time as shown (Denaj et al, 2024), (Kushta et al, 2022). Notice also that general test like Cangorous and others large scale exams are very traditional in the country, and overcome testing is physics for the country. Based on those elements we considered a CI test for analysing teaching and understanding mathematics on high school students by a lo scale pilot measurement. Similar use of this test has been used as Tool for Assessing Students' Reasoning Abilities and Understandings in general (Carlson et al, 2010), whereas similar test for mathematic subject have been developed and discussed straightforwardly (Epstein, 2013). The calculus Concept Inventory that we proposed to use herein as our instrument in this study has been developed by J Epstein (www.PhysPort.org) and provides a standard instrument for measuring knowledge in calculus and Mathematical clear reasoning. Notice that despite similarities of the CI test, the measurement of mathematical knowledge also shows serval dissimilarities, and specifics that should be addressed carefully for fruitful results. In this regard, we employed a comparative view with other CI tests conducted before and reported in (Prenga et al, 2022)], bearing in mind that mathematical knowledge plays the role of a broader knowledge measure, a feature that could be considered for Assessing Students' Reasoning Abilities and Understandings in general. In this regard, e Precalculus Concept Assessment (PCA) analysed in (Carlson et al, 2010) has been demonstrated to be useful for assessing pre-post learning and for comparing various approaches to teaching precalculus and college algebra. Therefore, we have limited our study on analysing a PCA version downloaded with permission by (www.PhisPort.org). This test consists of multiple-choice items, and therefore, the choice of the alternative is also a decision-making process, that displace arguments to item response theory (IRT) described by Embretson et al (2013), Wim J. van der Linden (2005), Van der Linden et al (1997). The analysis of a CI test can be performed through using Rasch analysis whose outputs consist of calibrated measure of student's ability, items difficulties and misfitting events mirror statistical inadequacies in the first insight, but represent guessing behaviour, confusing events, contextual and readability issues of the test, etc., that cannot be measured directly. Additionally, the overall behaviour of testing process and adequacy of the measurement can be checked be the analysis of test indexes, introduced in (Ding et al, 2006). We propose to extend the results obtained by the pilot test toward a more representative metrics, by using indexes as instruments for measuring representativeness of the students' competencies to understand the CI test, providing that participants mirror devotedly enough the whole community. By using indexes of a standard CI test such as PCA., we argue that every discrepancy observed regarding indexes ranges is due to the insufficient student's knowledge for the subject, which is being checked by the item, that enables us to acknowledge typical knowledge shortcomings indirectly. For measurement and analysis included in this work, we have conducted rigorous CI test in several classes at Faculty of Natural Sciences, University of Tirana, and Polytechnic University of Tirana some more lessened protocol exams, in other studies groups, used also we used informal conclusion from concreted surveys and questionnaires on high schools' professors and students. This last has been considered only for proposal drafting. It is worth underlining that elaboration of the data has been made by using our ad-hoc algorithms written in MATLAB, but interested readers can consider dedicated commercial software like (https://www.winsteps.com) or statistics guides (https://real-statistics.com).

Introductory elements of Rasch analysis

By nature, the CI test assessment is based on binary grades (0,1) corresponding to the correct/incorrect answers (Anderson, 2010), (Planinic, 2010). CI tests are organized on multiple choice alternatives, so, the student has nothing to calculate, and he(r) should be focused on conceptual knowledge only. After gathering raw data from the CI tests' outcomes, individual answers are shown by matrix elements $T(i,j) = (0,1)$. Unanswered questions are left blank. By a naïve approach, those values can be read as probability that student (i) could solve the test. However, when picking up an alternative, student might have guessed, or he failed to check right alternative despite he was strongly confident for it. So, common multiple choices test suffers from guessing or accidental failures. Another issue related to knowledge measurement is related to the scale. So, the knowledge differences between individuals (i) and (j) who won 90/100 and 70/100 scores respectively and (m) and (n) which won 30/100 and 10/100 respectively are obviously different, but numerically it appears the same. To deal with such problems, Danish mathematician D. Rasch has developed an elegant sociometric model based on dichotomous response variables which is known under his name, and later Andrich has advanced it for polytomous variables or Likert scale responses. Detailed arguments can be found in (Wright, 1997), (Coletta et al, 2019), (Planinic, 2016) etc. Here are in short, the core calculation steps. One start by calculating the success rate of student (i) and the fraction of students which answered correct for question (j) $P(i;j) = \frac{1}{N;n}\sum_{j;i=1}^{N;n} T(i,j)$). Those values as assumed as probability measures. Based on them one calculates their logistic counterparts known as logistics ability (β_i) and difficulty (δ_j) respectively.

$$\beta_i = ln\frac{\mathrm{P}(i)}{1-\mathrm{P}};\ \delta_j = ln\ \frac{1-\mathrm{P(j)}}{\mathrm{P(j)}} \tag{1}$$

Those quantities which are in logit units consist of linear and homogeneous measure of corresponding variables, see (Planinic, 2016) for full details. Now, the probability that student (i) would have solve item (j) is calculated straightforwardly based on IRT approach and homogeneity assumptions

$$P_e(i,j) \equiv P(\beta,\delta) = \frac{exp\ (\beta_i-\delta_j)}{1+exp\ (\beta_i-\delta_j)} \tag{2}$$

According to Rasch, under some assumption regarding distribution of the answers and abilities of students, for given difficulty of an item, the probability is unidimensional and depend solely on student ability (Planinic, 2016), (Prenga, 2024). Therefore, estimated value $P_e(i,j)$ is a better estimator of real probability for success compared to the initial (and naively evaluated) $T(i,j)$. For improving this estimation, the model proposes to replace iteratively matrices P_e^{n-1} with new ones P_e^nstarting from T. The procedure stops when a threshold criterion is met. First outputs of the model are final values

$$\beta_i^{final} = ln\frac{\mathrm{p}^{\mathrm{final}}(i)}{1-\mathrm{p}^{\mathrm{final}}(i)};\ \delta_j = ln\ \frac{1-p^{final}(\mathrm{j})}{\mathrm{p}^{\mathrm{final}}(\mathrm{j})} \tag{3}$$

By this idea, one concluded in a calibrated instrument for measuring knowledge, because the final outcome counts for all inhomogeneity issues and is capable of evidencing them progressively. Also, this measurement is more linear compared to the raw score. So, 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 $\beta_k = 1\ logit\ \ \beta_k = 2\ logit.$ The differences between final and original values contain interesting information and can be viewed as another estimator of the CI testing. Elevated deviances are called misfits and indicate the discrepancies between current behaviour of the system and preliminary assumptions for statistical approaches. Misfits' elements and items are classified as outlier-sensitive fit (outfit) and information-weighted sensitive (infit). A thorough description of those important elements can be found in (Anderson et al, 2010), (Planinic, 2010), (Liu, 2020), etc. Alos the reader can find very useful instruction in a detailed explanation provided in (https://www.rasch.org). for Seen as secondary indicators, infits and outfits could be used as averaged signals for "pathological individual" cases, that is to significant differences between original $T(i,j)$ and final $P_e(ij).So$, if $\beta_i < \delta_j$, $P_e(ij) < \frac{1}{2}$, and if $T(i,j)) = 1$, probably the check of the correct answer is made by accident, or by guessing. Also, $\beta_i > \delta_j$ and

$T(i,j)) = 0$, it is probably that student has failed by accident, or the item contain face or content validity issues regarding this student. Similarly, a discussion can be formulated for entire item misfitted issues. In this regard, the Rach analyse is more than a powerful measurement tool, it consists in a fruitful methodology of deeper studies in education issues, academic progress, teaching performance, and more. This feature can be easy evidenced by a brief overview of the application literature, some of them included in the references on this work. Interesting observation regarding those procedures employed in FCI studies for Albania, can be found in (Prenga, 2024), (Prenga et al, 2023).

Indexes of the Concept Inventory test

The indexes of tests are considered as a very efficient tools for evaluating the tests' goodness and testing environment itself. An instructive representation for this instrument is introduced by (Ding et al, 206). The calculating quantities for tests' indexes are estimators of integrity, significance, validity, discriminatory power, difficulty measure etc. By evaluating those indicators, researchers can judge about the significance of their results, but also, the quality of the tests, the presence of face or content validity etc, but one can use also the observed discrepancies between current values and desired ones as indirect indicators of students' knowledge and their understanding on the subjects being tests by a certain item (Prenga, 2024). To enable this capability, in this reference is proposed to use certified and standardised CI test, providing that that all subjects included in the CI test have been developed correctly and sufficiently during the school course. The quantities included in the set of the indexes are the following:

1. The item difficulty index and the test difficulty as their average, are given by

$$p_{item} = \frac{N_{CorrectAnswers}}{N_{students}};\ P_{test} = \frac{1}{N_{items}} \sum^{n_{items}} P_{item} \qquad (4)$$

By theory, desirable values for the parameter are in the range [0.3-0.9].

2. The *item discrimination index* measures the capability of the item to distinguish differences between students' knowledge. The average index reports discrimination capability of the test. It reads

$$d = \mathrm{f}\frac{(N_h^{x\%} - N_l^{x\%})}{N_{students}};\ D = \frac{1}{n_{items}} \sum^{items} d_{item} \qquad (5)$$

Here N_h is the number of students on the top-score list up to x%, and $N_l$is the number of the students with lowest scores up to x%. For $x = 50\%$ $f = 2$, and for $x = 25\%, f = 4$. Discriminatory indexes take their values in the interval [-1,1], and desirable values are $(d, D) >$ 0.3.

3. The *item's reliability index* measures the correlation of results for a given item with the whole test that is, its consistency. It is given by formulas

$$r_{item} = \frac{\tilde{x_i} - \tilde{x}}{\sigma(x)} \sqrt{\frac{p}{1-p}};r_{avg} = \frac{1}{n_{items}} \sum_{i=1}^{n_{items}} r_{item} \qquad (6)$$

$\widetilde{x_1}$ represent the average score for students who answered correctly, $\tilde{x}$ is the average of total score, p is the difficulty index of current and $\sigma(x)$ is the standard deviation of the total score. Desirable values are in the range $r > 0.2$

4. The *self-consistency of the test* (Kuder-Richardson index) is

$$\rho_{test} = \frac{n_{items}}{n_{items}-1} (1 - \frac{\sum^{n_{items}} \sigma(x_i)}{\sigma^2(x)}) \qquad (7)$$

$\sigma(x_i)$ and σ(x) are respectively the standard deviations of scores (x) for the item (i) and for the entire test. The values $\rho_{test} > 0.7$ are acceptable for the in-group measurement.

5. The *discriminatory power of the entire test* (Ferguson's delta) estimates the distribution of the scores over the possible range. It is given by the formula

$$\Delta = \frac{N^2 - \sum_i^{n_{items}} f_i^2}{N^2 - \frac{N^2}{n_{items}+1}} \quad (8)$$

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

Data and Methodology

From the pedagogical and academical point of view, the analysis of conceptual knowledge in mathematics, calculus and related subject in Albania is considered a new field of research. On the other side, studies and test for the overall knowledge in mathematics and science of nature are very common, and of remarkable quality. Also, periodical test on mathematics in the framework of traditional organisation and implementation of the new methodologies and curricula, are quite a normal activity among teaching and education agencies and activists. It resulted that we have very clear pattern of students' knowledge in mathematics, physics, calculus etc, but not lively evidence on the deep corresponding conceptual knowledge. This remark become more actual when considering remarkable changes in education practices and literatures for the high school studies in the country. Also, within our best knowledges, while several works have been elaborated last year regarding physics conceptual knowledge, see (Prenga et al, 2022), (Prenga, 2024), (Kushta et al, 2023), (Denaj et al, 2024), they have been in education branches mostly, letting un-covered the problems for bachelor studies in engineering branches. In short, we wanted to fill this gap by this initiating work presented herein. For this purpose, we have realised a dedicated CCI test (Calculus Concept Inventory) based on the on some engineering branches of the Polytechnic University of Tirana, Albania, and some FCI test data have been imported from other research recent activities with the permission of the organisers. Also, we performed a FCI test on a reduced sampler on High school students to compare the data. The CCI test is developed by J. Epstein and can be found in the PhysPort.net. Sometimes it is referred as PCA test -Preliminary Calculus Assessment test. It contains 22 elementary questions from basic calculus subjects. By a preliminary check we observed that all topics included therein have been developed in our high school level. Under such consideration, we do not claim objective Face Validity issues during the test. Considering the fact that in recent research relate to the conceptual tests has been underlined a low interest of students to complete testing and also a high rate of the fictive answering in non-curriculum-related tests in (Prenga, 2024), (Denaj, 2024) we have advised our students that the result of this test would be considered for further academical reference and also asked them to write the name on the test paper. By doing so, we have realised a testing procedure in common exam conditions. Hover, the participation has been voluntarily, as testing practises usually advised. Also, we gave an extra time to the student to fill inasmuch possible the empty answers, suggesting that letting blank the cell of answer is not wrong if student has no idea about the answering. Blank records have been rejected for calculating indexes but are included when using Rasch method and procedure. The initial measurement has been conducted on a group of 35 voluntary participants of students at engineering Branches from the Polytechnic University of Tirana. most of them belong to the mathematics engineering and physics engineering., it is very low number but, in this case, we have granted the randomizing procedure which not easy in these types of surveys, basically due to the voluntary participation, (Prenga et al, 2023). Also, it has been nearly impossible to realise pre and post-test in larger groups. another test, now for the physics conceptual knowledge has been developed at the same community by not exactly the same sample. It is reported in the table as FCI-1. We imitated also a direct comparison of the knowledge failure interdependencies based on the idea elaborated in (Prenga et al, 2023), by asking similar questions seen in different context, physics, and mathematics respectively and analysed the differences. The items were selected nearly identical, for example 'a) mathematics version: find scalar product of vectors $u = (1,2,3)$ and v= $(4,5,6)$; b) physics version: find the work done by force $F = (1,2,3)$ if displacement was $X = (4,5,6)$. For this case we obtained again significant difference on the answering them, but the number of participants was very low (13) so we considered it only as qualitative indicator. Just for mentioning, in this double test comprised by physics and mathematics similar parts we evidence 18% more scores in mathematics exam than in the similar test in physics, that is close to the result reported in (Prenga et al, 2023) for students on the education branches ate the Faculty of Natural Sciences, University of Tirana. Regardless of statistical issues on this preliminary measurement, we reaffirmed the problems remarks in original article that failures on conceptual knowledge in physics were not caused necessary by mathematical and calculus shortcoming, therefore we might compare two PCA and FCI test outcomes separately, similar as debated in (Prenga et al, 2024) for another community of students. Finally, for measuring the gain

$$g = \frac{\%CI_{PostTest} - \%CI_{PreTest}}{100 - \%CI_{PreTest}} \tag{9}$$

We have conducted a pre- and post-in our main testing exemplar, the CCI (PCA) calculus test.

Results

For analysing the data, we have contemplated main ideas and schemes introduced in the sections 2 and 3 above. We calculated indexes to recognise the tests' readability and usefulness in statistical sense, by assuming that if indexes of a certified and standardised tests are in desired range, the effect of heterogeneity on testing elements and other disturbing issues are minimised, so, the outcome of the test are useful for a broader community representation. Also, following similar arguments of (Prenga et al, 2023), misfitting events on the Rasch analysis' outcome can be interpreted as representative of core and substantial problem in education stage rather as consequence of direct heterogeneities on the group when the test is carried out. In the next step, we performed a schematic Rasch analysis for a deeper the basic mathematical conceptual knowledge analysis. As explained herein, this is a testing key step for validating our test procedure and qualification for generalisation of its outcomes, provided that several statistical prerequisites as randomness on selecting individuals and the homogenous distribution of the student's knowledge among them are difficult to be fulfilled by our group of interviewers. Also, we used the validity of the indexes as indicator of student's perception of CCI test, which in turn, can be considered as indirect measure of the education performance in general. Hence, by comparing the indexes of the different standard tests we might judge about overall student's knowledge in different filed of sciences. Procedurally, by employing the Rasch analysis we have calibrated the teste itself and based on its key statistical properties, final outcomes of the test as CI score, guessing level etc., have been used for further analysis and discussion. After performing those preliminary steps and procedures, we concluded in general observation and results. To empower a comparative analysis for knowledge problems in different field of science on our high school system and considering that several CI studies for physics knowledge have been realised so far, whereas on calculus conceptual knowledge there are no significant references, we have organised a pre-testing analysis using a double test on physics and mathematics. This test was constructed in such a way as only the context will make the difference. For example, we asked for evaluation of a derivative for a function which is continuous but has different shapes in different intervals of the support for calculus part, and next, we gave similar function for the velocity of point mass and asked to identify the acceleration, and the force exerted on point mass. We asked 10 questions in each of the tests so 20 in total, and each item was checked carefully to comply with standard programs of physics and calculus instructed in high school level. To avoid any problem regarding face validity issue, we have tried to randomise by selecting students from various first on the engineering bachelor branches several weeks after starting the studies. We chose this period arguing that some concepts in calculus and mechanics that have been elaborated in high school might need for soe complements, and also, to homogenise a little bit the knowledge provided to them. Also, if the student argues that the concept being asked was firmly unknown, we excluded his record for further analysis. Differently to the goal stated in (Prenga et al, 2023), we wanted also to check also the harmonization of the knowledge inventory by students, therefore, the test was developed unequally, and the items were mixed. The idea was that, if students were capable to interlink calculus concepts with physics concepts, he can try to review some solutions, or in the bast cases, he will solve correctly physics problem when he succeeds in the solving the matching mathematics item. So far, we started the discussion by evaluating test's indexes for the auxiliaries and our main tests, the CCI-PCA Unfortunately, we did not gather enough participants, and in total only 16 records were available for further analysis, Table 1. However, we can consider some of those finding because the randomizing procedure has been better than in other cases, despite the issues of the voluntary participation which by nature hamper the randomising but is impossible to resolve hereto. Initially we analysed the double test in dedicated on comparing relationship and mutual interaction physics and mathematics conceptual shortcomings. Or we asked to compute the scalar or vector product for tow given vectors for calculus part, and similar we asked to calculate the work done or the torque for given vectors of forces, displacement, the radius of forces and so on. etc. In this case, all records containing one item unanswered were excluded for part of the tests. Also, observed that the for the test carried out on the small but voluntary high school students' participants, all indexes were in the desired range, and few items resulted as misfitted.

Table 1. Comparison of the testing on mechanics and calculus algebra for students of the bachelor branch in engineering

	Ratio of correct answers	Misfitted ratio	Indexes	Guessing level	Reviewing ratio
General Physics	45%	12.5% (2 in 16)	Not all indexes in desired range Discrimination index 0.85	8%	15.6%
Elementary Calculus test part	58%	18.7% (3 in 16)	All in desired range	3%	6.5%

By this preliminary test we have observed a slight performance of students in calculus over physics. It reaffirms recent fining that physics conceptual knowledge does exhibit several problems, and in some case, the insufficiency of the laboratory support prevails as a factor, which is not the case for calculus learning. Interestingly, there are more inhomogeneities among students' knowledge in calculus compared to physics, as seen by the misfitted ratio, but for this point which is strongly a statistics-related origin, so we do not comment more because the number of participants was too small. The observation indicates that indexes, in the grade that they can be adopted for this small sample, won't be problem for calculus tests. This is enforced by the low ratio of guessing behaviour. We observed also that the ratio of reviewing (we permitted the reviewing of the answers, by asking just to highlight final decision) was higher in physics test, indicating that students probably have solved a physics item and after observing similarities with mathematics part, have reviewed it again. It indicates that students have stronger confidence in their calculus knowledge than in physics conceptual knowledge, and they probability believe that calculus concepts prevail on direct application. Notice that herein we didn't exclude misfitted items in calculating CI score, limiting therefore the rigor of some conclusions herein. However, using this group, we could not investigate for cause factors because the measurement is taken once without having the opportunity to change condition for examining the respective response. By this preliminary analysis, we observe that there are indicators that conducting a calculus testing would be fruitful at least compared to the physics conceptual knowledge test, and therefore, we can compare the results between them. Also, we have indicators that knowledge in mathematics seems to be more soldi compared to another science as physics and probably they are solid compared to the other too. In this sense, the results of the CCI test could be more generalizable in education quality perspective, because the CCI test outcomes are expected to have less sufferance or problems from the side effects or secondary sources. This is not to clear the way for the work, but just to clarify the validity of some results gathered form this low scale testing. Also, the results obtained in this preliminary test suggest organising a larger scale one to test some ideas elaborated herein in the level of hypotheses testing.

After this step, we have proceeded with main exam and analysis regarding the CCI-test, which is standardised and proven as an accurate instrument for measuring conceptual knowledge in mathematics for high school students. This test is carried out in two groups of students. The first sampler was com posed from high school students based on voluntary participation. We make also effort to realize an acceptable randomising for this sampling. His test is intended to evidence conceptual knowledge issues ate a broader level. The second test is conducted on students of the first of bachelor studies and all of them were enrolled in engineering branches. Notice that in general, the knowledges of students in this stage are mostly those absorbed from the high school studies, but those students are likely the ones that prefer mathematics and natural sciences more. For this group of students, we have conducted also a pretest and post-test to measure the gain on conceptual knowledge during university studies. Notice that in general, CI tests are intended to measure conceptual knowledge failures, the effect of the common-sense errors and other knowledge drawback, rather direct CI scores. I this regard, we were not focused on analysing the level of the CI scores. We considered also the fact that the failure on fulfilling statistical requirements for sampling procedures would hamper and even devaluates direct measures such as the CI scores-they report only the personalised group where students come from, so nothing to generalize! On the other side, several qualitative measures such as guessing behaviour, misfitted item ratio, indexes' features etc, would be useful for generalizing as indicatory features, or for using in comparative analysis. Notice direct scores of the test where generally below the 60% level that we considered as the threshold of basic knowledge in calculus, using similar argument as in (Hestenes et al, 1992). For bachelor students we obtained the scores at 52 % that is not too bad, even not good referring to the 60% threshold. For other groups we observed a high level of unanswered items that makes the estimation of the CI too fuzzy for consideration. Similar behaviours were reported for the FCI test by several references cited herein, so we qualify this just as a failure on procedures of the tests and not result of students' knowledge. It appears that volunteer students do not invest much attrition to answer difficult questions and prefer to finish the test quickly. Therefore, we do not qualify blank cells as a prove for failure on knowledge, although it suggests a degree of difficulty regarding the tests and particular items. So far, we have limited the assessment of the gain coefficient for the second group only. As explained above we piloted a CI-gain measurement, using similar arguments regarding the methodology and expected outcomes like in (Carlson et al, 2010), (Crooks et al, 2014), (Coletta et al, 2019). Inder such circumstances, the gain coefficient would report a composite effect of the academic advance, teaching efficacy and subject specifics. It makes the separation of the contribution difficult, so we have considered this parameter for qualitative evaluation only. Regarding realization of the test, we followed the procedure described in standard literature, see (Coletta et al, 2019)). The students are interviewed in beginning and at the end of the calculus courses, and results are gathered and interpreted collectively and in statistical basis. This procedure poses questionable issues regarding statistical rigor, because we have realized the measurement in an isolated part of the population-a group of students belonging to a certain branch, but nevertheless we believe that the results obtained for this small group represent in some extend the reality. Notice that during bachelor studies, students are not instructed specifically for subjects of the CCI test, but all concepts used therein are reimagined and rebuild during separate course of analyse, algebra and calculus. Also, the fearing for the influences of the statistical inadequacies for concluding remarks has been smoothed by employing the confirmation of indexes requirements by both tests, by assuming the argument of (Prenga et al, 2022) are applicable herein. The group chosen for measurement consists to 35 participants from the Faculty of Mathematics and Physics Engineering, 2023-2024. We were not able to gather more students for accomplishing sampling perquisites for several objective reasons. Notice that statistical prerequisites are very important to guarantee the reliability for this measurement if we consider that students that participate in the first and second test were not the identical. However, all the arguments and results of the analysis are conclusive for the group of students who are tested, and in any case, they are strong indicators for the whole population of students that were oriented toward applicative and engineering branches. After elaborating those critical elements regarding the representativeness of the outcomes, let's go to those last. Firstly, we highlighted that the Concept Inventory scores have advanced separately, that is, the progress after course was not exhibited only the algorithm solution advance, but on conceptual knowledge too. Here we obtained $CI_{PostTest} = 59.1$ (average 13 scores from 22) and $CI_{PreTest} = 36.36$ (average 8 points form 22), so the gain obtained

$$g_{CCI} = \frac{\%CCI_{PostTest} - \%CCI_{PreTest}}{100 - \%CI_{PreTest}} = 17.5\%$$

The moderate value g~18% obtained herein is lower than the gain coefficient of the FCI scores measured on similar community of students and at the same time interval, (Kushta et al, 2023). Notice that this parameter is a complex measure of mixed effect as elaborated above, so we do not explore on the weight of any of them. For an accurate measurement of this parameter, we must reorganize the test and some preliminary procedures especially, as to match to strictly defined prerequisites, and to exclude other influences, that remains for another work. So far, we might judge on this parameter by considering the initial knowledge level indicated by the CI score and assume that that conceptual knowledge shortcoming could be persistent, conditioning a low gain for low initial CCI scores. On the other side, low value on initial scores of CCI favour high g-value. The balance of those two opposite effects cannot be evaluated quantitatively, but their significance can be clarified if the test fulfils statistical requirement. Under our condition of measurement, we might argue that the gain is moderate and at the same range as the gain coefficient of the FCI scores for samples taken from the same community of students reported in (Prenga, 2024). It supports the idea that problems and challenges for improving conceptual knowledge on physics and mathematics are similar and of a more general nature. A good part of our interest has been paid for the general pattern of students' perception for a standard test in calculus and elementary mathematics. It can be achieved by analysing the indexes of the CCI test. Just for highlighting the significance feature of the test, we underline that the indexes for all tests fulfil the prerequisites of reliability, therefore all the findings are trustworthy in the indicatory sense. Particularly, it indicates that indexes are themselves good indicators of students' perception for the CCI test.

Table 2. The indexes for PCA. For comparison are shown indexes of a recent FCI from literature

Test Name	Details	Difficulty index	50 to 50 Discrimination index	Reliability Index (Point Biserial)	Self-consistency index	Discrimination power
CCI [PCA]	Sampler: High school students, 2023	0.52	0.63	0.21	0.66	0.82
CCI-PCA Pretest	Sampler: Bachelor, 2024	0.57	0.73	0.24	0.76	0.91
CCI-PCA post-test	Sampler: Bachelor, 2024, 2024	0.53	0.77	0.65	0.78	0.93
FCI-1	Sampler from high school students	0.58	0.65	0.44	0.66	0.87
Reference values		≥ 0.3	≥0.3	≥0.2	≥0.7	≥0.9

As seen in Table 2, indexes of the CCI test lie on desired (reference) range for both pre and post-test. Therefore, the CI gain is measurable and represents the group, and inasmuch this last represent a larger population, its value calculated herewith can attribute the reference gain for the CI in Calculus test in general. We observe that indexes of PCS are in valid range differing from the fact that for FCI test. Notice that FCI-1 test has been conducted on a larger group than our CCI, so this finding consists in an important indicator regarding overall students' perception of the CCI test. We have classified the fulfilling of the index's validity range as an indicator of acceptable knowledge on the subject being tested. Also, it qualifies its outcome as CI scores, misfitting ratio, guessing behaviour rates, as comparable features with similar ones obtained by other measurement as the FCI-test for example. So, it appears that (a) the CCI test has 'read better' the student's knowledge compared to FCI, or (b) their knowledge indicators in mathematics are better. Both alternatives (a) or (b) converge to the suggestion that students react better for a mathematics' test than for physics one. We might qualify this as an indicator that students understanding mathematics is not at the bottom of their overall knowledges. We talk for indicators, because the absolute knowledge is measured by the raw CI score, where we didn't observe significant difference with those found in older FCI tests. Current scores obtained on CCI pre-test was at 41.6% level, and FCI test scores reported in (Prenga et al, 2023) range 35%-42 % for different group of students interviewed. However, regarding to comparative analysis of the knowledge in mathematics and physics those values do not tell much because the two CI tests used herein are different instruments and therefore a direct CI scores comparison is not interesting. This is why we considered the comparing of indexes validity instead of direct comparing the corresponding CI scores for CCI and FCI tests. In this regard, CCI test is perceived formally well, indicating a solid basic structure of the elementary concepts by high schools. Notice that the pre-test has a lower discrimination power than the wanted value (above 0.9). Remembering that the tests that we used is by nature enough discriminatory and well structured, this finding indicates that for the group where the test is carried out there are several issues regarding the distribution of the abilities of students to solve this test. Whereas it penalizes quantitative analysis for the findings herein, it suggests that he outcome of the Rasch analysis should be taken with precautions. Noticing that Rash procedure is used more for calibration that for direct measurement, this observation suggests that when performing Rash measurement, we must exclude misfitted occurrences to improve a better matching of system features with theoretical assumption. However, this choice is not adequate for our case because the number of the students is too small for further reductions (35) and the number of questions is also not sufficiency high (it contains 22 items). It follows that quantitative outcomes resulting after employing the Rasch model for this test are open to doubt, but again, we highlighted thus element to avoid any subjectiveness and for clarifying the reader regarding strict numeric values, whereas regarding the dynamics and phenomena, there is no doubt regarding their certainty. Another intriguing parameter regarding efficiency of the CI test is related to the guessing behaviour which acts as a hidden feature. We observed that that in our CCI test the rate of the guessing behaviour was around 7%, remarkably lower than corresponding value of the FCI-1 test. On all those observations, it resulted that the CCI outcomes were at the same level or even better than the finding referred for the FCI outcome in the tests conducted across the country. It resulted that raw score and statistical ‘goodness’ are better for the Concept Inventory Calculus compared to Physics CI-test, suggesting different methods and efforts for improving student’s academic progress students.

Conclusions

By harmonizing the Rasch model with several physics’ models, we were able to measure in this observation and conclude regarding them. Firstly, we observed a slight dominance of CI and knowledge in Calculus compared to those gained in physics courses. This finding is consistent with a common belief among education experts and faculty professors that theoretical aspects as calculus concept can be assimilated easier compared to physics law and concepts. By our measurement and analysis, it resulted that for the group of students tested who embodied also some representative features, conceptual knowledge in calculus provided by the high school is at a moderate to low level, and its improvement after university courses is also moderate to low. Because advanced knowledge students prefer mostly studying abroad or following medicine branches and law, we acknowledge the upper measure as more representative, but also, it is only ‘moderate’ indicating that teaching and learning maths and calculus do have problems. This is very important issues because it conditionate not only the success of studies in broader spectre of branches but also, students’ preferences for maths and engineering branches. We observed that results of Conceptual Knowledge tests seem to be lower than scores obtained on common exams, that reveal the importance of conducting CI test for a better understanding of students’ knowledge on this fundamental science. By reading also the events such guessing behaviour as complementary information for students’ confidence in his own knowledge, we highlight that generally our students are clear and act undoubtedly, that support the use of CI test or testing and exams by alternatives for studying purpose only, because in general, the expected guessing would be around 7%. Also, the level of outfit occurrences for Rasch analysis is not extensive, that support the validity of our finding although rigid statistical prerequisites were not meet. When compared with FCI test that have been conducted in larger and broader scale, we can argue that the results obtained by our CCI test represent better pattern, regarding raw results and scores, and their representative and resolution power as measurement instrument. This support the idea that Conceptual Inventory test can be used effectively for measuring knowledge in mathematics and calculus, regardless the fact that the achievement on the algorithmic procedure instructed by the lecturers and methods can be sufficient for solving problems in most of the cases for in high school level.

References

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

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction*, *28*(2), 113-145.

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.

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

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

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Epstein, J. (2013). The Calculus Concept Inventory— Measurement of the Effect of Teaching Methodology in Mathematics. *Notices of the AMS, 60*(8), 1018-1026.

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

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

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

Liu, X. (2010). *Using and Developing Measurement Instruments in Science Education: A Rasch Modelling Approach*. Information Age Publishing, Charlotte, NC.

O'Shea, A., Breen, S., & Jaworski, B. (2016). The development of a function concept inventory. *International Journal of Research in Undergraduate Mathematics Education*, *2*, 279-296.

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.

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

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

Prenga, D. (2024). A Thematic Review on the Combination of Statistical Tools and Measuring Instruments for Analyzing Knowledge and Students' Achievement in Science. *European Modern Studies Journal, 8*(3), 687-706.

Prenga, D. et al. (2023). J. Phys.: Conf. Ser. 2872 [accepted, to be published]

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

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

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

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

Savinainen, A., & Scott, P. (2002). The Force Concept Inventory: a tool for monitoring student learning. *Physics Education, 37*(1), 45.

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), 010105.

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

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

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

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

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