A CONCEPT INVENTORY TESTING APPROACH FOR THE ASSESSMENT OF HIGH SCHOOL STUDENTS' KNOWLEDGE ON CLIMATE, EARTH SCIENCES AND ENVIRONMENTAL HEALTH HAZARDS

DODË PRENGA, SONILA BOÇI, RUDINA OSMANAJ

Department of Physics, Faculty of Natural Sciences,

University of Tirana, Albania

e-mail:dode.prenga@fshn.edu.al

Abstract

The teaching of climate and environmental issues to undergraduate students in our education system and usually elsewhere is primarily oriented towards educating them with responsibility and care for nature and therefore, the corresponding knowledge is offered through several subjects mostly as supplementary or illustrative acquaintance. These specifics impose changes in the corresponding learning structure compared to fundamental subjects such as physics, mathematics, etc., inflicting additional complexity on studying and measuring it. This work was initiated as an assessment of the level of knowledge of students on climate, environment and earth sciences within the framework of the 2024 National Conference on Climate and Environment of the Department of Physics, and then is expanded into a comparative analysis of the structure of learning and acquisition of knowledge which are offered through distributed sources, with those presented by structured and fundamental disciplines In this regard, this work also aims to evoke some methodologies for measuring the level of students' knowledge in general and identifying problems in the learning and teaching process. This analytical assessment was carried out through the CI Knowledge Inventory technique, index theory and the Rasch model.

Key words: *Concept Inventory, physics knowledge, index theory, the Rasch Model, sociometric measuring and instruments.*

Përmbledhje

Informimi mjedisor dhe pajisja e nxënësve te ciklit para universitar me dije për klimën dhe çështjet mjedisore tek ne dhe përgjithësisht gjetkë është orientuar kryesisht në edukimin e tyre me përgjegjësinë dhe kujdesin për *natyrën, që për pasojë, ofrohet nëpërmjet disa lëndëve kryesisht si dije plotësuese apo ilustruese. Këto specifika imponojnë ndryshime në strukturën e transferimit të dijeve në këtë rast krahasuar me ato të lëndëve bazë si fizika, matematika etj., gjë që bën matjen dhe studimin e tyre jo evident. Ky punim u iniciua si një vlerësim i nivelit të dijeve të nxënësve për klimën, mjedisin dhe shkencat e tokës në kuadër të Konferencës Kombëtare për Klimën dhe Mjedisin të 2024's të Departamentit të Fizikës, dhe me pas u zgjerua në një analizë krahasuese të strukturës së të mësuarit dhe përvetësimit të koncepteve në dijet mbështetëse dhe të ofruar në mënyrë të shpërndarë, me ato të ofruara nga disiplinat kryesore. Në këtë kuptim, punimi synon të evokojë edhe disa metodika për matjen korrekte të nivelit të dijeve të nxënësve në përgjithësi dhe identifikimin e problemeve në procesin e transferimit të dijeve. Ky vlerësim është realizuar nëpërmjet teknikës së analizës së të mësuarit konceptual, teorisë së indekseve dhe modelit Rasch.*

Fjalë kyçe: *Të mësuarit konceptual, teoria e indekseve, teknika Rasch, instrumentet dhe matjet sociometrike.*

Introduction

In our secondary education system, climate and environmental subjects are instructed purposedly in geography, and as illustrative or complementary part of the primary knowledge in physics, chemistry and biology. Though the structure of teaching and learning in this case is less organized compared to fundamental disciplines, plentiful information and data regarding climate issues attain students continually in a non-academic pathway. We assume that as result, the corresponding knowledge becomes consolidated, and therefore, it can be assessed in principle by using common instruments such as procedural or conceptual tests.

Remember that a procedural test assesses students' ability to follow pre-instructed steps, straight application of theoretical principles and formulas, whereas a conceptual knowledge test aims on measuring deep understanding and students' ability to use fundamental concepts for explaining phenomena or system behaviour. Conceptual knowledge tests are known also as Concept Inventory (CI) test, following the idea of Force Concept Inventory (FCI) analysis introduced by (Hestenes et al., 1992), and developed through numerous applications and theoretical works, (Crooks et al., 2014), (Andrich, 2005), (Savinainen, 2002)., etc. FCI test consists of a set of 30 multiple-choice items which are drafted in a such a way as to measure the deepness of knowledge in mechanics, contrasting to classical exercises which investigate students’ understanding on a certain part of the subject. Moreover, CI tests are used typically for investigating knowledge shortcomings, errors and common-sense interpretation, and problems in learning and teaching in general. So, in (Hestenes et al., 1992) has been emphasised the exclusive capability of the FCI to analyse the occurrence ‘six common sense errors in mechanics’, along of the assessment of conceptual knowledge. Next, the CI exemplars in other disciplines have been developed and expanded, enabling a diagnostic view of knowledge and its shortcomings, and adequate measurement of conceptual understanding of sciences. In this context, we believe that a qualitative assessment of information about climate and environment could be achieved by using similar methods and tools as in formal Concept Inventory practices, even though as mentioned briefly herein, the structure of this knowledge does not favour its’ rigour measurement.

Despite this, there are two factors that contribute to the consolidation of this knowledge, by activating a compensation mechanism in this respect. Firstly, the precaution of the teachers to accent civic responsibility to the environment on modern schooling. Second, the knowledge on climate and environment issues appears as an outcome of students’ critical thinking and active involvement., which provide a better academic proficiency in terms of discussion in (Vilia et al., 2017), (Kim & Sunderman, 2005), (Kumar et al., 2023) etc. Bearing in mind those elements we accept hereby that this specific type of knowledge can be measured and analysed using CI type test as the instrument. Additionally, considering our recent FCI measurements referred in (Prenga et al., 2023), (Kushta et al, 2023), (Hamolli & Prenga, 2024), a comparative pedagogical view can be developed. From this last feature and to encourage young scientists to contribute to those studies, we will exhibit in the following the principles of CI analysis, suggesting also for more detailed arguments (Hestenes et al., 1992), (Crooks et al., 2014), (Planinic et al., 2019) etc. Finally, the Concept Inventory on Climate knowledge test (CCI) used in this work is drafted in collaboration with teachers and experts of the field based on the relevant literature used in our high school system.

Drafting the ad-hoc Climate-Concept Inventory test

For drafting the test, beside the specifics mentioned above, we considered also that this atypical knowledge could be eventually affected by students’ beliefs, conscience and social responsibility. Therefore, the proposed items of the CI test were inspected and reviewed by teachers at the secondary school, who manage better to invoke non-academic factors in the assessment processes.

Noticing that the selection of alternatives in a multiple-choice test is a decision-making process according to Item Response Theory (IRT), see (Klymkowsky et al., 2020), (Planinic et al., 2019), for avoiding additional shortcomings related to this scoring inconveniences such as contextual confusion, terminology issues, etc., we drafted a set of basic questions for the CCI test, that was verified latter by using Indexes, see below. Each item has been based on the textbooks used in our high school education, such as Physics 10-11 (Pople, 2022), Physics 12 (Sang et al., 2022), Geography 10 (Herzig & Manaj, 2016), Chemistry 12 (Hill et al., 2022), Biology 10-11 (Jones & Jones, 2016). So, for informative and know how questions related to elements of the atmosphere composition, temperature, pressure, winds, the biosphere, the greenhouse effect, the hydrosphere etc., we considered Geography 10 (Herzig & Manaj, 2016); whereas for phenomena based topics such as electricity production, energy resources, thermal radiation , impact of power plants using fossil fuels, radioactive waste problems, we considered generally Physics curriculum for grades 10 and 11 (Pople, 2022).

Some of thinking-based questions have been selected also from basic programs of physics, which we are describing shortly herein. So, in the Physics for grade 12 (elective) (Sang et al., 2022) throughout the study of concepts, phenomena, and laws of physics, students explore environmental issues related to energy production methods, power plants and their environmental impacts. Chemistry provides information about their distribution in nature and their importance in the life cycle of various organisms (Hill et al., 2022). Phenomena related to environmental pollutants such as vehicle exhaust gases and methods to reduce them, acid deposition and its effect in the environment, the increase of the greenhouse effect and global climate change, including recommendation for using least toxic substances, are explained in Biology 10-11 (Jones & Jones, 2016) and Chemistry 12 (Hill et al., 2022).

We observed that climate and environmental information provided by regular curricula, despite the dispersed nature of sources and the spread over the year, is rich, qualitative, and include elements of conceptual knowledge and in deep learning in general. Therefore, we concluded that this knowledge can be measured and analysed similar to Physics or Mathematical by using Concept Inventory analysis, regardless of remarks that we listed in the preliminary stage of test's writing. Based on those arguments and prerequisites we drafted 30 initial items of the CCI test and piloted it to achieve a reliable instrument. The questions were basic mostly, and only for few of them some of precaution on judgment was required. So, when we ask for the correct proportion of the oxygen in the atmosphere, it is sufficient to remember what has been learned directly on the class; but when we asked to indicate the consequences of global warming by selecting alternatives such as *a=nothing;* b=…;,c=…;, *d=inflict changes that affect natural equilibria and ecosystems*; a thinking step is need.

Data and method

The final scope of the analysis has included also a comparative view of CI test and a discussion of some aspects of teaching and learning sciences in general. Therefore, beside the data collected on purpose, we used also some FCI tests results gathered in the framework of master's degree thesis by students of Physics Branch at the Faculty of Natural Sciences. For ethical reason based on the prior agreement with teachers and students, we are not disclosing the schools where CI measurement have been made, disclosing only the regions which were in Saranda, Tirana, Durrësi, Dibra and Kruja.

The CCI tests have been conducted in small groups of high school students from Dibra and Tirana. Notice that several low-scale tests have been conducted on small groups based on stratified and opportunity sampling as indicatory or pilot measurement. The elaboration of the data is realized by using our ad-hoc algorithms written in MATLAB/OCTAVE, but interested readers who do not admire writing codes, can consider alternative commercial software like WINSTEPS or even EXCEL based on statistical guides provided in dedicated pages such as (https://real-statistics.com) etc. Finally, because of several objective and subjective hitches on organizing the tests and the expensive cost for realising a full random sampler, we have conducted the measurement usually on certain group of students based on the convenience and voluntary sampling, and for the piloting stage we used purposive sampling.

Introductory elements of the Rash analysis

Consider the result of the CI test which roughly is organised in a table with N_{items} columns $N_{responders}$ rows. The matrix element M(i,j) of this table is the letter (a, b, c, d) which is selected as the answer by the student or could be also empty. This original table to input data matrix $T(i,j) = (0,1)$ where the correct answer is assigned by (1) and the wrong one by (0). Also, empty element could be present. Naively we might assume that those numbers represent the probability that student (i) resolves the item (j). The raw measure of students' knowledge (regarding this test) can be indicated by the corresponding mean $S_i = \frac{1}{Nr_{Items}} \sum_{j=1}^{Nr_{Itmes}} T(i,j)$. It gives also the estimated probability that the student understands perfectly the subject asked on the CI test. By averaging this quantity among all students, we have an estimator (score) of students' knowledge on that discipline. For the FCI test for example, if this average falls below 0.6, that is the CI score ae less than 60%, it is said that this group (or, if we have realised correctly a random sampling, the whole population), do not have sufficient knowledge in Mechanics! By averaging across the columns we have an estimate of the easiness of the test CI seen from our group, $S_j = \frac{1}{Nr_{Students}} \sum_{i=1}^{Nr_{Students}} T(i,j)$. However, those quantities are not correct in the terms of Item Response Theory. So we can contradict the assumption of $P(i,j) = T(i,j)$, that is, the probability that student (i) could answer correctly to question (j), by supposing that he(r) has selected the alternative by random and make it, or he/she failed to select the correct alternative at a bad moment of ambiguity, despite he practically knows the answer! Second, the presumption the linearity and homogeneity of the data scale looks also incorrect.

For this argument let's compare the differences of knowledge between individuals (A, B) who had 90/100 and 70/100 of total scores respectively, and (C, D) who had 30/100 and 10/100 respectively. In this scenario, student A (which has very good knowledge, hence, he has solved almost all the test) has resolved a very difficult problem compared to student (B) whereas (C) has resolved one extra easier item compared to (D)! In (Savinainen & Scott, 2002), (Sands et al., 2018), etc., the reader can obtain broadened arguments on this subject. Danish mathematician D. Rasch has invented an elegant sociometric model that is able to resolve those shortcomings. The model initially used dichotomous response variables which is known under his name (Rasch, 1961), and later Andrich has advanced it for polytomous variables or Likert scale responses, (Andrich, 2005), (Wright, 1982; 1977). Interestingly, the model can realise the calibration process of the measuring instrument, the CI, too. One starts by calculating students' success rate (a_i) and item's solution rate (e_j) or fraction of students who answered correctly the question: $a_i; e_j \leftrightarrow p(i;j) = \frac{1}{N;n} \sum_{j;i=1}^{N;n} T(i,j))$. The measure of the students' success is calculated directly $\frac{SuccessProbability}{FailureProbability} = \frac{p_i}{1-p_i}$ and similar is the easiness $\frac{p_j}{1-p_j}$ of the whole test. Next based on the IRT the probability of success given a cause y is a logistic function $p(1,y) = logit(y) \equiv \frac{e^y}{1+e^y}$, one calculates student's ability to resolve the test and item's difficulty by the following

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

which are measured in logit units. In (1) it is defined the logit difficulty measure of the test $ln\frac{1-e_j}{e_j}$ instead of the easiness $ln\frac{e_j}{1-e_j}$, but we will see that it is more practical. So, assuming that Conceptual Knowledge is unique, α_i and δ_j are basic indicators for the knowledge being tested from the students' group standpoint, therefore the estimated probability:

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

would express better the rate of success for a particular event (i ,j). However, the measurement is not perfect, nor is homogenous the group of students. Some bias and measurement errors are present in (2) so the instrument must be calibrated to minimise their effect. Therefore, to improve the significance of (2) for given data (1), one replaces iteratively matrices $P_e^{n-1}(i,j)$ with new ones $P_e^n(i,j)$ starting from $P^0(i,j) = T(i,j)$, by a straightforward update of the new values based on overall variance and residuals calculated using pairs $\{P_e^{n-1}(i,j), P_e^n(i,j)\}$, till a threshold criterion for the residuals is met, so

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

Ideally α_i^{final} and δ_j^{final} would match and be symmetrical to zero. Unmatching events and asymmetries are important information and findings. All deviances constitute information for the testing process, for shortcomings in testing and statistical issues, efficacy of teaching and learning and many more. Alos, the Rasch procedure transforms CI test to a calibrated instrument with linear and homogenous scale, that facilitate significantly the descriptive analysis. For example, recognising that ability shift between $\alpha_m = 3\ logit$ and $\alpha_n = 2\ logit$ is the same as the distinction between $\alpha_k = 1\ logit\ \alpha_k = 2\ logit$, we might analyse directly the histograms in logit units as classes or levels of knowledge.

Also, the differences between probability estimate $P^{final}(i,j)$ and original values can be viewed as another estimator of the CI testing applicability. Elevated outlier-sensitive fit (outfit) and information-weighted sensitive (infit) and are signals for insufficiency of statistics or abundance of pathological events. The discrepancies $P_e(i,j) - T(i,j)$ can be used as diagnostic tools for the adequacy of the testing process and occurrence of the irregular event etc. For example, if the ability of student (i) is smaller than the difficulty of the item (j) $\alpha_i^{final} < \delta_j^{final}$, $P_e(i,j) < \frac{1}{2}$ from the equation (2), and if $T(i,j) = 1$, the correct answer has been chosen by accident, or by guessing. Similarly, for $\alpha_i > \delta_j$ and $T(i,j) = 0$, the student has failed from contextual confusion, or other factors that ghosted students' knowledge, etc. In general, the Rash model and analysis improves the resolution power of CI test and extends the analysis to variables that are invisible to classical exams. Also, we might compare the Rasch analysis for CCI test with those of CI test on structured disciplines as Physics to employ a retrograde analysis for understanding knowledge transfer shortcomings in this last.

Considering that the Rasch procedure acts as calibration step for the test itself, the item's difficulty parameter for our CCI's will depicts also the level of information shared or diffused among student's communities, their understanding of the importance of the subject, level of integration of environmental topics in standard high school education in the country and many more. An intriguing extension and comparison can be enabled by using the idea elaborated in (Prenga, 2024), where the CI data are grouped to produce a multi-scale assessment that can be used in polytomous Rasch analysis. The polytomous Rasch model evaluates the probability that student would achieve the score level h, given the threshold parameter $\{\tau_j\}$ that measure the difficulty of obtaining scores h relative to h-1, by using the following estimate probability for success

$$P\left(x_{ij} = h | \beta_i, \delta_j, \tau\right) = \frac{exp[k\left(\beta_i^{(k)} - \delta_j^k\right) - \sum_{j=0}^{h} \tau_j]}{\sum_{k=0}^{m} exp[k\left(\beta_i^{(k)} - \delta_j^k\right) - \sum_k^h \tau_k]} \qquad (4)$$

In (Prenga, 2024) it is proposed to consider the threshold values as measure of learning advance under several appropriate assumptions.

Using Indexes for certifying the test

Realizing a straight random and representative sampling is expensive and impractical. Moreover, students are not willing so much to participate testing after school time, whereas during regular schooling such test are nearly impossible for administrative purpose. As by our recent experience, the number of students which fill the test without evoking their knowledges, just to finish it, is not neglectable. These inadequacies can hamper seriously the outcome of the test and jeopardises the Rash evaluation, because it localises pathological events after using their data in all running steps. To deal with statistical flaws of testing and its imperfections in general, the Indexes Theory can be very useful. The CI test's indexes are estimators of integrity, significance, validity, discriminatory power, difficulty measure etc., and can be used to overcome several impediments due to improper statistics. In Table 1 we have summarised the main indicators according to this theory, suggesting the reader to consult (Ding et al., 2006) for details. In this sense, indexes can be used also as preliminary stage of the Rasch CI test analysis, enabling reliable outcomes of this last.

Table 1. Summation of indexes

Index's name	Formulae	Desired range
Difficulty index	$\delta_i = \frac{n_{Correct}^{answer}}{N}$; $\delta_{test} = \frac{1}{n_i}\sum_{i=1}^{n_i} \delta_i$	0.3-0.9
Discrimination index	$d = \frac{(N_{highestScore}^{Succses} - N_{lowestScore}^{Succes})}{N}_{p\%}$ $D_{Test} = d_{avg}$	>0.3
Reliability index (Cronbach alpha)	$r_i = \frac{\tilde{x}_i - \tilde{x}}{\sigma(x)}\sqrt{\frac{\delta_i}{1-\delta_i}}$;$r_{avg} = \frac{1}{n_i}\sum_{i=1}^{n_i} r_i$	>0.2
Discriminatory power	$\rho_{test} = \frac{n_i}{n_i - 1}(1 - \frac{\sum_i^{n_i} \sigma(x_i)}{\sigma^2(x)}$	>0.7
Self-consistency index	$\Delta = \frac{N^2 - \sum_i^{n_i} f_i^2}{N^2 - \frac{N^2}{n_i + 1}}$	>0.7

Legend: i indicate test item, N is the number of students, n_i is the number of correct answering of item (i), f are frequencies, σ are standard deviances, $\tilde{x}$ are the respective mean.

Notice that statistical inadequacies on CI testing would be mirrored on the misfitting and devices calculated by the Rasch model, but at the same time, all final outcomes such as students' ability, items' difficulty, the threshold difficulty for multi-grade evaluation test, etc., would result affected.

Therefore, we used Indexes' analysis to apply a preliminary filter, in the sense that if indexes of a certain CI test were found out of desired range, we reviewed or re-conduct it, and quantitative assessment for outcomes of the Rasch model calculation has been effectuated only if indexes of the test were found in the valid range. Of course, we also analysed indexes of test separately for diagnosing its integrity, reliability, discrimination power, difficulty level etc.

Data analysis and discussion

Following the above arguments and discussions, we have assumed that under some circumstances, the processed information and high school students know how on climate and environmental topics can be approached to a latent variable that can be measured with similar procedures and instruments like we do for analysing Conceptual Knowledge in physics or Mathematics.

Piloting the Climate Test

The initial CCI test draft consisted of 30 questions as explained above and is piloted to a group of 13 students enrolled in Physics, Faculty of Natural Sciences, which consisted to a convenience sampling, because we were not able to realise the random sampling. We assumed that if our CCI test would not pass the 'indexes' diagnostic', it must be reviewed thoroughly, otherwise, we would follow with a partial revision or amendment. By this preliminary examination of the test validity, we observed some of the indexes were found out of the validity range according to the reference (Ding et al., 2006).

After analysing the indexes of the items themselves, we decided to remove 8 questions with highest difficulty index, presuming that too difficult items mirror the 'face validity' or 'contextual issues', due to unclearness of the formulation or other objective reason from student's point of view. As result, our CCI-test was reduced to 22 questions.

Next, we compared CCI test's outcomes with those of a certified and standardised CI test such as Force Concept Inventory test, using the data from previous surveys on conceptual physics knowledge conducted on groups of students from the branches of Informatics, Physics, Mathematical and Computer Engineering branch, at the Faculty of Natural Sciences, University of Tirana, 2023-2024. For this step, the CCI test was realized in a voluntary participation group from those branches, (27 participants, in two rounds, 12 and 15 students each).

Table 2. Indexes of FCI and Climate CI tests in pilot trial

Test Name	Size	Difficulty index	Discrimination index	Reliability index	Consistency index	Discrimination power
FCI	85	0.58	0.63	0.24	0.66	0.81
CCI	27	0.53	0.8	0.65	0.57	0.83
Desired		>0.3	>0.3	>0.2	>0.7	>0.9

We qualified the test for this analysis after observing that its indexes were within the validity ranges, Table 2. We observed also some dissimilarities whose do not constitute any disqualifying issue regarding this stap. So, the small difference between reliability and difficulty's indexes compared to FCI-test, is expected because of the relatively easy questions of the CCI items, and the lower value of self-consistency index indicates that some questions were not aligned appropriately with the whole test. Considering that the shortened version of 22 questions has passed the validating test according to indexes analysis, we qualified it as our legacy instrument for studying climate and environmental knowledge in a group of high school students.

Discussion forthe CCI test measurement

The survey for measuring and analysing students' knowledge on climate and environmental issues by using our ad hoc CCI test is conducted in several small groups of undergraduates in Vlora, Lezha, and Peshkopi district, totalling 104 participants. For comparative analysis and pairing FCI: CCI testing, both tests have been carried out a group 53 students in Peshkopia district, grace of a very fruitful collaboration of the teachers at the secondary school the hard work of our master's degree thesis student, (Nela, 2023). They asked to not be mentioned for ethical purposes, and we respect their will, but their collaboration and devotion to conduct a rigor testing, has been crucial. In the first stage we conducted an indexes analysis, Table 3.

For reference purpose, we have used also the results of a recent FCI testing conducted in a high school students (Duka, 2022), because it has been impossible do conduct successive test on high school students, and this reference is supported by very similar testing conditions in the sense of region, rules applied and the requirement for rigor testing. Initially we observed that indexes for a FCI test carried out in a high school, in similar conditions and circumstances fall within the validity range, whereas the difficulty index of the CCI test of 22 questions, was out of range. We observed that by excluding four more items from the 22-itmes CCI test, the difficulty index for the whole test comes within desired range (>0.3).

Regarding the occurrence of FCI indexes out of validity range in the recent similar measurements has been interlinked with the heterogeneity of various types, (Prenga et al., 2023). When discussing for CCI testing process, the non-adequate sampling, imperfections of the CCI test itself, and heterogeneities of knowledge among students do affect simultaneously the result, therefore we have chosen to act on the test itself to reduce the negative effect related directly with it. Therefore, for quantitative references, we considered the test shown in second column of Table 3, that is the CCI of 18 basic questions.

Table 3. FCI and CCI indexes. The asterisk * indicate that the test has been conducted in similar condition but not at the same group.

	Initial CCI, 22 items	Final CCI, 18 items	FCI*
Difficulty index	0.2122	0.5239	0.2389
Discrimination index	0.5675	0.7295	0.5035
Reliability index	0.1304	0.4356	0.1761
Discriminatory power	0.445	0.4453	0.4652
Self-consistency index	0.9723	0.9982	0.9527

In principle, the Table 3 indicate the validity of the CCI test, but we can use a retrograde approach for interpreting the values shown on each of three columns. We observe similarities between indexes of CCI-22 and FCI*. Ion the first view, it suggested that the result obtained by the CCI test is reliable and report a knowledge reality, like the standard FCI* does. Looking on the arguments for the FCI*test reported in (Prenga et al, 2023), this similarity suggests that from student’s perception perspective, the structure of the knowledge provided by different methodologies is similar.

4.3 A comparative analysis for FCI and CCI tests

Recognizing the features discussed in the preceding section, we have conducted the core measurement in a group of 53 students as mentioned above.

Students. We observe that in this case, the correspondences presented in Table 3, are confirmed and the similarities are emphasized. It resulted that indexes of our CCI-22 items test are very similar to the FCI test conducted in the same classes of the same school. Therefore, we recognise them as typical for conceptual knowledge in a more general view. Also, we can comment the findings interchangeably, considering the standard and certified nature of the FCI test and commentaries regarding them. In this regard, the result of the CCI test indicates for a basic to low level of knowledge on climate and environmental issues.

For the reduced CCI test of 18 questions, the score is around 42%, which looks like in several FCI measurement reported recently in (Prenga et al., 2023). Remembering that remining items on the shortened CCI test were basic, we comment this second similarity aside the resemblance on indexes as indicator of a persistent problem on the transmission of the conceptual knowledge in high school, rather than a problem of the dispersed source of lecturing climate and environmental topics.

Notice that the level of FCI score obtained for high school' students recently has been argued as basic in a relative scale, (Prenga et al., 2023), despite the classification of original scale by (Hestenes et al., 1992) where less than 60% indicates lack of fundamental knowledge in mechanics. From a retrograde view, those findings suggested that complex conceptual questions look very difficult for most students, regardless of the discipline, therefore, including such questions in an exam or test would reduce the assessment capacity of those instruments. In classical mechanics for example, using a simplified FCI test would elevate the resolution power of measuring conceptual knowledge for the groups of students who participate our testing.

Another intriguing comparison and findings resulted from the analysis of the student's abilities to answer CCI test. Again, there is a similarity between classes of students' abilities to resolve both testes, or equivalently, levels of knowledge on CCI and FCI. Both can be grouped into 5 'natural' categories, as seen from optimal histograms of abilities, which are calculated using the Rasch model, Figure 1. For readers' clearness, we remember that theoretically the optimal histogram result by fixing the number of bins by the condition $\frac{\partial}{\partial n_{binning}}[< \mu >; \sigma(\mu)] = 0$, where μ is the mean and σ its variance for a given variable x, here, the students' abilities $\beta_{student}^{final}$. However, this formula is not applicable here because of the small number of points (number of students, 53), that are grouped in the histogram. Here we employed a minimal effect idea, by trying several values of the binning number based on optimal corresponding entropy algorithm, see (Prenga et al., 2023) for details. Note that the small number of points (53 for the comparative tests shown in Figure 1) disfavour generalisation, but we obtained the same optimal histogram for the CCI test on the 104 participants sampler.

Therefore, it looks that natural levels of understanding for two different knowledge are similar, which is added to other similarities observed above. It merits attention, but more data and improved CI test are needed. Notice that for larger group for FCI tests we have obtained 6 classes, (Prenga, 2024) which nevertheless constitutes to a non-significant difference.

On the other side, this finding suggests that the resolution power of CI tests is not good enough, and therefore, for correctly measuring the knowledge on our groups of students, a better alternative would be by using procedural examination and test. It looks that our students are accustomed with procedural learning rather than going in deep to fundamental understanding of the concepts.

Features of the knowledges in climate and environmental issues

It resulted that some specific for our CCI test are obvious. Again, we used as echelon for this discussion, the results of the standardised FCI test.

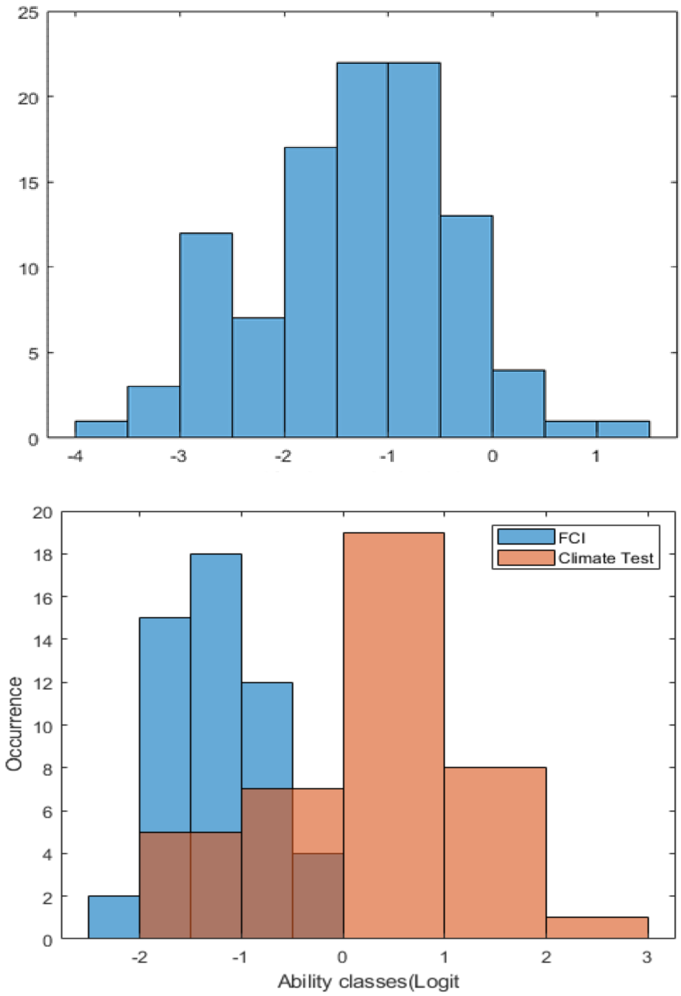

Figure 1. 10 Classes of the students' abilities (grades), upper figure, initial FCI test, 104 students. Lower figure, optimal classes for FCI and CCI conducted in the same group of 53.

So, the categories of knowledge on climate are distributed nearly symmetrical to a certain positive value (~0.5), whereas for FCI test, the abilities' classes are located on the negative part of the logit axis. This is an expected outcome if we consider that FCI test is basically a difficult and the displacement of the graphs would be the result of continuous adjusting of the centre during Rasch calibration, and the largeness of corresponding histograms mirror the nature of the knowledge dispersion among students. We observed that the spread of students' ability on FCI is almost the half of corresponding largeness of CCI ability histogram, Figure 2. This outcome mirrors the composition of the CCI test which contains very easy questions (information and fact), and some 'very difficult', even though we have removed 8 items that were perceived very hard. However, it resulted that students' abilities on the CCI test are more symmetrical than in FCI test, regardless of the bin size of the histogram. Referring the FCI as a certified instrument, those finding suggests reviewing the CCI and adding more questions, because the symmetrical shape of the corresponding distribution indicate for an accented homogeneity of the difficulties of the items of the CCI test. The effect of changing the number of questions is shown in Figure 2.

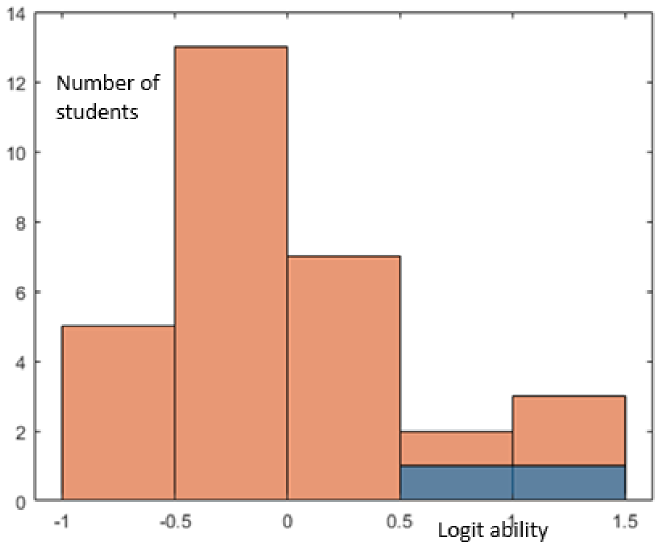

Figure 2.Histograms of the abilities for the test containing 18 questions and 22 questions. All misfitted events have been removed preliminarily.

The appearance of a skewness on the histogram when changing the number of questions suggest performing an investigation of this parameter in general. So, we observe that for CCI abilities, there are traces of skewed underline distribution in general, whereas for the FCI test the distribution looks symmetrical distribution, Figure 3. We relate this feature with specifics of the group of students that participate in the test, but a measurement in a larger sample is needed to avoid subjective conclusions. In principle, the skewness is related to the non-stationarity of the underlying distribution, which mirror the effect a couple of learning factors. Next, interesting features can be uncovered by analysing the deviances and abnormal events evidenced by Rash analysis. So, by checking the occurrences of the unexpected correct answering events $[P_{estimate}(i,j) < \frac{1}{2}, \cap T(i,j) = 1]$, we observed that a non-negligible part of students persist to answer CCI items by guessing, demonstrating low level of interest on completing a test.

Herein, we found a rate of guessing at 16% for the CCI test and around 18% for FCI test, which are high, and limit the accuracy of the measurement of conceptual knowledge by considering voluntary and participation without any benefit or other stimulating outcome. The guessing for FCI test is probability dictated from the difficulty of it items, but we were unable to identify in what extend the guessing comes from other reasons. On the other side, knowing that CCI test is significantly easier, the similar guessing rate observed for the CCI test indicate for a behavioural and intentional guessing. By crossing those findings together, we argued that the extra guessing observed in FCI must be related to the effect of common senses or pure random answers.

Despite statistical limitation of our surveys, the guessing rate observed for the CCI at 16% can be considered as an estimate of the limitation for similar survey accuracy, in the sense that if a test or survey isn't related to of any benefit or consequences, it is likely that at this level the tests' results would fail to measure the observable under scrutiny. From a more general consideration, those finding suggests that a non-negligible number of students behave as intellectually exhausted when undertaking a similar test (they might participate voluntary in the test, but they answered irresponsibly).

Next, we obtained that the difficulties of the CCI questions which were formulated purposedly to invite students to think before answering them, resulted at high values (in logit units). Notice also that the corresponding outfit and infit were also remarkable for those items. We suggested that those deviances mirror the insufficiency on illustrating corresponding phenomena during standard lecturing in high schools, which is expectable if we consider that all knowledge under analysis are provided through several disciplines.

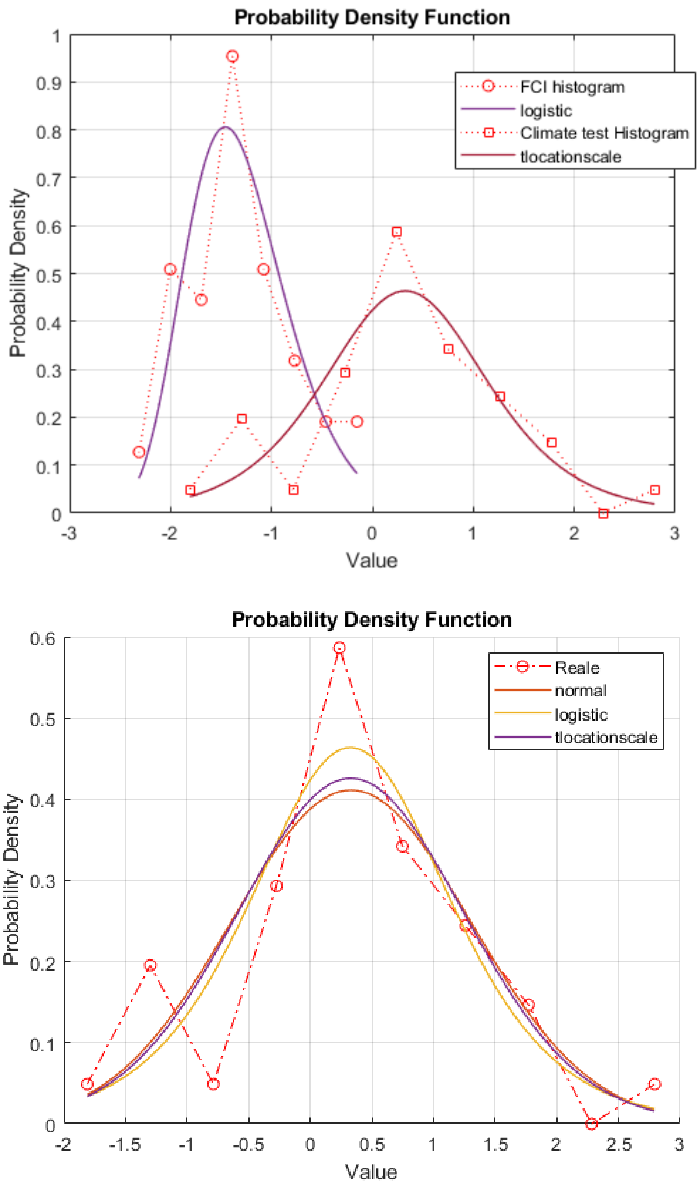

Figure 3. Traces of the underlying distribution of the students' knowledge on climate and environment.

However, the insufficiency of illustrating and demonstration in natural sciences in general should be taken under consideration by education care-taking agencies, regarding textbooks and classes-hours for those disciplines and specifically in physics. Next, two questions of the draft CCI based on very basic knowledge have been answered correctly despite that there was no detailed answer in the textbooks used in theory curricula, indicating that if adequate information is provided, students would adhere it, regardless of the sources. Another interesting finding consists to the structure of the difficulty threshold levels calculated for an ad-hoc polytomous items' test using the idea proposed in (Prenga, 2024). Notice that the parameter $\tau = [\tau_0, \tau_1, \tau_2, \tau_3, \dots \tau_n]$ in formulae (4) is interpreted as the threshold difficulty in basic Rasch model for polytomous variable, but in practice, it is difficult to realise a multi-scale answering CI tests.

To overcome this problem, we proposed to use our original CI test items to produce a polytomous-like variable. For this purpose, we have picked randomly 4 questions from the set of 18 items of the CCI test, we produced 10 'polytomous question' which by construction do have 5 values [0,1,2,3,4], aiming to mimic the level of knowledge identified herein. So, the assessment of a student regarding the 5-grade question is realised by adding up the scores obtained in the constituent items. By using (4) in the core algorithm of the Rasch model we evaluated the threshold vector for this fictitious test. The width of $w_{n,n-1} = \tau_n - \tau_{n-1}$ that evaluate the distance between the difficulty levels according to the polytomous Rash model, is recognised in this regard as an estimator of the learning performance or efficiency.

After performing such calculation and modelling we observed that corresponding difficulty level width $w_{n,n-1}$ do not change remarkable by the number of levels. It is unlike the corresponding finding for the FCI test and moreover, it differs significantly from recent estimation of this parameter using larger FCI test and referred in (Prenga, 2024). By commenting this feature of the difficulty threshold width, we conclude that teaching effort to rise the CCI knowledge from a certain level to the next one seems to be nearly constant. Again, the nature of the CCI test content has it weight in this behaviour, but also, we might presume that for basic information on our subjects, the efficacy of lecturing would be homogeneous for all levels. However, a robust analysis on this subject request a purposed and a carefully drafted CCI test, that can be realised by professors at the high school who are more familiarised with the matter.

We conclude this section by accepting that similarities and dissimilarities observed by employing the Rasch analysis for FCI and CCI tests can be used efficiently for a comparative analysis of teaching and learning on both subjects.

Conclusions

By performing the Rasch analysis for an ad-hoc test aiming on assessment of high schools' students' knowledge in climate and environmental topics (CCI test), we evidenced some similarities and dissimilarities with the standard FCI tests. Bearing in mind the different nature and structure of lecturing those subjects, we attribute similarities on the levels of knowledge to the increasing interest of the students for understanding the world around them, the appropriate auxiliary information provided by several sources, and to a relatively low level of conceptual understanding physics, resulting from several causes and factors.

Also, similarities on CCI knowledge parameters with FCI ones advocated that the gap on climate and environmental knowledge can be filled efficiently in the framework of teaching physics and other disciplines providing that the needed time space would be conferred. Despite the qualitative nature of those conclusions, we believe on their encouraging impetus for education researchers and scientists to conduct dedicated analysis of this category. We also recognise this finding as an indicator of evolving learning structure that suggests for a better harmonization of basic and fundamental knowledge in science with their common appearance materialised on the real life.

References

Andrich, D. (2005). The Rasch model explained. Applied Rasch measurement: A book of exemplars: Papers in honour of John P. Keeves, 27-59.

Crooks, N. M, & Alibali, M. W. (2014). Defining and measuring conceptual knowledge in mathematics. Developmental Review 34(4) 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.

Duka, K. 2022. Studimi i të kuptuarit koncpetual në fizikën newtoniane nëpërmjet testit FCI Teze Diplome, 2024. -The study of Conceptual Knowledge in Mechanics throughout FCI test, Master Degree thesis, 2022. Department of Physics, Faculty of Natural Sciences, University of Tirana.

Hamolli, L. & Prenga, D. (2024). Force concept inventory analysis by using indexes and Rasch model. Journal of Natural Sciences, 36, 210-230

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

Herzig, H. & Manaj, D. (2016). Gjeografi 10. Albas. ISBN 978-9928-02-747-4

Hill, G., Holman, J., Philippa G., Hulme, P. G. (2018). Kimia 12 me zgjedhje. (E. Andoni, M. Berisha, E.Tako, A.Hasimi, B.Guga, Trans.). Pegi. (Original book "Chemistry in context for Cambridge International AS and A Level" 2017). ISBN 978-9928-248-95-4

https://real-statistics.com

Jones, M. & Jones, G. (2016). Biologjia 10-11. (A.Trebicka, E.Hamzaraj, Trans.) Pegi. (Original book "Cambridge IGCSE Biology Coursebook" 2016) ISBN 978-9928-200-94-5.

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

Klymkowsky, M. W., & Garvin-Doxas, K. (2020). Concept inventories: Design, application, uses, limitations, and next steps. Active learning in college science: The case for evidence-based practice, 775-790.

Kumar, P., Sahani, J., Rawat, N., Debele, S., Tiwari, A., Emygdio, A. P. M., ... & Pfautsch, S. (2023). Using empirical science education in schools to improve climate change literacy. Renewable and Sustainable Energy Reviews, 178, 113232.

Kushta, E., & Prenga, D. (2023). Using Tests' Indexes to Improve the Assessment of the Conceptual Knowledge: a Case Study. International Journal of Education and Learning Systems, 8.

Nela, D. Vlerësimi i testeve të dijeve konceptuale në fizikë. Teze Diplome, 2024. -The assessment of CI test in physics, Master Degree thesis, 2024. Department of Physics, Faculty of Natural Sciences, University of Tirana.

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.

Pople, S. (2022). Fizika 10-11. (Z. Mulaj, B.Murtaj, Trans. and Adapt.). Erik botime. (Original book "Complete Physics for Cambridge IGCSE student book" 1999) ISBN 978-9928-130-40-2

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

Prenga, D. (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.

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, (Vol. 4, pp. 321-333). 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.

Sang, D., Jones, G., Chadha, G., Woodside, R. (2022). Fizika 12 me zgjedhje. (M. Ifti, Trans. and Adapt.). Mediaprint. (Original book "Cambridge International AS and A Level Physics Coursebook" 2014) ISBN 978-9928-08-339-5

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

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

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of educational measurement, 97-116.

Wright, B.D. & Masters, G.N. (1982). Rating Scale Analysis. Chicago: MESA Press.

www.physport.org/