PHYSICAL REVIEW SPECIAL TOPICS - PHYSICS EDUCATION RESEARCH **6**, 010103 (2010)

Rasch model based analysis of the Force Concept Inventory

Maja Planinic,* Lana Ivanjek, and Ana Susac

Department of Physics, Faculty of Science, University of Zagreb, Bijenicka 32, HR-10000 Zagreb, Croatia

(Received 20 May 2009; published 10 March 2010)

The Force Concept Inventory (FCI) is an important diagnostic instrument which is widely used in the field of physics education research. It is therefore very important to evaluate and monitor its functioning using different tools for statistical analysis. One of such tools is the stochastic Rasch model, which enables construction of linear measures for persons and items from raw test scores and which can provide important insight in the structure and functioning of the test (how item difficulties are distributed within the test, how well the items fit the model, and how well the items work together to define the underlying construct). The data for the Rasch analysis come from the large-scale research conducted in 2006-07, which investigated Croatian high school students' conceptual understanding of mechanics on a representative sample of 1676 students (age 17–18 years). The instrument used in research was the FCI. The average FCI score for the whole sample was found to be $(27.7 \pm 0.4)\%$, indicating that most of the students were still non-Newtonians at the end of high school, despite the fact that physics is a compulsory subject in Croatian schools. The large set of obtained data was analyzed with the Rasch measurement computer software WINSTEPS 3.66. Since the FCI is routinely used as pretest and post-test on two very different types of population (non-Newtonian and predominantly Newtonian), an additional predominantly Newtonian sample ($N=141$, average FCI score of 64.5%) of first year students enrolled in introductory physics course at University of Zagreb was also analyzed. The Rasch model based analysis suggests that the FCI has succeeded in defining a sufficiently unidimensional construct for each population. The analysis of fit of data to the model found no grossly misfitting items which would degrade measurement. Some items with larger misfit and items with significantly different difficulties in the two samples of students do require further examination. The analysis revealed some problems with item distribution in the FCI and suggested that the FCI may function differently in non-Newtonian and predominantly Newtonian population. Some possible improvements of the test are suggested.

DOI: 10.1103/PhysRevSTPER.6.010103 PACS number(s): 01.40.Fk

I. INTRODUCTION

The Force Concept Inventory (FCI) [1] is a well known and widely used conceptual test in mechanics. It was constructed on the basis of the findings of physics education research on student alternative ideas in mechanics and it is used to diagnose the prevalence of those ideas in different groups of students. This resulted in a wide use of the test, first in the USA (e.g., Refs. [2,3]) and then also in many other countries (e.g., Ref. [4]) throughout the world. The FCI is a multiple choice test of 30 questions that investigates student conceptual understanding of the Newtonian force concept with minimal use of mathematics. This test has a large impact on many physics teachers throughout the world since it demonstrated very clearly that students hold non-Newtonian ideas about force and motion both before and after instruction on Newtonian mechanics. The advantage of the test is that it can be easily administered to large samples of students, thus making its results even more shocking and significant. Over time the FCI has acquired a status of a standardized instrument for measurement of student conceptual understanding of mechanics. High FCI score is considered a strong, although not perfect, indicator of Newtonian thinking in students (not perfect since Newtonian physics requires more than just the ability to recognize one Newtonian among four non-Newtonian answers). However, low FCI score undoubtedly indicates poor conceptual understanding of mechanics. The FCI authors have suggested that the score of 60% can be considered a threshold for the development of Newtonian thinking [1,3]. Below that threshold student understanding of Newtonian concepts is insufficient for effective problem solving and such students have difficulties following physics courses at university level [3].

Since the FCI was first published it has been used in many physics courses throughout the world as a standard instrument for the assessment of student conceptual understanding of basic mechanics. Students were typically tested before and after instruction (pretesting and post-testing). Hake found in his large-scale study [2] that the average FCI student gains in a physics course could be correlated with the type of instruction implemented in the course. He defined normalized gain as $g=(x_{post}-x_{pre})/(100-x_{pre})$, where x_{pre} and x_{post} are pretest and post-test class averages in percent. Hake found that the highest gains were associated with interactive engagement courses, while all traditional courses achieved $g<0.3$. This implied that the FCI could also be used as a measure of instruction efficacy in promoting conceptual understanding of mechanics.

The FCI has proven to be an important instrument in physics education research and served as the model for the development of conceptual tests in other physics domains. Student FCI scores are frequently used as measures of student conceptual understanding of mechanics. However, it is important to realize that the meaning of those scores depends strongly on the structure and functioning of the test, which should therefore be investigated in great detail. Structure of

*Corresponding author; maja@phy.hr

the test refers to the distribution of items according to their difficulties. How test functions can be evaluated through the analysis of item fit to the model and analysis of how well the items work together to define the underlying construct. Since FCI is typically used on non-Newtonian and predominantly Newtonian populations of students its functioning should also be evaluated on these two kinds of populations. Some of the existing research on the FCI touches on those issues [5–13], but it does not give the complete picture of the FCI as a measurement instrument.

The Force Concept Inventory was constructed as an improved version of the mechanics diagnostics test for which classical test validity and reliability had been established [5]. Most items in the FCI were at least slightly changed over the years, several questions were completely removed, and several were added. The latest version of the test (which was also used in this study) is available on the web [6]. There were attempts to analyze the FCI with the factor analysis [7], in which no significant factors were identified. This finding provoked a debate in the physics education research (PER) community about what the FCI really measures [3,8]. One study investigated the influence of context on FCI items [9] and found that influence is not sufficient to affect normal use of the FCI as a diagnostic instrument. Another study investigated the quality of distracters in three FCI items (4, 9, and 11) using item response theory [10]. Item 4 was found to be inefficient, item 9 of medium efficiency, and item 11 efficient in discriminating students by their ability. In another study the scores on the FCI were compared with the scores on the force and motion conceptual evaluation [11]. The study found generally positive correlation between student scores on the two tests and also pointed to some discrepancies between them. Other studies found positive correlation between FCI gains and Scholastic Aptitude Test scores [12] as well as positive correlation between FCI gains and scores on Lawson's classroom test of scientific reasoning [13].

Since the FCI is undoubtedly one of the most widely used assessment instruments in physics education research it would be important to evaluate and monitor its functioning using various tools for test analysis. One of such tools is the stochastic Rasch model [14], which can provide important insight in the structure and functioning of tests.

For some years the physics education group at University of Zagreb has been conducting the FCI pretesting of the first year students at Faculty of Science. It was found that many of the students entering general physics courses still predominantly used non-Newtonian ideas in mechanics despite previous six years of physics instruction at school. This provided the motivation for a large-scale study which was undertaken by our group in 2006-07 [15]. In order to estimate the average level of conceptual understanding of mechanics in the population of Croatian students at the end of gymnasium (a type of Croatian high school which prepares students for universities), a representative sample of students was tested with the FCI. This has provided a large amount of data ($N=1676$) that was later analyzed with the Rasch model.

Additionally, a group of 141 first year students from the Faculty of Electrical Engineering and Computing at University of Zagreb, Croatia was also tested next year with the FCI. Their results are also included in this study.

The school system in Croatia consists of eight years of elementary school followed by four years of high school. After high school students can continue their education at different universities and colleges. There are several types of high schools in Croatia, but in this study we have focused on one type called gymnasium. Gymnasium graduates typically continue their education at university level. Gymnasiums can be of the general education type or they can specialize either in foreign or classical languages or in natural sciences and mathematics.

Physics is taught from the seventh grade of elementary school (age 12–13 years) until the fourth year of gymnasium (age 17–18 years) as a separate and compulsory school subject. In the seventh and eighth grade of elementary school students have two 45-min physics lessons per week. In gymnasium the number of physics lessons per week depends on the type of school. In the general education (GE) type schools and those which specialize in foreign or classical languages (FL or CL) students have two 45-min physics lessons per week throughout four years of schooling. In schools which specialize in natural sciences and mathematics (NSM) students have three 45-min physics lessons per week throughout four years of schooling. Mechanics is taught during the whole first year (age 14–15 years) in all types of schools.

II. DATA COLLECTION AND ANALYSIS

The population of students in the last (fourth) year of gymnasium had 12 366 students in 2006-07. Most of them (84.5%) were in the GE and FL or CL schools and only 15.5% were in the NSM schools. The sample of students in the study included 1676 students or 13.6% of the population. In the sample there were 429 students (25.6%) from the NSM and 1247 students (74.4%) from the GE and FL or CL schools. The sample represented proportionally different regions of the country. The number of participating schools was 54 which make 36% of all gymnasia in the country. Participating schools were chosen randomly from the list of all existing schools in a particular region. The principals of chosen schools were contacted, informed about the research, and asked for the permission to test their students. Almost all principals allowed the testing of their students. On the date of the testing that was agreed upon with the principal, one of the researchers would come to the school, supervise the testing, and collect the tests and answer sheets.

Students were tested in the period from October 2006 until February 2007, without any special preparation for the test. The latest improved version of the FCI available at the web at the time of the testing [6] was used. The test was carefully translated in Croatian.

The testing was anonymous, but each student was assigned a code so that they would later be able to identify their test scores. Students had to mark their answers on a special answer sheet. The allocated time for taking the test was 45 min. Participating schools were later informed about students' test scores and were asked to pass that information on to their students. No incentives, such as grades, were offered to students for taking the test. However, the purpose

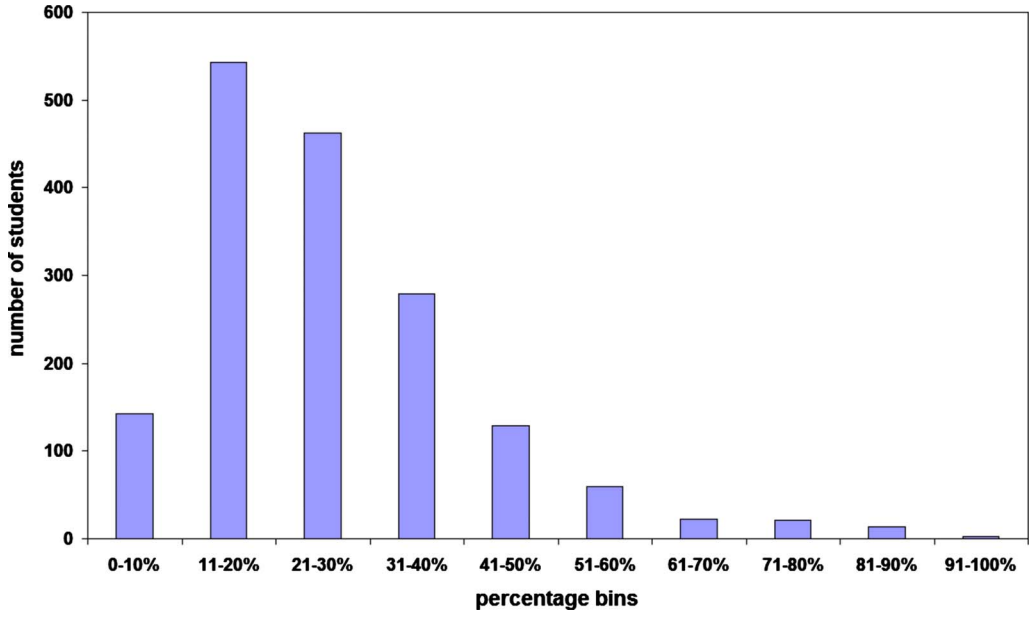

FIG. 1. (Color) Distribution of test scores for gymnasium students.

of the research and the importance of the test have been explained to them before the testing by the researcher present at the testing. Students generally showed interest in the test and wanted to know their score on it.

All students studied mechanics in the first year of gymnasium (age 14–15 years) and the testing was done in the fourth year, with the gap of 2.5 years between the instruction and the FCI testing. The gap between instruction on mechanics and testing is large, but during that time the students have been learning other physics topics which relied on Newtonian concepts. These concepts therefore should have been in use by students over the years since they studied mechanics and student understanding of those concepts should even have been refined through their application in other contexts.

A group of 141 first year students enrolled in introductory physics course at the Faculty of Electrical Engineering and Computing at University of Zagreb, Croatia was also tested in 2008 in the same way as gymnasium students. The students were tested after the instruction on mechanics.

After the classical analysis based on percentages of correct answers, both sets of data (gymnasium students and university students) were also analyzed using WINSTEPS 3.66 software [16] for Rasch analysis. Rasch analysis is a type of logistic regression analysis which can also be performed with generalized linear models and software which is based on those models [17].

III. RESULTS

A. Classical analysis

Distribution of scores for the whole sample of gymnasium students is shown in Fig. 1. The statistical information (arithmetic mean, median, standard deviation σ, and standard error of the mean $\sigma/\sqrt{N}$ for each distribution) for all groups of students is listed in Table I.

Figure 1 suggests that at the end of gymnasium students are still mostly non-Newtonians, as measured by the FCI standard. The average FCI score of the whole sample, as well as the scores of the two sample subgroups, is well below the threshold of 60%. We have found 4.7% of students in the sample with scores above 60%.

As expected, the scores of the students in the NSM schools are better than in the GE and FL or CL schools. Students who score above 60% come almost exclusively from the NSM schools.

Distribution of scores for university students is shown in Fig. 2. It is evident that, contrary to gymnasium students, this

TABLE I. Statistical information for all groups of students (NSM stands for gymnasiums which specialize in natural sciences and mathematics, GE for gymnasiums of general education type, and FL or CL for gymnasiums which specialize in foreign or classical languages).

	N	Arithmetic mean (%)	Median (%)	Standard deviation (%)	Standard error of the mean (%)
GE, FL or CL, and NSM	1676	27.7	23.3	15.2	0.4
GE and FL or CL	1247	24.8	23.3	12.3	0.3
NSM	429	36.2	33.3	19.2	0.9
University	141	64.5	63.3	20.8	1.8

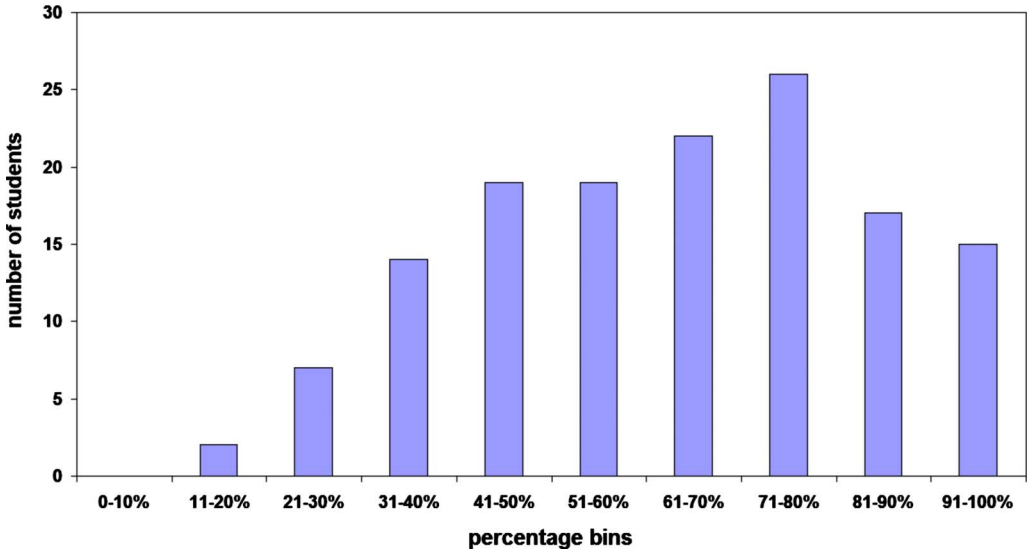

FIG. 2. (Color) Distribution of test scores for university students.

group of university students, with the average score of 64.5%, is predominantly Newtonian, with about two thirds of the students at or above the Newtonian threshold of 60%.

B. Rasch analysis for the gymnasium sample

The Rasch model [14] is a mathematical model developed by the Danish mathematician Georg Rasch around 1960 in an effort to bring the measurement in social sciences closer to the standards of measurement in physics. As a means of test analysis, Rasch measurement parallels physical measurement processes by being largely concerned with the construction of linear measures along specific unidimensional constructs. The important characteristic of the Rasch model is that it allows users to create an interval scale of scores for both the item and person measures.

The first requirement is that the variable to be measured with a test (e.g., understanding of the Newtonian force concept) is specified and described by a set of well chosen test items. The model assumes that the test is unidimensional, meaning that each item probes only the measured variable and not something else. The unidimensionality requirement is, of course, only an approximation since completely unidimensional test is almost impossible to construct but it can be realized well enough for practical purposes. The location of a person along the measured variable is described by a measure called person ability, which gives the information about the intensity of the measured variable that the person possesses. Each item in the test should require different intensities of the variable in question. Items are described by another measure called item difficulty. A good test should contain items of various difficulties, more or less equally spaced along the whole range of abilities of the targeting sample.

What happens when a person of ability B_n meets an item of difficulty D_i is determined by a probabilistic relation between the two measures. The probability P_{ni} of a correct answer of person n to item i (the Rasch model for dichotomous items) is given as [18,19]

$$P_{ni} = \frac{e^{B_n - D_i}}{1 + e^{B_n - D_i}}.$$

The probability of the correct answer is governed by the difference $B_n - D_i$. If person ability equals item difficulty ($B_n = D_i$) the probability of the correct answer is 0.5. If person ability greatly exceeds item difficulty ($B_n \gg D_i$) the probability goes to 1 and in the opposite case ($B_n \ll D_i$) the probability becomes 0.

Item difficulties and person abilities are calculated from raw test scores for items (from the number of correct answers to an item) and persons (from the number of correct answers given by a person). The construction of measures, which is performed by WINSTEPS or other Rasch model software, starts from the estimation of person ability B_n and item difficulty D_i. The first step in estimating B_n is the conversion of the raw score fraction of correct answers (p) into log odds $\ln[p/(1-p)]$ (odds being success-to-failure ratio). To estimate item difficulty D_i the same procedure is applied to the fraction of students who answered the item correctly. The model defines the unit of measurement called logit (log odds unit) in which both the item difficulties and person abilities are measured. The obtained estimated measures are expressed on the logit scale with the average item measure arbitrarily set at 0. The estimates are then corrected for effects of variance and iterated against each other until they meet a preset convergence criterion and give a set of internally consistent item and person parameters. For more detailed information about the Rasch model see, for example, Refs. [18,19].

The measures are linear, which is another very important characteristic of the Rasch model. For example, a person with the ability of 3 logit has three times more ability than a person with the ability of 1 logit. This is obviously very different from scores expressed as percentages, where it is impossible to say that a person who scores 30% on a test has three times more ability than a person who scores 10% on the same test. Percentages can reflect the correct ranking of persons or items but not the correct intervals between their

abilities or difficulties. Therefore, percentages are not linear in the variable which they represent [18,19]. The linearity of measures, on the other hand, is very important because meaningful arithmetic operations can only be performed with linear measures, thus enabling comparisons and statistical studies.

Each item and person measure comes with its Rasch standard error which indicates the uncertainty of the estimate. The estimates are more precise if the number of persons and items is large and if there is good targeting of the test on the distribution of students. Items and persons on the edges of their distributions will typically have larger uncertainties than central items and persons.

When item and person calibrations are obtained they are placed on a vertical ruler (Fig. 3) that measures person ability and item difficulty on the same logit scale. On the right-hand side of the ruler are the FCI items sorted by difficulty, with the most difficult items on the top and the easiest items on the bottom of the plot. On the left-hand side of the ruler are persons, sorted by their abilities (success on the FCI), with the most successful students on the top. It is obvious from Fig. 3 that the test was very difficult for the students since the distributions of item difficulties and of person abilities are significantly shifted with respect to each other. The mean item difficulty is 1.26 logit above the mean person ability. Ideally, the test should be centered on the target population. This plot also immediately shows the ordering of items according to their difficulty. Items with negative calibrations are easier, and those with positive calibrations are more difficult than the item average whose difficulty is set at zero. The spacing between items is also very important. Items should not be too close in difficulty because otherwise one item is not distinctly separate from the next. But the separation between individual items should also not be too large to avoid large gaps between the items.

Inspection of Fig. 3 reveals that the width of the test is about 4 logit, whereas the width of the person distribution is almost 8 logit. About two thirds of the items are in the region between −1 logit and +1 logit, but only approximately one third of all persons can be found in this range. Many items in the middle of the test are of very similar difficulty, but there are fewer items in the very easy and the very hard regions of the test. For this sample of students there are enough hard items but there are not enough easy items.

Once the abilities and difficulties are estimated, WINSTEPS calculates the theoretical probabilities for the success of each person on each item and compares them with the observed scores. The differences between the two are called residuals, and they are used to evaluate the fit of data to the model [19]. Rasch analysis programs usually report fit statistics as two chi-square ratios: infit and outfit mean square statistics. Outfit is based on the conventional averaged sum of squared standardized residuals, whereas infit is an information—weighted sum which gives more value to on-target observation. Large infit value on a particular item indicates that some persons of the ability which is close to the difficulty of the item have not responded in the way consistent with the model. Large outfit value of an item indicates that persons who are far in ability from the difficulty of the item have responded in an unexpected way. For example, large outfit of an easy item means that some able students have unexpectedly failed on this item. Large outfit of a hard item means that some students of low ability have unexpectedly succeeded on that item. Large infit values are generally considered more problematic than large outfit values. In this study observed outfit values are larger than infit values, so only outfit analyses are presented.

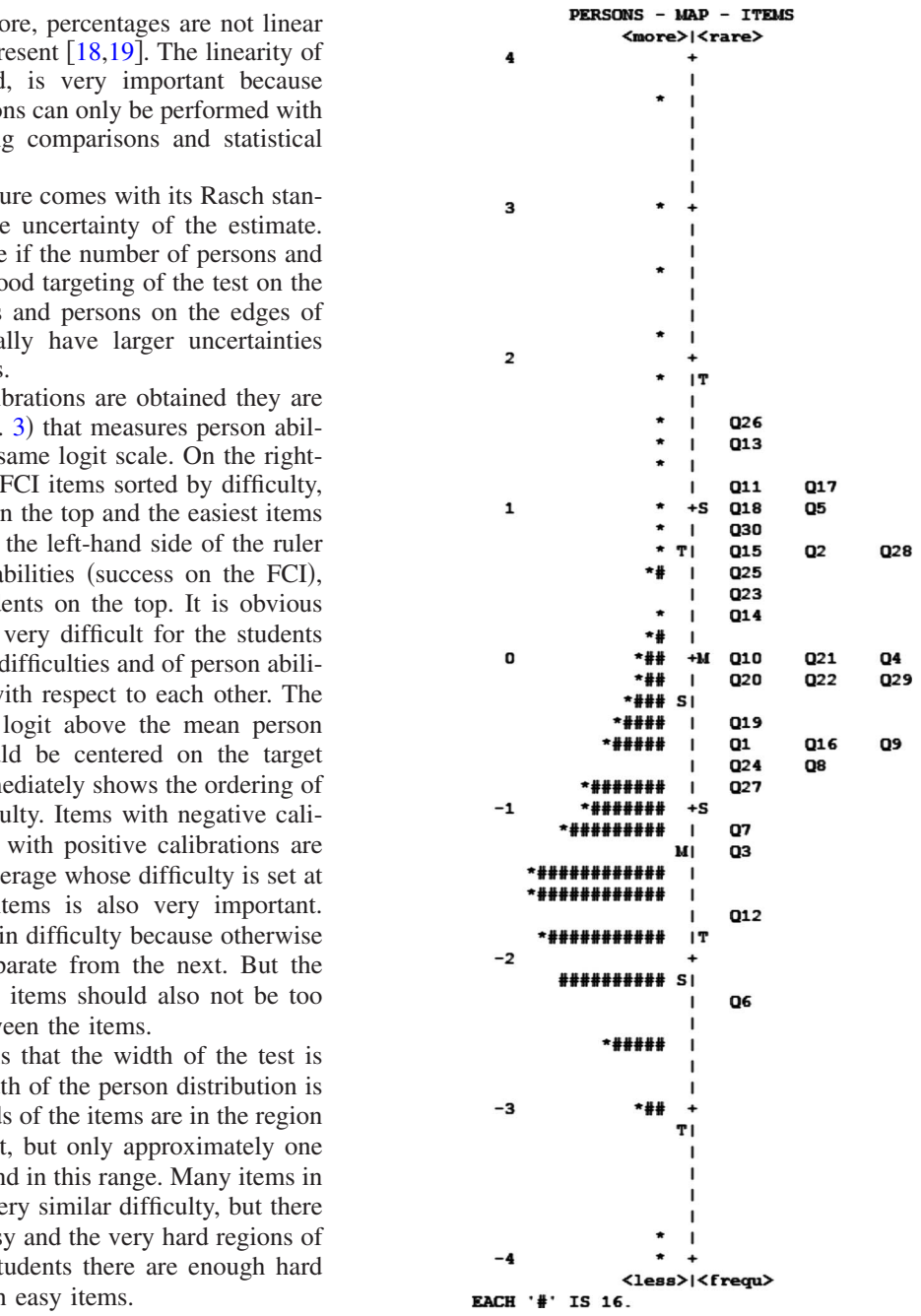

FIG. 3. Item-person map for gymnasium students. The left-hand side shows distribution of student abilities and the right-hand side shows the distribution of item difficulties. Items are labeled as Q1–Q30. M, S, and T are labels for the mean value, one standard deviation, and two standard deviations of each distribution. Each # represents 16 students and each * less than 16 students.

The expected value of both infit and outfit is 1. Items which are sufficiently in accordance with the Rasch model to

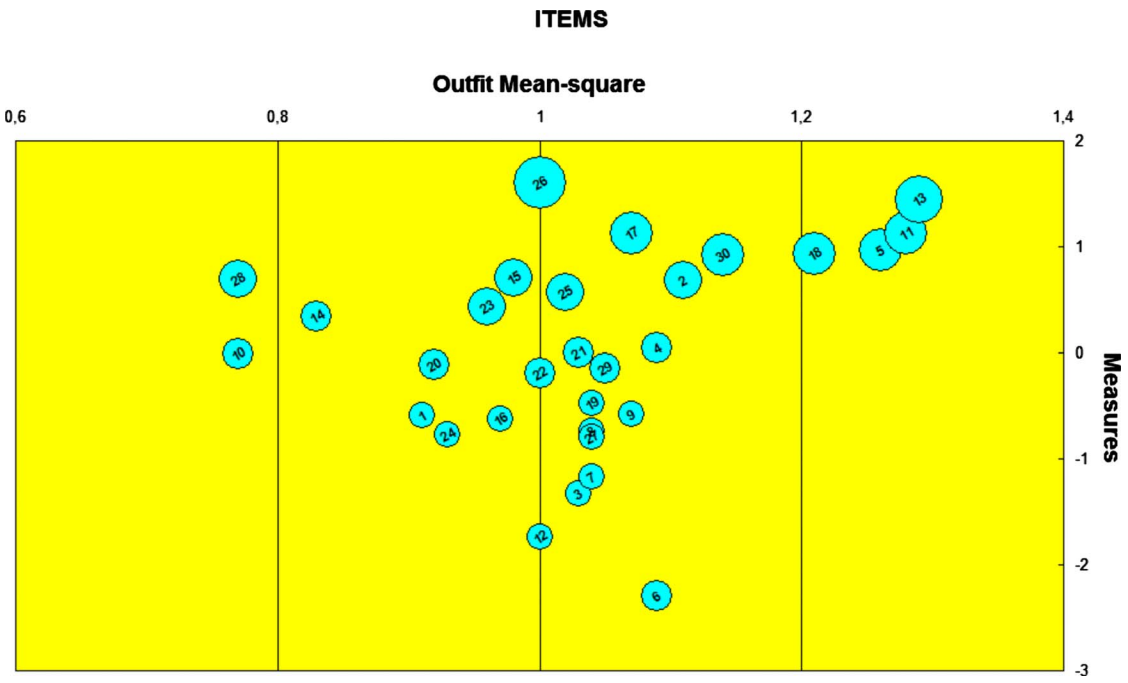

FIG. 4. (Color) The bubble chart for the gymnasium students showing outfit mean square statistics (MNSQ) vs item measure.

be productive for measurement will have infit and outfit values between 0.5 and 1.5 [20]. Items with significant infit or outfit (1.2–1.5) should however be inspected more closely to find out reasons for their misfit. When the data fit well they indicate that the subscale items all contribute to a single underlying construct, but largely misfitting items do not contribute to the underlying construct, either because they are badly formulated or because they measure something different than the rest of the items.

To inspect closer the structure of the test a bubble chart [19] of items is presented in Fig. 4. Each item is represented with a circle, whose size is proportional to the Rasch standard error of item's calibration. Smaller circles represent items with smaller uncertainty of calibration. It is useful to think of items in a bubble chart as "stepping stones" [19] which define the direction of the underlying variable. In a well constructed test circles on the bubble chart are clearly separated but not with very large gaps. The horizontal distance of circles from the expected outfit value of 1 indicates how well each item fits to the model. Ideally, items should be close to the central axis of the bubble chart.

Inspection of Fig. 4 reveals that some circles overlap, and many are very close in difficulty, thus making the ordering of the items unclear. It is also noticeable that the circles become larger for harder items on the top of the diagram because these calibrations were estimated from a smaller number of responses.

Items which are far away from the expected value of 1 are either items located at the left ends of bubble chart (items 10 and 28), which show more regular answer pattern than predicted by the model, or items located at the right end of the chart (items 5, 11, 13, and 18), which show too unpredictable answer pattern. In deterministic models one expects a very regular pattern of answers: student should succeed on all items of difficulties below their ability and fail on all those which are above their ability. Being a stochastic model, the Rasch model expects a certain level of irregularity in the data. Too much regularity in students' answer pattern to an item can be a sign of dependency or redundancy in the item. However, items with too unpredictable answer patterns are usually considered as much bigger threat for the validity of measurement than too deterministic items. Since items 5, 11, 13, and 18 are all hard items whose moderately large outfit (less than 1.3) is probably caused by lucky guesses of some low ability students this is not a sign of a serious problem for the measurement.

The concept of unidimensionality is very important for the Rasch model. All items are expected to work together and define a single underlying construct. The content of the items can be considered the empirical definition of this construct [20]. With the help of item point-measure correlations (Table II) it can be checked if the construct is present in the test. The point-measure correlation of an item is the correlation between the Rasch person ability measures and persons' responses to the item [20]. WINSTEPS computes these correlations as Pearson product-moment correlation coefficients [20]. We want all items to point in the same direction; therefore their correlations should all be positive. The size of correlations can indicate which items contribute more to the construct and which contribute less. Table II shows that the FCI defines an underlying construct on the gymnasium sample—all correlations are positive, although generally not very large (due to poor targeting of the test on the sample). Figure 4 shows that there are no grossly misfitting items.

The scale that perfectly fits the Rasch model is unidimensional, has adequate separation, has items that are not calibrated too far apart, and has individual items that all contribute to the underlying construct. For the FCI on the gymnasium sample it can be concluded that the unidimensionality requirement has been realized sufficiently well and

TABLE II. Statistical information for items for the Rasch analysis of the gymnasium sample. Displayed are total raw score, item measure in logit, Rasch standard error, infit and outfit MNSQ statistics, and point-measure correlation for each item.

Entry No.	Total score	Measure	Rasch S.E.	Infit MNSQ	Outfit MNSQ	Correlation
1	606	−0.59	0.05	0.96	0.91	0.43
2	261	0.68	0.07	1.07	1.11	0.29
3	872	−1.33	0.05	0.99	1.03	0.38
4	409	0.05	0.06	1.08	1.09	0.31
5	208	0.97	0.08	1.07	1.26	0.25
6	1206	−2.29	0.06	1.05	1.09	0.28
7	812	−1.17	0.05	1.02	1.04	0.35
8	656	−0.74	0.05	1.02	1.04	0.36
9	601	−0.58	0.05	1.07	1.07	0.32
10	424	−0.01	0.06	0.84	0.77	0.53
11	183	1.13	0.08	1.05	1.28	0.25
12	1023	−1.74	0.05	0.97	1.00	0.38
13	143	1.44	0.09	0.99	1.29	0.27
14	334	0.34	0.06	0.90	0.83	0.46
15	256	0.70	0.07	1.03	0.98	0.33
16	619	−0.63	0.05	0.96	0.97	0.42
17	183	1.13	0.08	0.97	1.07	0.33
18	215	0.93	0.08	1.09	1.21	0.24
19	568	−0.48	0.05	1.03	1.04	0.36
20	453	−0.11	0.06	0.94	0.92	0.43
21	422	−0.00	0.06	1.04	1.03	0.34
22	475	−0.19	0.06	0.99	1.00	0.38
23	313	0.43	0.07	0.97	0.96	0.39
24	668	−0.77	0.05	0.94	0.93	0.44
25	283	0.57	0.07	0.96	1.02	0.38
26	125	1.60	0.10	0.91	1.00	0.37
27	678	−0.80	0.05	1.04	1.04	0.35
28	258	0.69	0.07	0.90	0.77	0.46
29	465	−0.15	0.06	1.04	1.05	0.34
30	216	0.92	0.08	0.98	1.14	0.33

that all items work together, but several problems are noticeable: poor targeting of the test on the population, too small separation of items in the middle of the test, too small width of the test, and the lack of easy items.

C. Rasch analysis for the university sample

The FCI is usually used as pretest and post-test in introductory physics courses. The pretest population is in most cases predominantly non-Newtonian with low average scores on the FCI. The gymnasium sample in this study is an example of such population. However, after the instruction, the post-test population is expected to become predominantly Newtonian if the instruction is efficient in promoting conceptual understanding. It is therefore important to see how the FCI functions on both kinds of populations (non-Newtonian and predominantly Newtonian). To check the functioning of the FCI on the predominantly Newtonian sample of students we have performed Rasch analysis of the FCI post-test scores of 141 first year university students who scored an average of 64.5% on the test.

The functioning of the test on this population can be analyzed starting again from the item-person map (Fig. 5). The test is relatively easy for this population, with the mean ability of the sample being 0.93 logit above the mean difficulty of the items in the test. Table III shows that all correlations are again positive, but this time larger than in the case of the non-Newtonian population, due to the better targeting of the test on the sample. The distribution of student abilities is about 6 logit wide in this sample, while the distribution of item difficulties is about 3.5 logit wide. Here the lack of hard items is quite apparent, while most of the items are found in the middle of the test. How items contribute to the measurement of the underlying construct can be judged from Fig. 6 which shows the outfit bubble chart for the university sample. We can see from Fig. 6 that items 7, 18, and 29 have

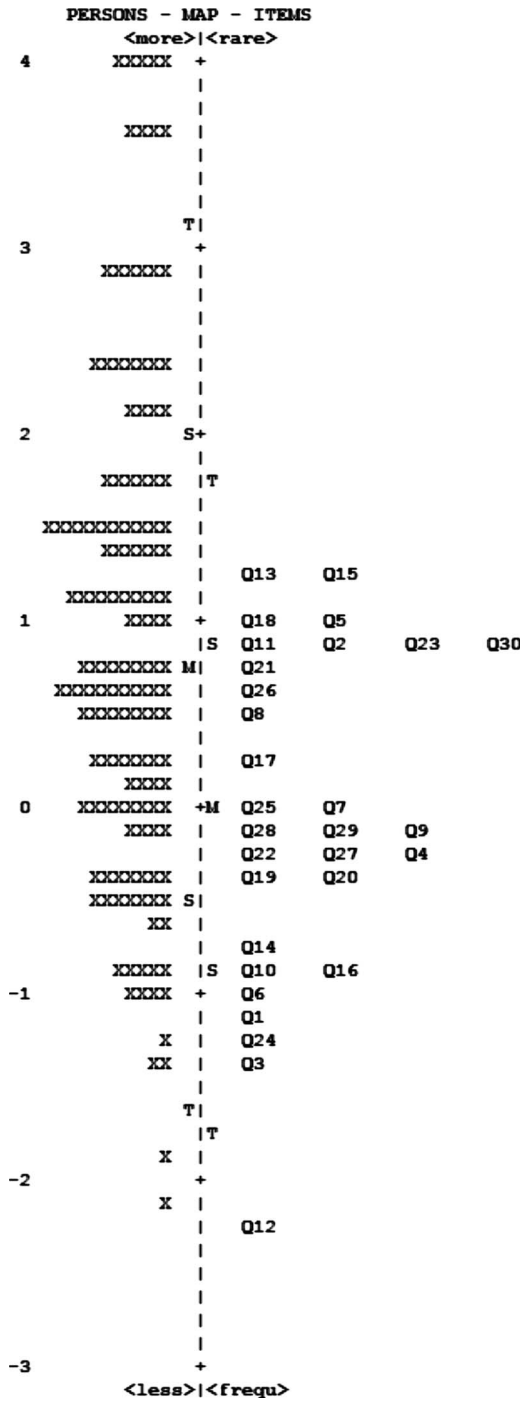

FIG. 5. Item-person map for the university students. The left-hand side shows the distribution of student abilities and the right-hand side shows the distribution of item difficulties. Items are labeled as Q1–Q30. M, S, and T are labels for the mean value, one standard deviation, and two standard deviations of each distribution. Each "x" represents one student.

the largest outfits, while items 18 and 29 in addition also have significant infits (Table III).

In item 18 a boy is shown swinging on a rope, moving toward equilibrium position. Students have to decide which combination of the four suggested forces (a downward force of gravity, a force exerted by the rope pointing toward the center of rotation, a force in the direction of the boy's motion, and a force pointing away from the center of rotation) acts on the boy. All wrong answers include the force in the direction of motion. The significant infit value of item 18 implies that some students of high ability, which means students who are predominantly Newtonians, have unexpectedly failed on this item and included in their answers the force in the direction of the boy's motion. That is surprising because other FCI items in which students typically express similar alternative ideas, such as items 13, 17, and 30, have too small infit and outfit, meaning that they are even too regular in discriminating between the students of low and high abilities. So why do some Newtonians fail on item 18? A possible reason could be that good students notice that—unlike in items 13, 17, and 30—in item 18 there is actually a force in the direction of motion, a component of the gravitational force. This might be a source of confusion for some students. Item 18 should be further investigated and maybe reformulated.

In item 29 students are asked which combination of the three suggested forces (a downward force of gravity, an upward force exerted by the floor, and the net downward force of the air) acts on the chair at rest. It seems that the net downward force of the air was confusing for students. Some otherwise successful students unexpectedly failed on this item because they have included this force in their answers. However, this item really does not test student understanding of Newton's laws but rather their understanding of the effects of atmospheric pressure on the bodies in air. It seems that this item does not test the same variable as the rest of the items. Although item 29 does not degrade measurement, the FCI would probably be a more coherent instrument if this item was excluded from the test.

Another problematic item is item 7. Item 7 basically tests the same thing as item 6 (student understanding of kinematics of circular motion), asking students to predict the trajectory of the object in circular motion after restraints which enable circular motion are removed (ball exits from the horizontal frictionless circular channel in item 6 or stone on a string moving horizontally in a circular path flies off after the string breaks in item 7). One would expect similar results on both questions, however item 7 is 1.03 logit more difficult than item 6. The relatively large outfit value of item 7 was caused by a number of otherwise successful students who have failed on this item but usually succeeded on item 6. It is possible that in item 7 students were confused because they knew that the trajectory of the stone after the string breaks will not be a straight but a curved line. Some students asked researchers during the testing what the figure in item 7 represented. There are indications that the situation described in item 7, as well as its distracters, was not clear to all students.

From comparisons of item positions in Figs. 3 and 5 and item measures from Tables II and III it is noticeable that several items (items 6–8, 10, 14, 17, 21, 26, and 28) significantly change their difficulty (by more than three standard errors) from non-Newtonian to Newtonian sample, therefore making the item difficulty order different in the two analyses. Since clearly distinguishable item difficulty levels are at least three standard errors apart [19], items which change their position by more than three standard errors should be care-

TABLE III. Statistical information for items for the Rasch analysis of the university sample. Displayed are total raw score, item measure in logit, Rasch standard error, infit and outfit MNSQ statistics, and point-measure correlation for each item.

Entry No.	Total score	Measure	Rasch S.E.	Infit MNSQ	Outfit MNSQ	Correlation
1	117	−1.14	0.24	1.01	0.86	0.35
2	70	0.82	0.19	0.87	0.84	0.57
3	123	−1.33	0.25	0.96	1.10	0.24
4	98	−0.24	0.20	1.05	1.05	0.39
5	59	0.97	0.19	0.99	1.04	0.50
6	115	−1.03	0.23	1.01	0.83	0.37
7	92	0.00	0.19	1.13	1.33	0.35
8	78	0.52	0.19	1.09	1.11	0.42
9	95	−0.12	0.20	1.10	1.18	0.36
10	111	−0.82	0.22	0.85	0.67	0.48
11	67	0.93	0.19	0.96	0.93	0.53
12	131	−2.23	0.33	0.99	0.73	0.26
13	58	1.27	0.19	0.80	0.75	0.63
14	109	−0.73	0.21	1.07	0.90	0.36
15	58	1.27	0.19	1.10	1.16	0.45
16	113	−0.93	0.22	0.94	0.96	0.39
17	85	0.26	0.19	0.96	0.86	0.50
18	66	0.96	0.19	1.20	1.39	0.37
19	100	−0.33	0.20	1.12	1.22	0.34
20	102	−0.41	0.20	0.99	1.20	0.39
21	72	0.74	0.19	1.15	1.17	0.40
22	98	−0.19	0.20	0.96	0.83	0.46
23	67	0.93	0.19	1.10	1.10	0.44
24	119	−1.26	0.24	0.86	0.59	0.44
25	93	−0.04	0.19	0.84	0.73	0.55
26	75	0.63	0.19	0.81	0.72	0.61
27	98	−0.24	0.20	1.10	1.08	0.37
28	96	−0.16	0.20	0.77	0.62	0.58
29	95	−0.12	0.20	1.18	1.34	0.31
30	68	0.89	0.19	0.83	0.75	0.60

fully examined. Of those items special attention deserve items 6–8 which show the largest displacements from pretest to post-test and which become significantly more difficult (relative to other items) on post-test compared to pretest.

This analysis suggests that the FCI may function differently on non-Newtonian and predominantly Newtonian populations. There seems to be a difference in the FCI construct in these two populations (item order does not remain the same). This could be an issue for the use of the FCI as pretest and post-test, but it should be further investigated on other student samples.

IV. CONCLUSIONS

The aim of the large-scale research undertaken in Croatia was to estimate the level of conceptual understanding of mechanics in the population of Croatian students in the final year of gymnasium. It can be concluded that the large part of the Croatian gymnasium students are still non-Newtonians when they finish high school and enter universities.

Similar problems are found in other countries as well (e.g., US-A [2,3] or Germany [4]). It appears that it is not easy to achieve high FCI scores, especially in general student population. Even though we have not attempted in our study to establish a link between the type of instruction and the FCI scores, we have come across anecdotal evidence that where interactive teaching methods are used scores are higher than the average, but to confirm that link further research would be required.

The Rasch model based analysis of the FCI provided some important insights in the structure and functioning of the test on two different kinds of populations: non-Newtonian and predominantly Newtonian. Generally it can be concluded that the test has succeeded in defining a sufficiently unidimensional construct for each population. The items in the test all work together and there are no grossly

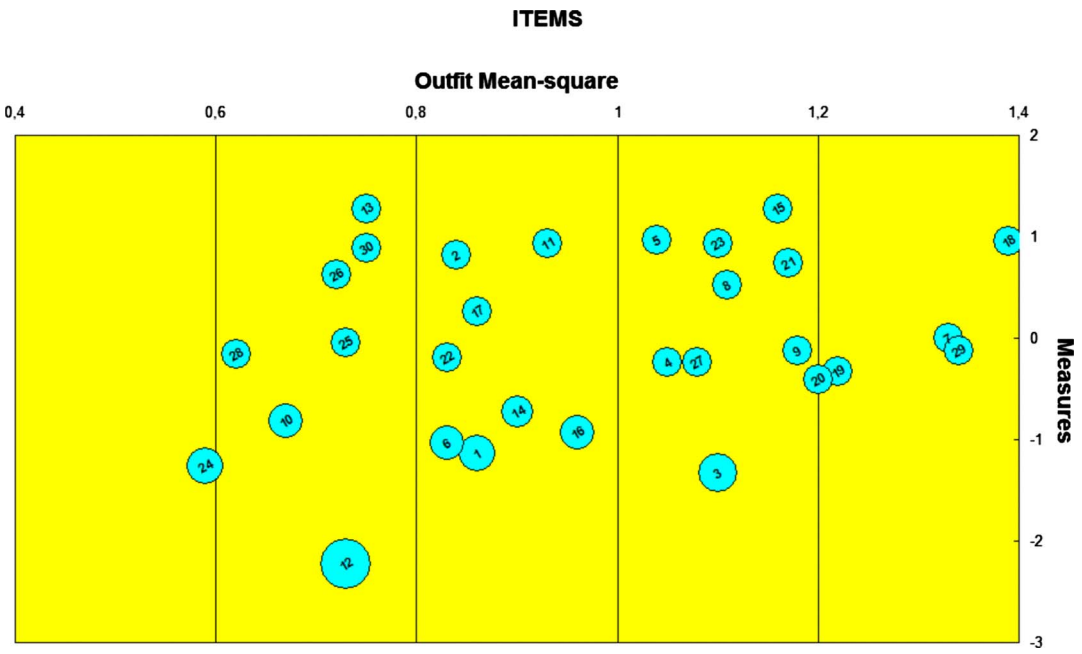

FIG. 6. (Color) The bubble chart for the university students showing outfit mean square statistics (MNSQ) vs item measure.

misfitting items which would degrade measurement. Some items with larger misfit do however require further examination, among them especially items 7, 18, and 29. The test is usually used as pretest and post-test on two qualitatively very different populations of students, non-Newtonian and predominantly Newtonian. This analysis suggests that the test may function differently on these two populations. The item difficulty order does not remain the same on both samples, as it should be if the construct is unchanged and well defined by the test. The item difficulty order is also not clear enough since many items are very close in difficulty. Items which change their position in the test significantly should be carefully examined, especially items 6–8.

The possible change in construct is not so surprising since the two populations may be quite different in their way of thinking about mechanics. Some issues which are very difficult for non-Newtonians can become quite easy for Newtonians (e.g., the idea that motion with constant velocity requires no net force). Some other issues may be hard for both groups. For measurement purposes we want the test construct to remain the same, but in practice we will often find that the construct has changed between two testing occasions, especially after some intervention has happened, such as the instruction on the topic of the test. Items that change their position in the test exhibit the so-called differential item functioning (DIF), which is quite common in the testing practice. It does not necessarily mean that the test is useless for measurement. We can determine student scores on pretest and post-test using only those items which are stable or we can construct the common scale from pretest and post-test data and measure students according to that scale on both occasions. On the other hand, the change in the construct can be informative of instruction efficiency in different areas measured by the test and of item quality (some items can exhibit DIF because they are poorly written or biased).

The width of the test is not sufficient to cover the whole range of abilities of both populations. There are too many very closely spaced items in the middle of the test and not enough items at the extremes of the test. This could be remedied by removing some of the items from the middle of the test and adding new items at the extremes. A possible solution could also be the construction of two different tests, one for pretest and another for posttest, which could be linked by a certain number of common items to enable the comparison of student scores on both tests. The Rasch model could provide linear measures of student abilities on both tests on the same scale, and the difference in ability measures on pretest and posttest could be used as the direct measure of student gain. Hake's normalized gain [2], which is generally used as the measure of student progress in the course, may also be influenced by the nonlinearity of raw scores expressed as percentages.

The Force Concept Inventory is an important instrument which has contributed very much to the development of physics education research and to the change of physics teaching practice throughout the world. It is a measurement instrument in physics education research, and as such it has to be calibrated, inspected, and carefully monitored, just like measurement instruments in physics. The Rasch model can be used as a powerful tool for monitoring and improving the functioning of the FCI, as well as of other diagnostic tests used in PER.

ACKNOWLEDGMENTS

This research is a part of the scientific project (Grant No. 119-0091361-1027) funded by the Croatian Ministry of Science, Education and Sports. We thank Planinka Pecina, Vlado Halusek, and Zeljko Jakopovic for their help with collection of data for this study. We also thank Mike Linacre for many helpful answers to our questions regarding Rasch analysis and the use of WINSTEPS software.

[1] D. Hestenes, M. Wells, and G. Swackhammer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[2] R. R. Hake, Interactive engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[3] D. Hestenes and I. Halloun, Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).

[4] T. Wilhelm, Verständnis der newtonschen mechanik bei bayerischen elftklässlern-ergebnisse beim test "force concept inventory" in herkömmlichen klassen und im würzburger kinematik-/dynamikunterricht, Physik und Didaktik in Schule und Hochschule, PhyDid 4, 47 (2005).

[5] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, Am. J. Phys. **53**, 1043 (1985).

[6] http://modeling.la.asu.edu/R&E/Research.html

[7] D. Huffman and P. Heller, What does the force concept inventory actually measure? Phys. Teach. **33**, 138 (1995).

[8] P. Heller and D. Huffman, Interpreting the force concept inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).

[9] J. Stewart, H. Griffin, and G. Stewart, Context sensitivity in the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **3**, 010102 (2007).

[10] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCaley, Testing the test: Item response curves and test quality, Am. J. Phys. **74**, 449 (2006).

[11] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the force and motion conceptual evaluation and the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **5**, 010105 (2009).

[12] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Interpreting force concept inventory scores: Normalized gain and SAT scores, Phys. Rev. ST Phys. Educ. Res. **3**, 010106 (2007).

[13] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

[14] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (Danmarks Paedagogiske, Copenhagen, 1960).

[15] M. Planinic, L. Ivanjek, A. Susac, P. Pecina, R. Krsnik, M. Planinic, Z. Jakopovic, and V. Halusek, *GIREP-EPEC Conference 2007: Selected Contributions* (Zlatni rez, Rijeka, 2008), pp. 295–300.

[16] J. M. Linacre, WINSTEPS Rasch measurement computer program, Winsteps.com, Chicago, 2006.

[17] R. Dittrich and R. Hatzinger, Using "GLIM" for computing the Rasch model and the corresponding multiplicative Poisson model, http://www.ihs.ac.at/publications/ihsfo/fo167.pdf

[18] B. D. Wright and M. H. Stone, *Best Test Design* (MESA Press, Chicago, IL, 1979).

[19] T. G. Bond and C. M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (Lawrence Erlbaum, Mahwah, NJ, 2001).

[20] J. M. Linacre, A users's guide to WINSTEPS, www.winsteps.com