Interpreting gains and losses in conceptual test using Item Response Theory

Jean-François Parmentier, Brahim Lamine

Interpreting gains and losses in conceptual test using Item Response Theory

Jean-François Parmentier[1, *] and Brahim Lamine[2, †]

[1]*Université de Toulouse, UPS, IRES, F-31400 Toulouse, France*
[2]*Université de Toulouse, UPS-OMP, CNRS, IRAP, F-31028 Toulouse, France*
(Dated: September 13, 2015)

Conceptual tests are widely used by physics instructors to assess students' conceptual understanding and compare teaching methods. It is common to look at students' changes in their answers between a pre-test and a post-test to quantify a transition in student's conceptions. This is often done by looking at the proportion of incorrect answers in the pre-test that changes to correct answers in the post-test – the gain – and the proportion of correct answers that changes to incorrect answers – the loss. By comparing theoretical predictions to experimental data on the Force Concept Inventory, we shown that Item Response Theory (IRT) is able to fairly well predict the observed gains and losses. We then use IRT to quantify the student's changes in a test-retest situation when no learning occurs and show that i) up to 25% of total answers can change due to the non-deterministic nature of student's answer and that ii) gains and losses can go from 0% to 100%. Still using IRT, we highlight the conditions that must satisfy a test in order to minimize gains and losses when no learning occurs. Finally, recommendations on the interpretation of such pre/post-test progression with respect to the initial level of students are proposed.

I. INTRODUCTION

Conceptual tests are widely used by physics instructor to asses students' conceptual understanding and compare teaching methods. In particular, the Force Concept Inventory [1] (FCI) evaluate student's mastering of Newton laws [2]. It consists of 30 multiple-choice questions where incorrect answers are based on the most frequently answers given by students in interviews. Many topics are covered by the FCI : kinematics, identification of forces and the three Newton's laws [1, 3]. Instructors usually use the raw score or the Hake gain [2] to evaluate global student's progression. Item Response Theory (IRT) provide a more theoretically grounded measure of student's progression [4–6]. Over the past decade, IRT have been applied with success to concept inventories, in particular to the FCI [7–11]. Student's raw score or student's proficiency given by IRT provide a global measure of the acquisition of the Newtonian concepts.

A closer look to student's answer in a test-retest situation has shown that while the total score to the test is highly reliable, 31% of the student's answers change from test to retest, suggesting weak reliability for individual answers [12]. Looking how answers of students change between a pre-test – before instruction – and a post-test – after instruction – using a database embedding more than 13 000 students' answers, Lasry et al. [13] revealed a strong positive correlation between the initial score and the proportion of incorrect answers on the pre-test that were changed to correct answers on the post-test – the gains. A symmetric result was found for the losses – the proportion of correct answers on the pre-test that were changed to incorrect answers on the post-test, strongly and negatively correlated to the initial score. This result suggests that students with higher prior level learn more and forget less than students with lower prior level.

In this article we show that IRT can be used to qualitatively predict those experimental data while offering another interpretation of the previous results. The observed correlation mainly comes from inherent properties of the test rather than reflecting the level of progression of students. We show in particular that the student's proficiency progression, as obtained by IRT, increases for low proficiency students, a conclusion at the opposite of the previous interpretation.

The article is organized as follow : section II provides definition of gains and losses; section III introduces IRT theory and the underlying assumptions; section IV compares theory's predictions with experimental data; section V exploit IRT to predict answer's changes; finally section VI and VII discuss and conclude this work.

II. GAINS AND LOSSES

Consider the situation of students taking a same test two times : the first one before instruction and the second one after instruction. It is hoped that the score of each student increases, so that a part of answers which were initially wrong becomes correct. Following Lasry et al. [13], we define the gain G as the proportion of incorrect answers on the pre-test that change to correct answers on the post-test. Similarly, the loss L is defined as the proportion of correct answers on the pre-test that change to incorrect answers on the post-test. We then introduce IC_i as the proportion of students who change from an incorrect (I) to a correct (C) answer at the question i and I_i as the proportion of initial incorrect answers. gains and losses are then defined by $G = \overline{IC_i}/\overline{I_i}$ and $L = \overline{CI_i}/\overline{C_i}$, where $\overline{(\,.\,)}$ denotes the average over the questions of the test. C_i is the proportion of initial correct answers to question i so that $\overline{C_i}$ is the average pre-test score of the students. Using data
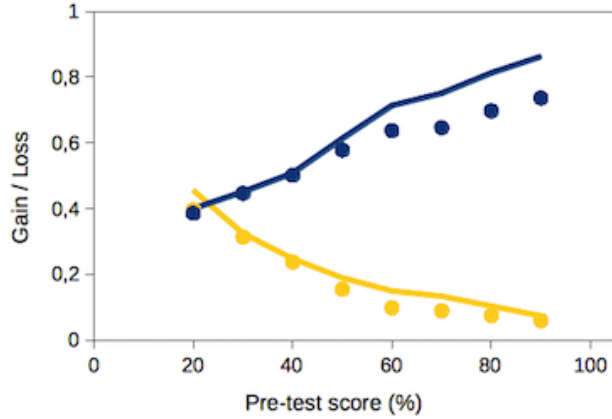
FIG. 1. Gain (blue) and loss (yellow) as a function of pre-test score at the FCI. Points are measurements from a large pool of students [13] and lines are theoretical predictions using questions parameters of the IRT analysis obtained in [9].

FIG. 2. Item characteristic curves for questions 1 (dashed line) and 13 (continuous line) of the FCI. Questions parameters are taken from [9].

from more than 13,000 students' answers on the Force Concept Inventory (FCI), Lasry et al. [13] measured dependance of gains and losses with prior knowledge (pre-test score). As shown in Fig. 1, students with higher prior knowledge have higher gain and smaller loss than students with lower prior knowledge. In order to interpret these results, it is first necessary to draw the same graph when no learning occurs. That is to say when the same test is taken two times consecutively, with student not memorizing their previous answers and not having learned anything between the two tests. We show in the next sections how IRT is able to answer this question.

III. THE ITEM RESPONSE THEORY

Item Response Theory (IRT) belongs to the family of latent trait modeling [14]. In those models, each student is described by a number of latent traits, also call proficiencies. The answer of a student to a question is thought of as the result of the interaction between the capabilities of the person taking the test and the characteristics of the test items. The score of a student to an item is modeled by a probabilistic function of his proficiencies and the item's characteristics. A consequent number of knowledge and skills are always necessary to give a correct answer [15] but in many cases, only one proficiency is sufficient to determine the student score. This is call unidimensional Item Response Theory but is often simply called IRT. This assumption was shown to be valid to model student's answer to the FCI [8, 9] and will be assumed in the following.

Let's note θ the proficiency of a student. Each question i is modeled by a function $P_i(\theta)$ which describes the probability of a student with proficiency θ to correctly answer to the question i. P_i functions, called item characteristic curves, are often assumed to be generic "S-shape" functions (see Fig. 2), called logistic function, whose varia-
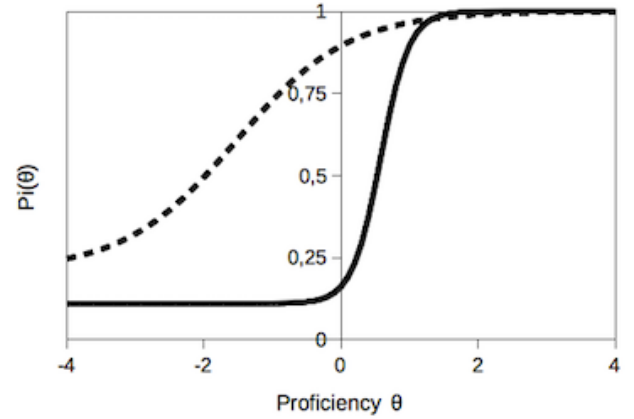
tions characterize each questions. In the three-parameter item model, $P_i(\theta)$ is given by

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp\left[-1.7\, a_i(\theta - b_i)\right]} \ , \qquad (1)$$

where a_i, b_i and c_i are parameters of the question : a_i is its discrimination power, b_i its difficulty and c_i the probability of guessing. The parameters are estimated by statistical techniques using a large pool of students answers. Other models exist such as the two-parameter model ($c_i = 0$), the Rasch model ($c_i = 0$ and $a_i = 1$) and the non-parametric kernel smoothing approach [16]. All these models have been applied to the FCI [7–11]. For instance, P_i functions for question 1 and 13 of the FCI are plotted in Fig. 2. Question 13 is more difficult than question 1 ($b_{13} > b_1$) so its curve is more "on the right" of the graph. Its discrimination is also larger ($a_{13} > a_1$) so that the S-shape is steeper. Finally, the guessing parameter is lower ($c_{13} < c_1$), as seen on the value of P_i when θ goes to $-\infty$.

The true score (in %) of a group of students with proficiency θ is given by $S(\theta) = \overline{P_i(\theta)}$. Because of the probabilistic nature of IRT, the score $S(\theta)$ for a given proficiency θ differs from the observed score of a student with that proficiency θ – the number of correct answer given by the student divided by the number of questions. The true score $S(\theta)$ is only recovered as an average over a large number of equal-proficiency student's individual observed scores. The observed score is also named the raw score and one strength of IRT is to convert this raw score, which is a discrete bounded variable, into a continuous unbounded variable, θ, which is assumed to be an interval scale – i.e. a scale which can be used to quantify a progression or a difference of proficiency between students [4].

IV. IRT PREDICTION OF GAINS AND LOSSES

The objective of a course is to increase student's proficiency. Let's write θ_{pre} the proficiency of a student before instruction and θ_{post} its proficiency after instruction. By definition, the probability of choosing the correct answer to the question i during the pre-test is $P_i(\theta_{pre})$. For the same reason, this probability is $P_i(\theta_{post})$ for the post-test. For a wide group of student with the same proficiencies, we get $I_i = 1 - P_i(\theta_{pre})$ and $IC_i = (1 - P_i(\theta_{pre})) P_i(\theta_{post})$. Reporting these equations into the definition of the gain and the loss leads to

$$G = S_{post} - \frac{\overline{\delta P_i(\theta_{pre}) \, \delta P_i(\theta_{post})}}{1 - S_{pre}} \, , \qquad (2)$$

$$L = (1 - S_{post}) - \frac{\overline{\delta P_i(\theta_{pre}) \, \delta P_i(\theta_{post})}}{S_{pre}} \, , \qquad (3)$$

where $\delta P_i(\theta)$ is the difference between probability of success of question i and average test score S for a given proficiency :

$$\delta P_i(\theta) = P_i(\theta) - \overline{P_i(\theta)}. \qquad (4)$$

By definition $\overline{\delta P_i(\theta)} = 0$. In the particular case when $\theta_{pre} = \theta_{post}$ (i.e. when no instruction occurs), $\overline{\delta P_i \, \delta P_i}$ is the variance of the P_i's for a given θ and is a characteristic of the test.

Equations (2) and (3) show that IRT enables us to predict measured values for G and L once θ_{pre}, θ_{post} and all the P_i's are known. However, data of Lasry et al. [13] give values of G and L as functions of S_{pre} so informations about θ_{pre}, θ_{post} and all the P_i's function are missing.

First P_i functions are taken form literature. Using the three-parameter model, Wang and Bao [9] performed an IRT analysis of the FCI using their own database of $2\,800$ student's answers, leading to the knowledge of the 30 P_i functions. The measurements obtained by Wang and Bao with their students can be used for any students because characteristics of questions are independent of the population used to obtained them. This property is known as parameter invariance [17]. Hence there P_i functions are used here.

Secondly, foreach values of S_{pre} we estimated S_{post} from data of Lasry et al. [13] using

$$S_{post} = S_{pre} \, (1 - L) + (1 - S_{pre}) \, G \, , \qquad (5)$$

which comes from the definition of G and L and the fact that $S_{pre} = \overline{C_i}$.

And finally θ_{pre} and θ_{post} are estimated by reversing the relation giving S as a function of θ : $S(\theta) = \overline{P_i(\theta)}$. This is an approximation where the observed raw score is assumed to be equal to the true score. The sample of Lasry et al. [13] contains $13\,000$ students divided into 9 bins leading to an average of $1\,400$ students for each raw score. In this case the hypothesis of equating the raw score to the true score seems reasonable.

Figure 1 shows that eqs. (2) and (3) match fairly well the experimental measurements, indicating that IRT is able to correctly predict gains and losses. Discrepancies can be attributed to both uncertainties of measurements of P_i and to an unperfect parameter invariance. Such a case can occur in particular when the hypothesis of unidimensionality does not hold. As shown by Scott and Schumayer [3], while a unique proficiency can be used to describe student's characteristic, a 5 dimensional model seems preferable. Our results show that a one-dimensional model is able to give the global tendency for the gain and the loss. A more detailed analysis is reported for future work.

As seen in Fig. 1, gain is an increasing function of student's initial score. A tempting interpretation is to say that students with higher initial knowledge learn more than students with lower initial knowledge. The reverse is also true for loss : students with higher initial knowledge have lower loss than students with lower initial knowledge. However this argument implicitly assumes that gains and losses are zero when no learning occurs. We now show that this is not the case, which at least makes the previous conclusion unsecured. To do so, we use IRT to estimate G and L when $\theta_{post} = \theta_{pre}$, using equations (2) and (3). Results are plotted in Fig. 3, which clearly show that even when no learning occurs gain is an increasing function of the pre-test score and raise up to one. Similarly, loss goes down from one to zero as pre-test score increases. For a pre-test score value of 50% both gains and losses have the same value around 35%. Such a change in student answers at the same question has been observed between two successive passes of the FCI [12]. Reported values of gains and losses were 18% and 20% for a population mean score of 47%. Discrepancy between their experimental measures and IRT prediction could largely be attributed to a memory effect because students took the tests two times in the same week so they may have memorized some of there initial answers. At the contrary, our IRT model assumes the independence between the test-retest, i.e. that students have not memorized any of their previous answers.

V. PROPORTION OF ANSWER'S CHANGE

In order to interpret why gains and losses can have such high values even when no learning occurs, we focus directly on the global proportion of answer's change. In a test-retest situation, we have :

$$\overline{IC_i} = \overline{CI_i} = S \, (1 - S) - \overline{\delta P_i^2} \, , \qquad (6)$$

where $S = S_{pre} = S_{post}$. The explicit dependence of S and δP_i with $\theta_{pre} = \theta_{post}$ have been omitted for clarity. The first term of the right hand side of equation (6) is a parabolic function of S and does not depend on the considered test. Hence, for any conceptual test, this part is identical. The second term on the right hand side of equation (6) depends on the item characteristic curves and consequently on the test. Values of $\overline{IC_i}$ have been
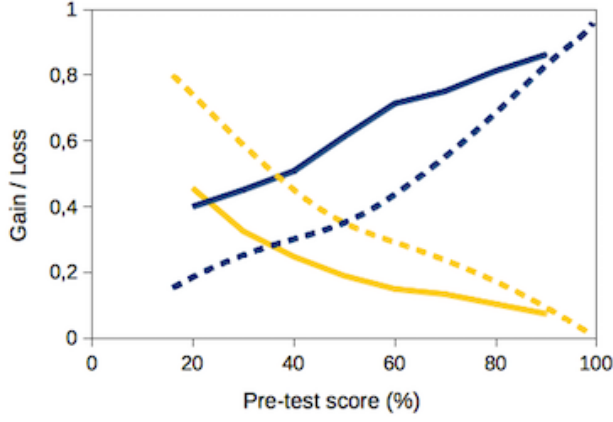
FIG. 3. Gain (blue lines) and loss (yellow lines) as a function of pre-test score at the FCI. Continuous lines are IRT predictions when learning occurs, dashed lines are IRT predictions when no learning occurs (i.e. assuming $\theta_{post} = \theta_{pre}$).
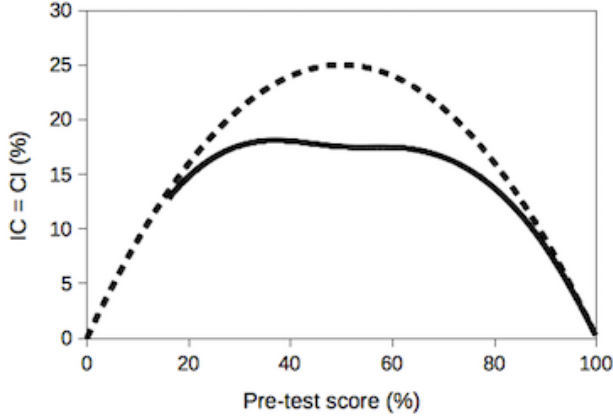

FIG. 4. Proportion of answer's changes from a right (resp. wrong) answer to a wrong (resp. right) answer (continuous line) for the FCI. Dashed line is $S(1-S)$.

plotted for the FCI as a function of the score in Fig. 4. It is clear that in this case, the contribution of $\overline{\delta P_i^2}$, while not negligible, is rather small. Consequently, for a group of students with a true score of 50 %, nearly 18 % of answers change from correct (resp. incorrect) to incorrect (resp. correct) in a test-retest situation. This result has a consequence on the reliability of the test and on the interpretation of gains and losses. In order to interpret gains and losses in term of learning outcome, their values should be as small as possible in a test-retest situation. As a consequence, values of $\overline{IC_i}$ should also be as small as possible. Because the first term of equation (6) does not depend on the test, one can only influence the $\overline{\delta P_i^2}$ term in order to make it as high as possible (so that $\overline{IC_i}$ decreases). It immediately leads to the conclusion that one has to choose questions – therefore the P_i's functions – in order to maximize values of $\overline{\delta P_i^2}$ for all θ.

In order to understand how to choose those P_i's func-

tions, we consider the simple case of a test with only 3 questions. Three different cases are considered, each one corresponding to a particular set of P_i's functions. The three cases are named test A, B and C and their item characteristic functions are plotted in Fig. 5 (left column). For each θ, the proportion of answer's change is given by $\overline{CI_i} = \overline{P_i(1 - P_i)}$, where $\overline{(\,.\,)}$ denotes the averaging over the 3 questions of the test. Hence, each individual question i has a contribution of $P_i(1 - P_i)$. This contribution is null when $P_i = 0$ or 1 and has a maximal value of 0.25 when $P_i = 0.5$.

Test A has three questions whose characteristic curves overlap for a wide range of θ. As a consequence, for a wide range of θ all individual questions will contribute to the proportion of answers that change. For instance, for a true score of 50% ($\theta = 0$), $P_1(\theta) = 0.88$, $P_2(\theta) = 0.5$, and $P_3(\theta) = 0.12$, leading to $P_1(1 - P_1) = P_3(1 - P_3) = 0.1$ and $P_2(1 - P_2) = 0.25$. Hence, for a score of 50%, the proportion of change, which is the average of these three values, is about 15%. The representative curve of $\overline{IC_i}$ is very similar to the one obtained for the FCI, indicating that a lot of item characteristic curves of the FCI overlap, as already noted in previous studies analyzing the FCI using a unidimensional IRT [7–11].

At the opposite, test C has three questions whose characteristic curves do not overlap - i.e. the range of θ where these functions go from a value close to 0 to a value close to 1 are well separated (see Fig. 5). As a consequence, each question will contribute separately to the proportion of answer's change. For instance, for a true score of 50% ($\theta = 0$), $P_1(\theta) \simeq 1$, $P_2(\theta) = 0.5$, and $P_3(\theta) \simeq 0$, leading to $P_1(1 - P_1) = P_3(1 - P_3) \simeq 0$ and $P_2(1 - P_2) = 0.25$. Hence, for a score of 50%, the proportion of change – which is the average of the P_i values – is $0.25/3 \simeq 0.08$. This value is much smaller than for test A. In a test with N separated questions, the maximal value of $\overline{IC_i}$ is $0.25/N$ and is obtained for values of $S = 0.5/N$, $1.5/N$, ... , $(N - 0.5)/N$. In a test with $N = 30$ separated-questions, maximal value for $\overline{IC_i}$ is about 1%. Hence the change of answers occurs very rarely, and values of gains and losses remain very small.

Finally test B shows the transition between test A and the extreme case of test C.

VI. INTERPRETATION OF GAINS AND LOSSES WHEN LEARNING OCCURS

According to the discussion of the previous section, the interpretation of gains and losses should be separated in two extreme cases : when a wide majority of item characteristic curves overlap – like in test A – and when none of the item characteristic curves overlaps – like in test C.

In the first case, $\overline{\delta P_i^2}$ is small and equations (2) and (3) reduce to $G = S_{post}$ and $L = 1 - S_{post}$. Hence the gain is more or less the post-test score and does not add any supplementary informations on student's learning. One can still want to isolate the part of the gain due to in-
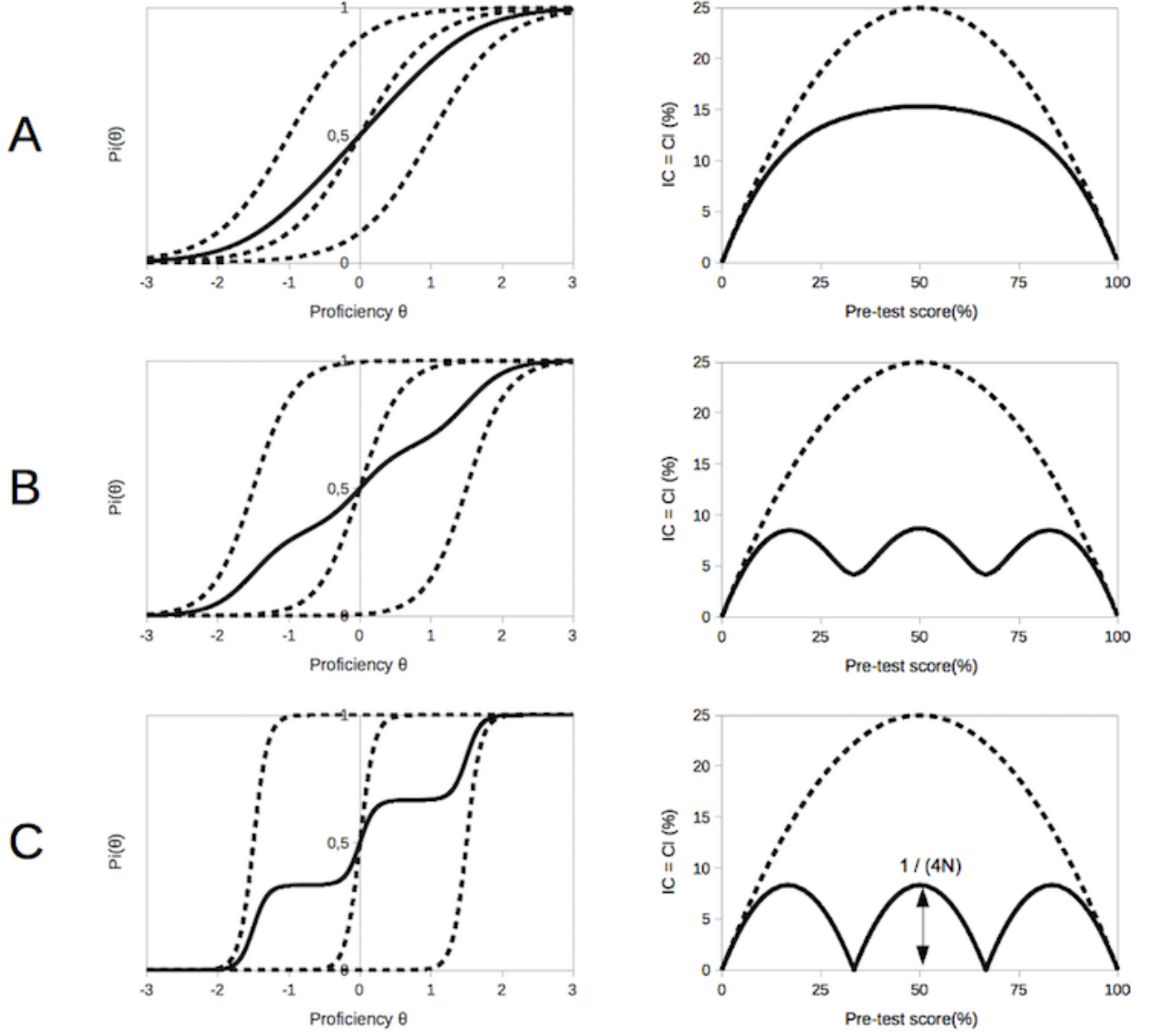
FIG. 5. Each row corresponds to given tests (A, B or C) comprising 3 questions. Left : item characteristic curves of the three questions (dashed lines) and true score (continuous lines) as functions of proficiency θ. Right : proportion of answer's change $\overline{IC_i} = \overline{CI_i}$ (continuous line) and $S(1 - S)$ (dashed-line) as functions of the true score S.

struction by defining $\Delta G = G_{\text{learning}} - G_{\text{no learning}}$. In the case of type A test, $\Delta G = S_{post} - S_{pre} = g_{raw}$, leading to the so-called raw gain (because $G = S_{pre}$ when no learning occurs). The analysis of Lasry et al. [13] data shows that g_{raw} is a decreasing function of the pre-test score. Does it mean that students with lower initial knowledge gain more than students with higher initial knowledge ? No because student's post score is limited to 100% so the raw gain g_{raw} tends to zero when the pre-test score tends to 100%. Also the score is an ordinal scale and not an interval scale [4–6]. As a consequence, the raw score can only lead to a sorting of students but an increase of 1 point for a student with a low initial score does not reflect the same learning than an increase of 1 point for a student with a high initial score. A correct

comparison of progress has to invoke an interval scale such as the student proficiency θ introduced in the previous sections [4–6]. Fig. 6 plots the raw gain as a function of the pre-test score for given values of students learning increase $\Delta \theta$. As seen on this figure, a given value of g_{raw} corresponds to various value of student's progression $\Delta \theta$, depending of the initial student's score.

In the second case (test of type C), where all questions are well separated, the proportion of questions that changes when no learning occurs is nearly null – it is lower than 5% for $N \geq 5$. Assuming a student positive progression $\Delta \theta = \theta_{post} - \theta_{pre}$ greater than the error range of all questions (i.e. $\forall i, \ \Delta \theta \gg 1/a_i$ with a_i the discrimination power), the number of answers that change from
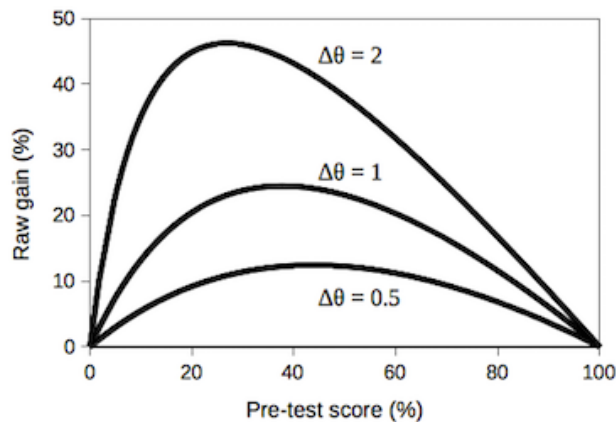
FIG. 6. Evolution of the raw gain with initial pre-test score for three fixed values of student's learning $\Delta\theta$. The raw gain corresponds to ΔG for a type A test.
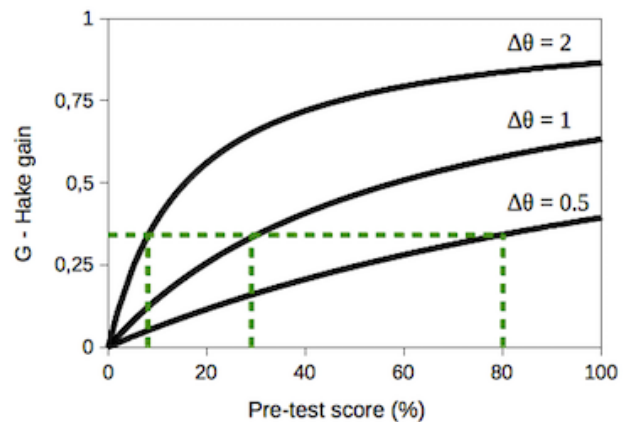

FIG. 7. Evolution of the Hake gain with initial pre-test score for three fixed values of student's learning $\Delta\theta$. Green dashed line is $G = 0.34$ and correspond to $\Delta\theta = 2$ for S=8%, $\Delta\theta = 1$ for S=30% and $\Delta\theta = 0.5$ for $S = 80\%$. The Hake gain corresponds to ΔG for a type C test.

incorrect to correct is $S_{post} - S_{pre}$ leading for the gain to

$$G \;=\; (S_{post} - S_{pre}) \,/\, (1 - S_{pre}) = G_{\text{Hake}} \; . \qquad (7)$$

Interestingly, one recovers in this limit the Hake's gain [2], which can be interpreted as the proportion of questions changing from incorrect to correct in a test comprising seperated item response curves (like test C). The number of answers that change from correct to incorrect is null and $L = 0$. However, like the raw gain, the Hake gain is not an interval scale [6] and has to be taken with due care when comparing student's progression, as already emphasized. To illustrate this, let's consider an hypothetical test where the true score is a logistic function of the proficiency : $S = (1 + \exp(-\theta))^{-1}$. This model is characteristic of a test where question's difficulties are distributed over the proficiency scale following a gaussian law : there are few easy questions, few hard questions and a wide majority of questions with an intermediate level of difficulty. The Hake gain is plotted on Fig. 7 as a function of the pre-test score for various fixed value of student's learning $\Delta\theta$ that are typical of student's learning (see for instance Fig. 8 for typical values of $\Delta\theta$ in a mechanic course). As can be seen, the gain is an increasing function of the pre-test score for a fixed value of student's learning. Hence, the fact that the gain is larger for initial high level students than for initial low level students does not necessarily reveal that the initial high level students have learned more. Moreover, a given value of G corresponds to various value of student's progression $\Delta\theta$, depending of the initial student's score. As shown in Fig. 7, a fixed value of the gain – for instance 0.34 – correspond to a strong learning for low pre-test score ($\Delta\theta = 2$ for S=8%), a medium learning for medium pre-test score ($\Delta\theta = 1$ for S=30%) and a low learning for high pre-test score ($\Delta\theta = 0.5$ for $S = 80\%$). This clearly shows that the Hake gain should not be used to compare student's progression when they have different pre-test score, even in test of type C.

Table I summarizes values of G and L for the two limit

Type of test	G	L	ΔG
A	S_{post}	$1 - S_{post}$	g_{raw}
C	G_{Hake}	0	G_{Hake}

TABLE I. Summary of gains and losses for the different types of test. $\Delta G = G_{\text{learning}} - G_{\text{no learning}}$ is the difference in gain between a situation when learning occurs and a situation when no learning occurs, that is to say the part of the gain which is due to learning.

cases. As can be seen, ΔG reduces to the raw gain for type A tests and to the Hake gain for type C tests.

We conclude this section by discussing the efficiency of instruction with respect to the initial level of the students. As already emphasized, the proficiency θ has good properties [4–6] and hence could be used to determine the learning $\Delta\theta$ of a student, $\Delta\theta = \theta_{post} - \theta_{pre}$. This increase of proficiency is plotted in Fig. 8 as a function of the pre-test score for the data of Lasry et al. [13]. We have evaluated θ using the scores by inverting the relation $S(\theta)$. According to Lasry et al. [13], uncertainties on pre-test scores, gains and losses are about 2%, leading to uncertainties on the post-test score of the same order of magnitude. These uncertainties lead to uncertainties on the proficiencies, particularly for low or high scores due to the 'S' shape of the curve, and are represented in Fig. 8. If θ is assumed to be the good scale for measuring the learning, Fig. 8 clearly shows that learning decreases as the pre-test score increases. This is an opposite conclusion with the first interpretation of the evolution of gains and losses with pre-test score, but in accordance with the evolution of g_{raw} with pre-test score. It seems to state that our teaching methods are more efficient on students with low prior knowledge. We recall that this result is based on data from more than 13,000 students who had taken the FCI at the beginning and at the end of an introductory physics course in a large variety of
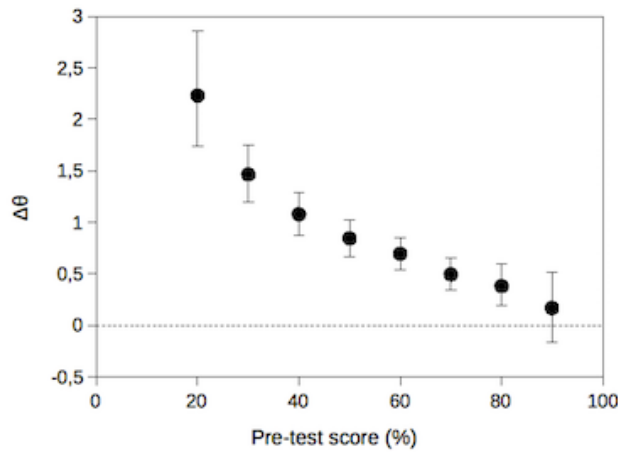
FIG. 8. Evolution of student's learning ($\Delta\theta = \theta_{post} - \theta_{pre}$) with the pre-test score evaluated from data of Lasry et al.[13].

institutions: US high schools (10,007) , three Canadian two-year colleges (971), a US public university (1560) and three top-tier private universities (884) [13]. Due to possible correlations between students' prior knowledge and student's institution, this could reflect a difference between institutions. But this also could mean that it is more difficult in an introductory physics course to give the same increase of learning to students with high prior level knowledge than to students with low prior level knowledge. This discussion is out of the scope of this article but in order to answer this question one would have to evaluate $\Delta\theta$ for each student in a group follow-ing the same course with the same teacher, plotting the same curve as in Fig. 8 and finally perform a comparison across institutions.

VII. CONCLUSION

We have shown that IRT is able to fairly well predict experimental measurements of gains and losses with the FCI when learning occurs. In addition, IRT shows that values of gains and losses for the FCI are rather high even when no learning occurs. The reason being that item characteristics curves overlap. All errors associated to individual questions contribute together to the probability of answer's change, leading to a difficult interpretation of gains and losses. In such a case the gain is more or less the post-test score and does not reveal that initial high level students have learned more that initial low level students.

In the case where item characteristic curves do not overlap, answer's changes are very low, the gain reduces to the Hake gain while the losses drop to zero.

We have shown that the effect of instruction can be assessed by looking to the proficiency increase instead of looking to the gain increase. The proficiency increases more for low-level student (i.e. low pre-test score).

ACKNOWLEDGMENTS

[1] D. Hestenes, M. Wells, and G. Swackhamer, The Physics Teacher **30**, 141 (1992).

[2] R. R. Hake, American Journal of Physics **66**, 64 (1998).

[3] T. F. Scott, D. Schumayer, and A. R. Gray, Physical Review Special Topics - Physics Education Research **8**, 020105 (2012).

[4] B. D. Wright and J. M. Linacre, Archives of physical medicine and rehabilitation **70**, 857 (1989).

[5] B. D. Wright, Educational Measurement: Issues and Practice **16**, 33 (1997).

[6] C. S. Wallace and J. M. Bailey, Astronomy Education Review **9** (2010), 10.3847/AER2010024.

[7] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. Mc-Cauley, American Journal of Physics **74**, 449 (2006).

[8] M. Planinic, L. Ivanjek, and A. Susac, Physical Review Special Topics - Physics Education Research **6**, 010103 (2010).

[9] J. Wang and L. Bao, American Journal of Physics **78**, 1064 (2010).

[10] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, American Journal of Physics **80**, 825 (2012).

[11] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Physical Review Special Topics - Physics Education Research **11**, 010112 (2015).

[12] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, American Journal of Physics **79**, 909 (2011).

[13] N. Lasry, J. Guillemette, and E. Mazur, Nature Physics **10**, 402 (2014).

[14] A. Kamata and D. J. Bauer, Structural Equation Modeling: A Multidisciplinary Journal **15**, 136 (2008).

[15] M. Reckase, *Multidimensional Item Response Theory* (Springer New York, New York, NY, 2009).

[16] J. O. Ramsay, Psychometrika **56**, 611 (1991).

[17] A. A. Rupp and B. D. Zumbo, Educational and Psychological Measurement **66**, 63 (2006).