How to quantify the efficiency of a pedagogical intervention with a single question

Jean-François Parmentier

How to quantify the efficiency of a pedagogical intervention with a single question

Jean-François Parmentier*

CERFACS, 42 Avenue Coriolis, 31057 Toulouse, France and Université de Toulouse, UPS, IRES, F-31400 Toulouse, France

(Received 3 June 2018; published 12 November 2018)

In many situations, the change in the conceptual understanding of students is measured using a single question. This is, for instance, the case in peer instruction where students answer twice to the same questions, before and after the discussion phase. Using item response theory and assuming that students proficiencies are normally distributed, it is shown that the Cohen's *d* effect size characterizing the change of mean proficiencies can be estimated by taking 0.6 times the log of the odds ratio of class scores. Moreover the polychoric correlation coefficient between students' answers is suggested as an additional indicator to detect abnormal changes in scores when its value is below 0.3. Taken together, these two indicators give both a precise measurement of a pedagogical intervention—a peer discussion or something else—and a coefficient of security to detect random answers or poor writing of questions. The application is made to the evaluation of peer discussions that took place in an introductory mechanics course taught using peer instruction.

DOI: 10.1103/PhysRevPhysEducRes.14.020116

I. INTRODUCTION

Concept inventories are widely used in education research to evaluate changes in conceptual understanding related to a specific intervention [1]. In this case they are used twice: the first one before the intervention—generally a full semester course—and the second time after the intervention—at the end of the semester. For instance, the Force Concept Inventory (FCI) [2] evaluates students mastering of Newton' s laws [3]. It is composed of 30 multiple-choice questions where incorrect answers are based on the most frequent answers given by students during interviews. Concept inventories are difficult to design [1,4] and their administration takes time and requires some caution [4].

Instructors or researchers can measure the change in conceptual understanding of students due to a pedagogical intervention using only one question. A typical example is the use of Peer Instruction (PI), an evidence-based, interactive teaching method, widely used by science teachers [5,6]. A class taught with PI is divided into a series of short presentations, each focused on a central point and followed by a related conceptual question, called a ConcepTest, which probes students understanding of the ideas just presented. Students are first given one or two minutes to formulate individual answers and report them to the instructor using classroom response systems such as clickers. Students then discuss their answers with others sitting around them. A few minutes afterwards, the instructor calls for an end to the discussion and polls students for their answers again, which may have changed based on the discussion. Finally, the instructor explains the answer and moves on to the next topic. In this case, the change of conceptual understanding is measured by the two votes, and the pedagogical intervention is the discussion. Many kinds of interventions can be measured in this way. For instance, instead of a discussion with peers, the instructor can give a hint, or give additional time to think about the question [7]. Variations of the instructor's directions for the discussion can also be tested, such as telling students to reach a consensus with their peers [7]. In online learning environments, an intervention can be the use of an online discussion board to discuss the question with peers, or to display to each student only a few selected particular rationales written by previous students [8,9].

The efficiency of an intervention can be analyzed qualitatively, for instance, by listening to the student's dialogues [10,11]. While this technique is powerful, it is extremely time consuming. Hence it is dedicated to research applications. In this article, the efficiency of an intervention is evaluated through the variation of students' answers between the first vote—before the intervention—and the second vote. The analysis leads to two quantitative outputs: the Cohen's *d* effect size of the learning and a correlation index enabling us to detect a global guessing behavior or poor item formulations. A main advantage of using students answers is that they can be computed

*parmentier@cerfacs.fr

automatically, without any human intervention. Our two indicators are based on latent trait modeling, i.e., item response theory, in order to give the most reliable information.

Potential applications of the proposed method include the following:

- Teachers who have to create many ConcepTests for their courses and then to select, year after year, which questions to keep and which ones to move or discard.
- Researchers who want to test variations of the instructor's instructions in peer instruction quantifying how many times a particular method is better than the others.
- Software designers and online-learning teachers wanting to select the best among various strategies.

The article is organized as follows: Sections II and III deal with traditional measures based on the variation of students' scores and highlight their limitations, Secs. IV and V introduce item response theory and assumptions on students' conceptual understanding, and, finally, Secs. VI, VII, and VIII introduce how to calculate new indicators of progression of students on conceptual understanding related to the question that are based on item response theory. Section IX presents an application to items used in a course taught using PI.

II. USUAL MEASURES OF DIFFERENCES BETWEEN PROPORTIONS

Pre- and postintervention scores are the proportions of students who answer correctly to the question. Note that p_1 is the proportion of students who answer correctly at the first vote, and p_2 the proportion of students who answer correctly at the second vote, i.e., after the pedagogical intervention. Usual measures to compare the effect of an intervention are the risk difference, the risk ratio, and the odds ratio [12]. All those measures are based on a probabilistic point of view and are often used in epidemiology to reduce the risk for people to suffer from a particular disease.

The risk difference is simply the difference in risk (probability) of an event between two groups. In this case, the risk difference is $\mathrm{RD} = p_2 - p_1$.

The risk ratio, also called the relative risk, is the ratio of two risks. In this case it is given by $\mathrm{RR} = p_2/p_1$. For instance, if there are 40% of students who answer correctly at the first vote, and 70% at the second one, the risk ratio is 1.75. It means that students are 1.75 times more likely to have a correct answer after the pedagogical intervention than before it.

Where the risk ratio is the ratio of two risks, the odds ratio is the ratio of two odds. Here, the odds of success after discussion is $p_2/(1-p_2)$, while the odds of success before discussion is $p_1/(1-p_1)$. Using the same example as previously, before the intervention, students are $0.67 = 0.4/(1-0.4)$ times more likely to give a correct answer than to give an incorrect one. After the intervention, they are $2.3 = 0.7/(1-0.7)$ times more likely to give a correct answer than to give an incorrect one. The ratio of the two odds is defined by $\mathrm{OR} = p_2/(1-p_2) \times (1-p_1)/p_1$. In our example, it is equal to $2.3/0.67 = 3.5$.

III. WHAT'S WRONG WITH THOSE MEASURES?

The risk difference is not an interval scale [13]: the signifiance of a particular value of RD depends on the initial prescore. For instance, a RD of 10% does not have the same signifiance whether the initial score is 10%, 50%, or 90%. Hence the RD cannot be a correct indicator to compare the efficiency of an intervention independently of the initial score.

The risk ratio suffers from the same problem. A RR of 1.5 does not have the same meaning whether the initial score is 10%, 50%, or 90%—it is even impossible to get a RR of 1.5 if the initial score is 90%.

The odds ratio is also not an interval scale. However, as it will be shown in the next sections, its logarithm value is, under some assumptions that will be detailed later, an interval scale.

RD, RR, and OR can be used to classify two interventions starting with the same initial prescore. For instance, if the interventions A and B start both with an initial score $p_1 = 40\%$ but intervention A leads to an increase of 20% while intervention B leads to an increase of 10%, it can be concluded that intervention A is better than the intervention B. However, these measures cannot be used

- to compare two interventions starting with two different initial scores.
- or to quantify how many times an intervention is greater than another, even if they both start with the same initial score. For instance, while intervention A leads to an increase of 20% and intervention B leads to an increase of 10%, it cannot be concluded that intervention A is twice as good as intervention B.

In the following sections, item response theory will be used to build a quantitative indicator that could overcome these limitations.

IV. 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 called 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 one's proficiencies and some item's characteristics. A consequent amount 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, often simply called IRT.

In unidimensional IRT, students are described by a single continuous unbounded variable, called the proficiency. One claim of IRT is that this proficiency is an interval scale. This means that an equal difference in proficiency always has the same signifiance, independently of the initial proficiency. Hence this scale could be used to measure the efficiency of a pedagogical intervention, whether the initial score to the item is 10%, 50%, 90% or any value. It can also be used to calculate how much one intervention is better than another.

The aim of a question is to test students' understanding of a particular concept. Note that θ is the proficiency of a student which is measured by the question. A greater value of θ means a greater understanding of the associated concept. While the understanding of a concept is sometimes thought of as a binary variable, IRT assumes that it is a continuous scale. Application and validity of IRT to physics questions have been illustrated by analyzing score patterns of students to the Force Concept Inventory [16–19], the Mechanics Baseline Test [20], the Force and Motion Conceptual Evaluation [21], the Brief Electricity and Magnetism Assessment [22,23], and the Continuous Time Signals and Systems Concept Inventory [24].

In IRT, an item is modeled by a function $P(\theta)$, which describes the probability of a student with proficiency θ to give the correct answer to the item. The P function, called the item characteristic curve, is often assumed to be a generic "S-shape" function, called a logistic function, whose form characterizes each question. In this work, the P function is assumed to follow the two-parameter logistic item model, called the 2PL model:

$$P(\theta) = \frac{1}{1 + \exp\left[-1.7a(\theta - b)\right]}, \tag{1}$$

where a and b are parameters of the item: a is its discrimination power, and b its difficulty. Usually these parameters are estimated using statistical techniques on a large pool of students' answers on many items. For instance, Wang *et al.* [17] use pattern responses of 2800 students on the 30 FCI items. In this case the proficiency is what is commonly measured by all the items [15].

Apart from the two-parameter model chosen here, standard models used to describe the P function are the Rasch model—also called the 1PL model—or the 3PL model. Descriptions of these models and our reasons for the selection of the 2PL model are discussed in Appendix A.

In the framework of the latent trait modeling, an item is seen as an imperfect measuring tool of proficiency. The proficiency is a latent variable because it is not directly observed. What is observed is the answers of the students to the item. Note X a particular answer which could be true—if the student answers correctly—or false—if the student answers incorrectly. Hence X is a dichotomous categorical variable. Latent model is written as [14]

$$X = \begin{cases} \text{True} & \text{if } \theta + \epsilon \geq b, \\ \text{False} & \text{if } \theta + \epsilon < b, \end{cases} \tag{2}$$

where ϵ is the measurement error. This measurement error has a null mean and a standard deviation equal to $1/a$. The item difficulty is seen as a threshold value and the discrimination coefficient represents the quality of the measure. In the case of a null measurement error, the logistic item characteristic curve Eq. (1) reduces to an Heaviside function centered in $\theta = b$. In this case the student gives a correct answer if and only if its proficiency is greater than the difficulty of the item.

V. DISTRIBUTION OF STUDENTS PROFICIENCIES

The proportion of correct answers to a question depends on the distribution of the students' proficiencies. In latent trait modeling and IRT, it is often assumed that this latter is normally distributed. This assumption will be used in the following, i.e., $\theta \sim \mathcal{N}(\bar{\theta}, \sigma_\theta^2)$. Average proficiency $\bar{\theta}$ and standard deviation σ_θ are unknown. However, using a logistic approximation to the cumulative normal distribution [25], they can be related to the proportion of correct answers p (see Appendix B):

$$\frac{\bar{\theta} - b}{\sigma_\theta} \simeq 0.6 \frac{\sigma_Y}{\sigma_\theta} \ln\left(\frac{p}{1-p}\right), \tag{3}$$

where $\sigma_Y = \sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}$. The ratio between σ_Y and σ_θ depends on the measurement error of the item. Using data from Lasry *et al.* [26], this ratio was estimated to 1.09 (see Appendix C). Hence this ratio will be assumed to be equal to 1, i.e., measurement errors have a negligible effect. As a consequence, the parameters of the distribution of the

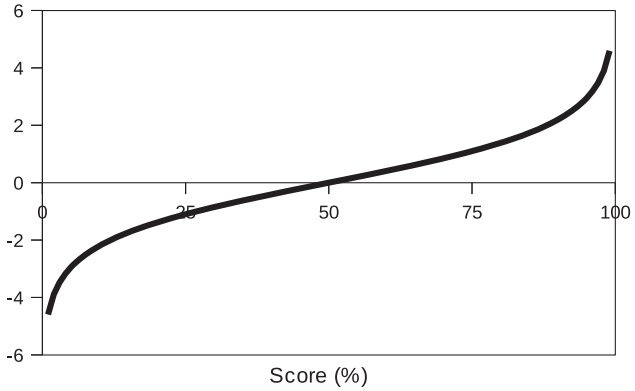

FIG. 1. The logit function. Vertical abscissa is $0.6 \ln[p/(1-p)]$, where p is given by the horizontal abscissa.

students' proficiencies are linked to the proportion of correct answers by

$$\frac{\bar{\theta}-b}{\sigma_\theta} \simeq 0.6 \ln\left(\frac{p}{1-p}\right). \tag{4}$$

The right-hand side is 0.6 times the logit of p defined by $\text{logit}(p) = \ln[p/(1-p)]$. The logit function transforms a probability between 0 and 1 in a value between $-\infty$ and $+\infty$ (see Fig. 1).

VI. EFFECT SIZE

Let us go back to our initial problem. Before the pedagogical intervention, the proficiencies of students are noted θ_1 and assumed to be normally distributed: $\theta_1 \sim \mathcal{N}(\bar{\theta}_1, \sigma_1^2)$. After the intervention, their proficiencies are θ_2 and are also assumed to be normally distributed: $\theta_2 \sim \mathcal{N}(\bar{\theta}_2, \sigma_2^2)$. Applying Eq. (4) to pre- and postscores leads to

$$\frac{\bar{\theta}_2 - \bar{\theta}_1}{\sigma_1} = 0.6\left(\frac{\sigma_2}{\sigma_1}\text{logit}(p_2) - \text{logit}(p_1)\right). \tag{5}$$

The ratio σ_2/σ_1 is unknown. However, estimations show that it can be assumed to be close to 1 (see Appendix D). Hence, assuming $\sigma_2 \simeq \sigma_1 = \sigma$, Eq. (5) gives the Cohen's d effect size [27]:

$$d = \frac{\bar{\theta}_2 - \bar{\theta}_1}{\sigma} = 0.6 \ln\left(\frac{p_2}{1-p_2}\frac{1-p_1}{p_1}\right). \tag{6}$$

Equation (6) estimates the difference of means of two normal distributions using the areas under the curves that are beyond a threshold value, as illustrated in Fig. 2. The horizontal axis represents students' proficiency. The Gaussian is the proportion of students with a given proficiency. The vertical dotted line corresponds to the difficulty of the item. Students with a higher skill level than the item's difficulty correctly answered the question. The proportion of correct answers is therefore represented by the gray area on the right of the vertical dotted line. In the second vote, the distribution of students' proficiency shifted to the right, increasing the proportion of students with higher proficiency than the item's difficulty. Equation (6) therefore makes it possible to evaluate the offset between the two distributions from the areas under the respective curves.

The Cohen's d effect size is perhaps the most commonly used effect size metric and is broadly used in education research and many other fields. Estimating the change of scores in terms of the Cohen's d effect size allows us to use the rules of thumb for interpreting the efficiency of the intervention in terms of very small to huge [27,28]. Figure 3 plot isovalues of the effect size corresponding to these

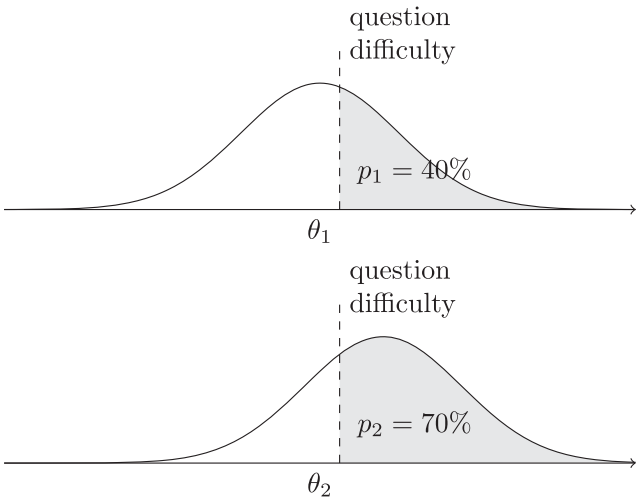

FIG. 2. Distributions of students' proficiencies. Top, before the intervention; bottom, after the intervention. The vertical dashed line is the question difficulty. Proportions of correct answers correspond to gray areas.

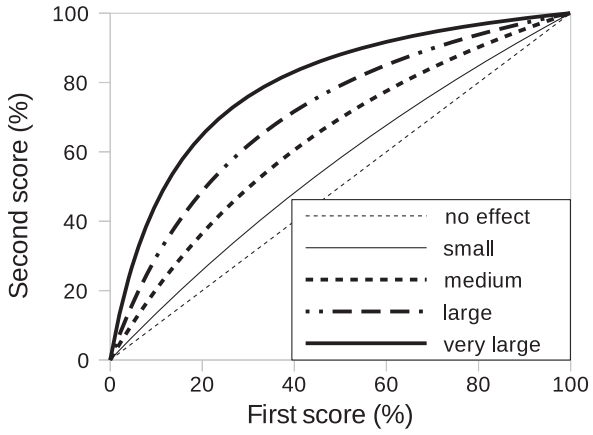

FIG. 3. Isovalues of the Cohen's d effect size (Eq. (6): no effect ($d = 0$), small ($d = 0.2$), medium ($d = 0.5$), large ($d = 0.8$), and very large ($d = 1.2$).

degrees in the (p_1, p_2) plot, enabling us to have a quick estimate of the value of the effect size of an intervention. Moreover, effect size enables comparisons to other teaching methods [29] and should be used rather than other methods such as the Hake's g [30].

VII. TETRACHORIC CORRELATION COEFFICIENT

The correlation coefficient between the students' answers before and after the intervention can enlighten us on the nature of the votes. For instance, if all students have the same increase of proficiency, the correlation coefficient between $\bar{\theta}_1$ and $\bar{\theta}_2$ is equal to 1. However, one can expect that depending on various inhomogeneities, this increase of proficiency is not the same for all students,

TABLE I. Contingency table.

X1\X2	False	True
False	N_1	N_3
True	N_2	N_4

but fluctuations should remain small. On the opposite side, a low correlation coefficient between answers at the first and the second vote should warn us that something wrong could have happened. Maybe the students have voted randomly at one of the two votes—or both of them—or maybe the measurement errors are huge (see Appendix E), perhaps due to a poor writing of the question. As a consequence, a correlation coefficient close to 1 is a good thing—due to measurement errors one cannot expect a coefficient greater than 0.85 (see Appendix C)—and a small correlation coefficient indicates that the results of the votes should be interpreted cautiously. Section IX shows that a value of 0.3 can be used as a rough threshold.

The traditional Pearson correlation coefficient cannot be used with students' answers because they are categorical variables (either true or false) and not continuous ones. The tetrachoric correlation coefficient has been especially developed to deal with categorical data explained by latent variables [31]. It is a product-moment correlation between two unobserved quantitative variables that have each been measured on a dichotomous scale. It assumes that the contingency table of the observed variables, here X_1 and X_2, comes from two correlated random variables that are normally distributed, here Y_1 and Y_2, where $Y_i = X_i + \epsilon$. The thresholds and the correlation coefficient between Y_1 and Y_2 are estimated using the maximum likelihood (ML) technique [31].

The statistical software R includes a dedicated library named polychor to estimate this correlation coefficient perform using the ML method. However, depending on the purpose, an estimate of this correlation coefficient can be sufficient and is obtained from the contingency table [32]:

$$\rho \simeq \cos\left(\frac{\pi}{1+\sqrt{(N_1N_4)/(N_2N_3)}}\right), \tag{7}$$

where N_1, N_2, N_3, and N_4 are the components of the contingency table (cf. Table I). Their sum is equal to the total number of students: $N_1 + N_2 + N_3 + N_4 = N$.

VIII. CONFIDENCE INTERVALS

The evaluation of d and ρ using Eqs. (6) and (7) are made from the observed proportions of correct answers and the corresponding contingency table. Hence they are estimators of the true values based on the theoretical proportions obtained only for an infinite number of students. Suppose that they are calculated using a sample of 10 students. It is clear that the value obtained will be a poor indication of the real effect size of your pedagogical intervention. For research applications, such as determining which one of two methods leads to the greatest effect size or if an intervention has a non-null effect size, confidence intervals are needed. A confidence interval is an interval that might contain the true value of the estimated parameter that would have been obtained with an infinite number of students drawn from a theoretical distribution—in our case the Gaussian distributions of students' proficiencies (shown in Fig. 2). Confidence intervals are given with a given confidence level. For instance, a 95% confidence interval has a 95% chance to contain the true value. This means that if the same experiment—first vote, pedagogical intervention, and second vote—was repeated on numerous samples of students, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter—d or ρ—would tend toward 95%. The greater the confidence level is, the wider the confidence interval. A 95% confidence interval is included in the 99% confidence interval calculated from the same sample.

For a given number of students N, the observed proportions of correct answers p_1^{obs} and p_2^{obs} follow approximative normal distribution laws of means p_i and variances $p_i(1-p_i)/N$ (due to the normal approximation of the binomial distribution when $N \gg 1$). Assuming that the standard deviation of p_i^{obs} remains law behind p_i, the logit function can be linearized around p_i. Hence the observed logit function L_i^{obs} of p_i^{obs} follows an approximative normal distribution of mean $\text{logit}(p_i)$ and variance:

$$\sigma_{L_i}^2 = 1/(Np_i(1-p_i)). \tag{8}$$

The observed effect size also follows an approximative normal distribution. Hence a 95% confidence interval of d can be estimated by

$$[d^{\text{obs}} - 1.96\sigma_d, d^{\text{obs}} + 1.96\sigma_d], \tag{9}$$

where σ_d is the standard deviation of d, estimated using the observed values p_i^{obs}:

$$\sigma_d^2 = 0.6^2(\sigma_{L_1}^2 + \sigma_{L_2}^2 - 2\rho_L\sigma_{L_1}\sigma_{L_2}). \tag{10}$$

The correlation coefficient ρ_L between L_1^{obs} and L_2^{obs} is unknown. Numerical simulations were performed in order to estimate it. For a given set of values of the number of students N, the correlation coefficient ρ, a proportion p_1, and an effect size d—or, equivalently, a proportion p_2 -, N samples were drawn from the bivariate normal distribution (θ_1, θ_2) assuming equal variances, a correlation coefficient ρ and expected values given by Eq. (4)—the question difficulty b was arbitrarily set to 0. Then the observed students' answers X_1 and X_2 and the corresponding contingency table were calculated using Eq. (2) assuming

a null measurement error. Those steps were repeated 10 000 times in order to calculate the correlation coefficient ρ_L. In order to cover a wide range of all possible parameters, this process was repeated for $N = 200$ and 1600; $\rho = 0.3$, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9; $p_1 = 0.3$, 0.4, 0.5, 0.6, and 0.7; and $d = 0.2$, 0.5, and 1.2, leading to a total of $10\,000 \times 210$ simulations performed. Results show that the correlation coefficient ρ_L between L_1^{obs} and L_2^{obs} is very close to the correlation coefficient ρ_p between p_1^{obs} and p_2^{obs}: the absolute difference between these two correlation coefficients is less than 0.005. Moreover, the correlation coefficient ρ_p is given by the traditional Pearson correlation coefficient between X_1 and X_2—assuming a value of 0 for an incorrect answer and 1 for a correct answer. Hence, ρ_L is given by

$$\rho_L = \frac{N_4/N - p_1 p_2}{\sqrt{p_1(1-p_1)p_2(1-p_2)}}, \tag{11}$$

where N_4 is a component of the contingency table (cf. Table I). When calculating the confidence interval with Eq. (9), σ_{L_1}, σ_{L_2}, and ρ_L are evaluated using the observed values p_1^{obs} and p_2^{obs} in Eqs. (8) and (11).

In order to validate Eq. (9), numerical simulations were performed in the same way as previously. For fixed values of N, ρ, p_1, and d, the number of times where the true value of d falls in the confidence interval was counted in 5000 simulations. This process was repeated varying the values of ρ, p_1, and d as previously (N was set to 200). The true effect size d is on average 94.9% of the times in the confidence interval, validating the approach.

Equation (9) shows that the size of the confidence interval is proportional to $1/\sqrt{N}$. As an illustrative example, a population of $N = 200$ students was generated, with $p_1 = 40\%$, $p_2 = 70\%$, and $\rho = 0.7$. The observed proportions of correct answers were $p_1^{\text{obs}} = 43\%$ and $p_2^{\text{obs}} = 74\%$. The corresponding observed effect size was 0.81 with a 95% confidence interval of [0.61, 1.01]. The true effect size is 0.75 falling in the confidence interval. Another population of 400 students was generated using the same values for p_1, p_2, and ρ. The observed scores were $p_1^{\text{obs}} = 37\%$ and $p_2^{\text{obs}} = 71\%$, leading to an observed effect size of 0.85 with a 95% confidence interval of [0.71, 1.00]. Once again, the true effect size falls in this (smaller) confidence interval.

Confidence intervals of ρ using the maximum likelihood method are outputs of the R library. However, its role is only to warn the instructor to conduct more investigations and it is not a guarantee of a good or bad change of scores. Hence confidence intervals are not necessarily needed and the rough estimation given by Eq. (7) could be sufficient.

IX. APPLICATION TO PEER INSTRUCTION

Let us consider a practical application case of an introductory mechanics course taught in a French École d'Ingénieurs from January 2016 to April 2016. This course was composed of ten lectures using Peer Instruction followed by tutorials in small groups. During lectures, all 190 students had a personal clicker. On average, 3.7 full Peer Instruction processes (first individual vote, peer discussion, and second individual vote) were performed during a lecture, leading to a database of 37 pre- and postdiscussion scores to the ConcepTests. Because of a participation rate at each vote around 80%, the average number of students answering twice to a ConcepTests is 125 (with a standard deviation of 24).

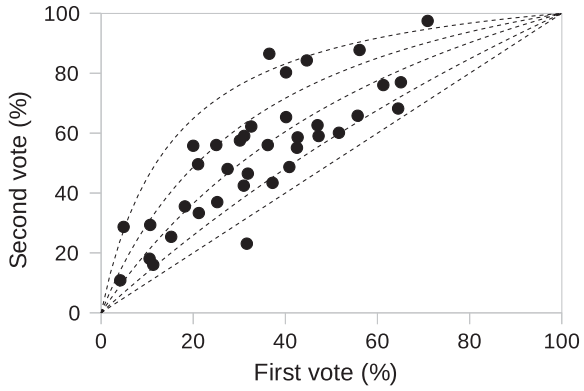

FIG. 4. Results of the peer discussion for the 37 ConcepTests. Each point is the proportion of correct students' answers to a given ConcepTest. Lines are isovalues of the Cohen's d effect size (no effect, small, medium, large, and very large effect).

Results of votes are plotted in Fig. 4. For each ConcepTests, the effect size is estimated from the pre- and postdiscussion scores using Eq. (6). The average effect size is 0.67, a value between the medium (0.5) and large (0.8) limits. One-quarter of the effect sizes are below 0.3 and one-quarter above 0.75. Figure 4 shows that almost all discussions lead to an effect size between small and large, with three above a very large effect.

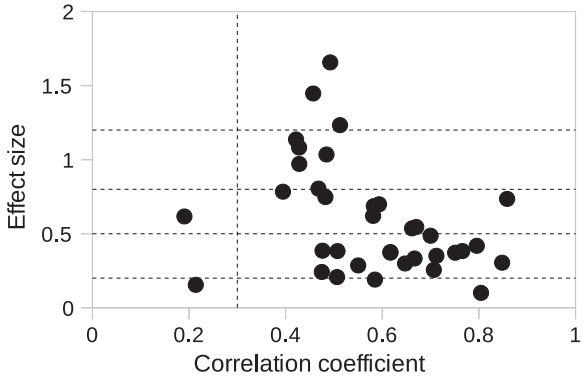

FIG. 5. Effect size as a function of the correlation coefficient estimated using ML for the 37 ConcepTests. Horizontal lines correspond to values for small (0.2), medium (0.5), large (0.8), and very large (1.2) effect sizes. The vertical line corresponds to $\rho = 0.3$.

From students' answers to items, the polychoric correlation coefficients were calculated using both the maximum likelihood method and the approximative value using Eq. (7). Both methods led to similar results (see Appendix F). Efficiency of the discussion process to all items are plotted in a d-ρ diagram in Fig. 5. Most efficient discussions are at the top of the diagram. One item lead to a negative effect size and is not represented on this plot. As seen in the diagram, two items led to a poor correlation coefficient around 0.2. Hence for these two items, there is a high probability that students had voted randomly at one of the two votes—or both of them—or that students had not completely understood the questions. From these results, a threshold value of $\rho = 0.3$ is suggested in order to detect from abnormal answers.

X. CONCLUSION

Assuming that the observed scores to a given item are explained by a latent distribution, the Cohen's d effect size was expressed in terms of the observed scores before and after the intervention has occurred [Eq. (6)]. One of the main advantages of this evaluation is that this effect size has good measurement properties: it can be used to compare different kinds of interventions—whether the scores at the first vote were the same or not—and to calculate the ratio of efficiency between two different interventions. Moreover, it is broadly used in education research and many other fields. While they were many assumptions used to derive Eq. (6), I advocate that it should be used to quantify the change of the scores instead of other indicators, such as the RD, RR, OR or the Hake's g because (i) it is theory grounded on item response theory and probabilistic thinking, (ii) comparisons with other educational studies using the Cohen's d effect size can be performed and (iii) using the Cohen's d effect size along with confidence intervals belongs to the recommended practice of "the new statistics" [33].

Precision of the estimated effect size from the observed data is performed using confidence intervals. Equation (9) gives the 95% confidence interval but other levels of confidence can be calculated by changing the 1.96 value to the corresponding one. The confidence interval is needed in order to demonstrate that a pedagogical intervention has a non-null effect. It is also required to classify two interventions: their confidence intervals should not overlap.

Moreover, the polychoric correlation coefficient was also suggested to detect abnormal behaviors when its value is below 0.3. This threshold value is, at the moment, only a rule of thumb grossly estimated from specific data. Consequently, it should be seen as a first step toward the definition of a more precise rule.

These two indicators are easy to calculate from students' answers and can be implemented in any software, leading to an automatic estimation of the effect of the pedagogical interventions. Results can be plotted in a d-ρ diagram to compare efficiency of different pedagogical interventions.

Finally, while the paper has been focused on dichotomous scores, the approach can be easily extended to partial scores (see Appendix G).

APPENDIX A: RATIONALE FOR CHOOSING THE 2PL MODEL

The 1PL assumes that the discrimination power a is set to an arbitrary fixed value—usually equal to 1. When analyzing a full test composed of multiple items such as a concept inventory, the 1PL and 2PL models differ because the 1PL model assumes that all items have the same discrimination power—i.e., $a = 1$ for each item—while the 2PL model allows them to have different values. However, in our case, only one item is considered. Hence the 1PL model and the 2PL model are equivalent due to the invariance property [34].

The 3PL model extends the 2PL one by adding a guessing parameter. In the 2PL model, when the proficiency θ goes to minus infinity, the probability to give a correct answer goes to zero, as stated in Eq. (1). In the 3PL model, this probability goes to a constant value—greater than zero but lower than 1—called the guessing parameter. While being attractive, assumptions and the interpretation of this model have been criticized [35,36]. Moreover, assuming a 2PL model still allows us to detect the presence of random votes as shown in Sec. VII. Another reason for not selecting the 3PL model is that it is not compatible with standard latent trait modeling—Eq. (2) does not hold. As a consequence, the tetrachoric correlation coefficient cannot be calculated. And, finally, a good item should not lead to guessing behaviors because all possible answers reflect common students' answers so that each student votes for the answer they believe to be correct. So guessing behaviors are expected to be infrequent. All these reasons led us to select the 2PL model.

APPENDIX B: RELATION BETWEEN THE OBSERVED PROPORTION OF CORRECT ANSWERS AND THE POPULATION PARAMETERS

The proportion of correct answers to a question depends on the distribution of the students' proficiencies. In latent trait modeling and IRT, it is often assumed that this latter is normally distributed. This assumption will be used in the following; i.e., $\theta \sim \mathcal{N}(\bar{\theta}, \sigma_\theta^2)$. Average proficiency $\bar{\theta}$ and standard deviation σ_θ are unknown. However, they are related to the observed proportion of correct answers. A simple relationship between those variables is derived in this section.

Let us note $Y = \theta + \epsilon$, which represents the imperfect measurement of proficiency θ with the error ϵ. Following Eq. (2), the proportion of correct answers is given by the proportion of students that have a Y greater than b:

$$p = \int_b^{+\infty} f_Y(Y)dY, \tag{B1}$$

where f_Y is the probability distribution function of Y. No exact mathematical expression can be found for the distribution f_Y because Y is the sum of two variables with different probability density functions: θ, which is normally distributed, and ϵ which follows a logistic distribution. However, the logistic distribution is very close to the normal distribution [25]. Hence the 2P Logistic model Eq. (1) can be replaced by the 2P Normal Ogive model [37,38] that assumes that the error term ϵ is normally distributed. As a consequence, Y is also normally distributed: $Y \sim \mathcal{N}(\bar{\theta}, \sigma_Y^2)$, with $\sigma_Y = \sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}$. Hence, p is given by

$$p = \frac{1}{\sqrt{2\pi}\sigma_Y} \int_b^{+\infty} e^{-\frac{(Y-\bar{\theta})^2}{2\sigma_Y^2}} dY = 1 - \Phi\left(\frac{b-\bar{\theta}}{\sigma_Y}\right), \tag{B2}$$

where Φ is the cumulative distribution function of the standard normal distribution. This function can be approximated using a logistic function [25]:

$$\Phi(x) \simeq \frac{1}{1+e^{-1.7x}}. \tag{B3}$$

Reporting Eq. (B3) into Eq. (B2) leads to

$$p \simeq \frac{1}{1+\exp[-(1.7/\sigma_Y)(\bar{\theta}-b)]}. \tag{B4}$$

The remaining Eq. (B4) leads to

$$\frac{\bar{\theta}-b}{\sigma_\theta} \simeq 0.6\frac{\sigma_Y}{\sigma_\theta}\ln\left(\frac{p}{1-p}\right), \tag{B5}$$

where $0.6 \simeq 1/1.7$.

APPENDIX C: ESTIMATION OF THE MEASUREMENT ERROR

In order to estimate the measurement error in Eq. (2), we use the data from Lasry *et al.* [26] who administrated the FCI twice in a row to 100 students. They reported the average contingency table for the 30 items (cf. Table II). From this contingency table, the tetrachoric correlation coefficient was estimated to 0.85 (95% CI = [0.82; 0.86]). This last one is equal to

TABLE II. Average contingency table of the test-retest for the 30 items of the FCI (data from Lasry *et al.* [26]).

X1\X2	False	True
False	43%	10%
True	8%	39%

$$\rho = \frac{\sigma_\theta^2}{\sigma_Y^2}, \tag{C1}$$

leading to $\sigma_Y/\sigma_\theta \simeq 1.09$.

APPENDIX D: ESTIMATION OF THE CHANGE OF VARIANCE

Estimating σ_1 and σ_2 from the contingency table alone is not possible because any values of those variances could lead to the same contingency table. Hence, in order to estimate the ratio σ_2/σ_1, other data are needed. Using the FCI, we estimated the proficiency—as measured by the FCI—of two groups of 1st year students, once at the beginning of a mechanical course, and the other at the end of the course, i.e., the end of the semester. The two courses both used Peer Instruction during lectures.

The first group was composed of 210 students and their initial pretest score was 12.4 (SD = 5.8). At the end of the semester, the average score was 15.7 (SD = 6). Using IRT, we estimate the distributions of θ before and after the course and found $\bar{\theta}_{\text{pre}} = -0.72$ and $\sigma_{\text{pre}} = 1.15$ and $\bar{\theta}_{\text{post}} = 0$ and $\sigma_{\text{post}} = 1$. The ratio between the two standard deviations is 0.87.

The second group was composed of 183 students. FCI average scores were 10.1 (SD = 3.9) at the pretest and 13.2 (SD = 4.4) at the post-test. Proficiencies were estimated to $\bar{\theta}_{\text{pre}} = -1.18$ and $\sigma_{\text{pre}} = 0.85$ and $\bar{\theta}_{\text{post}} = -0.42$ and $\sigma_{\text{post}} = 0.75$. The ratio between the two standard deviations is 0.88.

In both cases, the standard deviation remains close to 1 after one semester of teaching. This is an indication that this ratio could be close to 1 when looking only at the effect of a single small intervention—such as a discussion with peers.

APPENDIX E: CORRELATION COEFFICIENTS

The correlation coefficient between Y_1 and Y_2 is given by

$$\rho = \frac{\rho_\theta}{1+(\sigma_\epsilon/\sigma_\theta)^2} = \frac{\rho_\theta}{(\sigma_Y/\sigma_\theta)^2}, \tag{E1}$$

where ρ_θ is the correlation coefficient between θ_1 and θ_2. The correlation coefficient ρ is low if the correlation between θ_1 and θ_2 is low or if the measurement errors are huge (i.e., $\sigma_\epsilon \gg \sigma_\theta$).

APPENDIX F: VALIDITY OF THE APPROXIMATIVE ESTIMATION OF THE POLYCHORIC CORRELATION COEFFICIENT

The polychoric correlation coefficient was estimated from both the ML technique and Eq. (7) for the 37 items and results are plotted in Fig. 6. Both techniques lead to

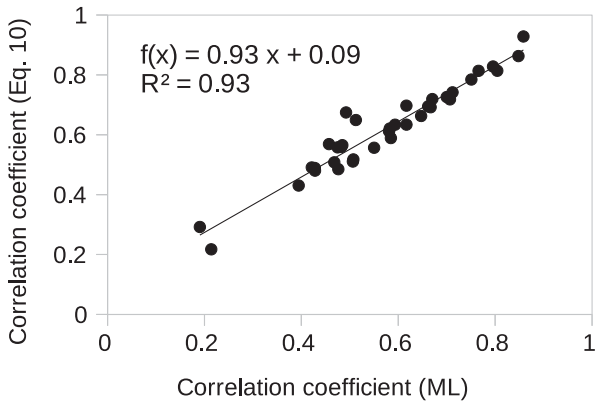

FIG. 6. Values of the correlation coefficient using Eq. (7) as a function of its value using the maximum likelihood method.

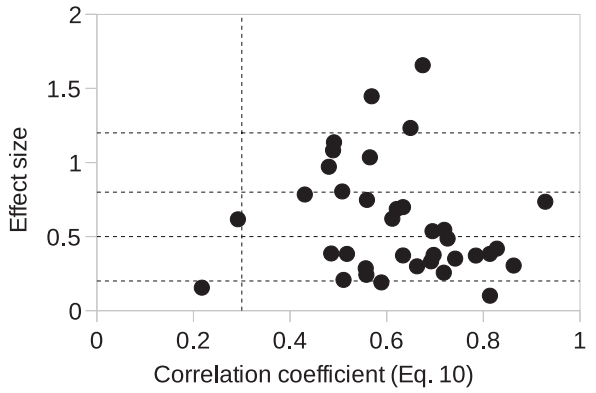

FIG. 7. Effect size as a function of the correlation coefficient estimated using Eq. (7) for the 37 ConcepTests. Horizontal lines correspond to values for small (0.2), medium (0.5), large (0.8), and very large (1.2) effect sizes. The vertical line corresponds to $\rho = 0.3$.

similar values. Equation (7) slightly overestimates the value given by the ML method, especially when the effect size is greater than 1.

The d-ρ diagram using Eq. (7) is plotted in Fig. 7. The threshold value of $\rho = 0.3$ can still be used to define correlation coefficients that are too low.

APPENDIX G: TAKING INTO ACCOUNT POLYTOMOUS RESPONSES

In the previous sections, only true or false answers were considered—i.e., dichotomous data. This section shows how to evaluate an effect size when multiple scores can be obtained on an item. This is the case, for instance, by taking into account partially correct responses or scoring rubrics, as shown in Table III, or even Likert scales.

A first approach could be to convert the answers to dichotomous scores, for instance, by setting the answer to true only if the highest score was obtained. While this could be a quick solution to the problem, more advanced methods can be used in order to obtain a more precise evaluation of the effect size. The graded response model (GRM) [39] and the partial credit model (PCM) [40] were designed to take into account graded response data. Graded response data consist of a score that is an ordinal number, typically ranging from 0 to M, where higher scores represent better performance on the item. Both the GRM and the PCM rely on logistic 2PL models and they are relatively similar.

TABLE III. Example of a scoring rubric for an item.

0	No response or the response is incorrect
1	Partially correct response
2	Completely correct response

TABLE IV. Example of the calculation of the effect sizes for partial scores between 0 and 3.

Score	P_k (pre)	P_k (post)	P_k^* (pre)	P_k^* (post)	d
0	20%	5%	100%	100%	⋯
1	40%	25%	80%	95%	0.90
2	31%	41%	40%	70%	0.75
3	9%	29%	9%	29%	0.84

The GRM models the probability to obtain a score equal or greater than a given value using a 2PL function:

$$P_k^*(S \geq k) = \frac{1}{1 + \exp\left[-1.7a(\theta - b_k)\right]}, \qquad \text{(G1)}$$

where S is the score obtained and k ranges from 1 to M. The probability to obtain a score equal to k is given by

$$P_k(S = k) = P_k^*(S \geq k) - P_k^*(S \geq k + 1). \qquad \text{(G2)}$$

The probability to obtain a score greater or equal to 0 is 1 and the probability to obtain a score greater or equal to $M + 1$ is 0. Hence, there are M different P_k^* functions and $M + 1$ unknown parameters: $a, b_1, \ldots, b_M$. As all the P_k^* are modeled using a 2PL model, Eq. (6) can be used to estimate the associated effect sizes, leading to M estimations of the true effect size. Finally, the average value of these M effect sizes can be used to get a final estimation of the true effect size. This is illustrated in Table IV with hypothetical scores obtained by some students. Scores lie between 0 and 3 and the corresponding fractions of students who obtained those scores are reported in columns 2 and 3. The 3 corresponding effect sizes are calculated using values of P_k^*.

A similar process can be used if a PCM is used instead of a GRM, with $P_k^* = P_k(S = k)/[P_k(S = k - 1) + P_k(S = k)]$.

Concerning the correlation coefficient between θ_1 and θ_2, the polychoric correlation coefficient generalizes the tetrachoric one to take into account multiple categorical

variables [31]. Hence it can be directly used with data using an R or Python library. However, assumptions behind this correlation coefficient are only in agreement with the GRM and not the PCM. Hence we recommend for the purpose of the study to use, preferentially, the GRM to calculate the effect size.

[1] J. Libarkin, Concept Inventories in Science, in National Research Evidence on Promising Practices in Undergraduate Science, Technology, Engineering, and Mathematics (STEM) Education Workshop 2, 2008, https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_072624.pdf.

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

[3] 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).

[4] A. Madsen, S. B. McKagan, and E. C. Sayre, Best practices for administering concept inventories, arXiv:1404.6500.

[5] C. H. Crouch and E. Mazur, Peer Instruction: Ten years of experience and results, Am. J. Phys. **69**, 970 (2001).

[6] A. P. Fagen, C. H. Crouch, and E. Mazur, Peer Instruction: Results from a range of classrooms, Phys. Teach. **40**, 206 (2002).

[7] N. Lasry, E. Charles, and C. Whittaker, Effective variations of peer instruction: The effects of peer discussions, committing to an answer, and reaching a consensus, Am. J. Phys. **84**, 639 (2016).

[8] F. Silvestre, P. Vidal, and J. Broisin, in *Design for Teaching and Learning in a Networked World*, edited by G. Conole, T. Klobučar, C. Rensing, J. Konert, and Lavoué, É (Springer International Publishing, Cham, 2015), Vol. 9307, pp. 339–351.

[9] S. Bhatnagar, N. Lasry, M. Desmarais, and E. Charles, *Proceedings of the European Conference on Technology Enhanced Learning* (Lyon, France, 2016).

[10] M. C. James and S. Willoughby, Listening to student conversations during clicker questions: What you have not heard might surprise you!, Am. J. Phys. **79**, 123 (2010).

[11] A. K. Wood, R. K. Galloway, J. Hardy, and C. M. Sinclair, Analyzing learning during Peer Instruction dialogues: A resource activation framework, Phys. Rev. ST Phys. Educ. Res. **10**, 020107 (2014).

[12] M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein, *Introduction to Meta-Analysis* (John Wiley & Sons, New York, 2011).

[13] S. S. Stevens, On the theory of scales of measurement, Science **103**, 677 (1946).

[14] A. Kamata and D. J. Bauer, A Note on the relation between factor analytic and item response theory models, Structural Equation Modeling: A Multidisciplinary J. **15**, 136 (2008).

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

[16] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).

[17] J. Wang and L. Bao, Analyzing force concept inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

[18] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the Force Concept Inventory into two equivalent half-length tests, Phys. Rev. ST Phys. Educ. Res. **11**, 010112 (2015).

[19] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).

[20] C. N. Cardamone, J. E. Abbott, S. Rayyan, D. T. Seaton, A. Pawl, and D. E. Pritchard, Item response theory analysis of the mechanics baseline test, AIP Conf. Proc. **1413**, 135 (2012).

[21] R. M. Talbot, Taking an item-level approach to measuring change with the force and motion conceptual evaluation: An application of item response theory, School Sci. Math. **113**, 356 (2013).

[22] M. Planinic, Assessment of difficulties of some conceptual areas from electricity and magnetism using the Conceptual Survey of Electricity and Magnetism, Am. J. Phys. **74**, 1143 (2006).

[23] L. Ding, Seeking missing pieces in science concept assessments: Reevaluating the Brief Electricity and Magnetism Assessment through Rasch analysis, Phys. Rev. ST Phys. Educ. Res. **10**, 010105 (2014).

[24] J. R. Buck, K. E. Wage, and M. A. Hjalmarson, *Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop (DSP/SPE), 2009* (IEEE, Bellingham, WA, 2009), pp. 726–730.

[25] S. R. Bowling, M. T. Khasawneh, S. Kaewkuekool, and B. R. Cho, A logistic approximation to the cumulative normal distribution, J. Indust. Eng. Management **2**, 114 (2009).

[26] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, The puzzling reliability of the Force Concept Inventory, Am. J. Phys. **79**, 909 (2011).

[27] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (L. Erlbaum Associates, Hillsdale, NJ, 1988).

[28] S. S. Sawilowsky, New effect size rules of thumb, J. Mod. Appl. Stat. Methods **8**, 597 (2009).

[29] J. Hattie, The applicability of Visible Learning to higher education, Scholarship Teach. Learn. Psychol. **1**, 79 (2015).

[30] J. M. Nissen, R. M. Talbot, A. N. Thompson, and B. Van Dusen, A comparison of Hake's g and Cohen's d for analyzing gains on concept inventories, arXiv:1612.09180.

[31] U. Olsson, Maximum likelihood estimation of the polychoric correlation coefficient, Psychometrika **44**, 443 (1979).
[32] N. J. Castellan, On the estimation of the tetrachoric correlation coefficient, Psychometrika **31**, 67 (1966).
[33] G. Cumming, The New Statistics, Psychol. Sci. **25**, 7 (2014).
[34] A. A. Rupp and B. D. Zumbo, Understanding parameter invariance in unidimensional IRT models, Educ. Psychol. Meas. **66**, 63 (2006).
[35] E. S. Martn, G. del Pino, and P. De Boeck, IRT models for ability-based guessing, Appl. Psychol. Meas. **30**, 183 (2006).
[36] M. von Davier, Is there need for the 3PL model? Guess What?, Meas. Interdiscip. Res. Perspect. **7**, 110 (2009).
[37] D. N. Lawley, On Problems connected with Item Selection and Test Construction, Proc. R. Soc. Edinburgh, Sect. A **61**, 273 (1943).
[38] F. Lord, A theory of test scores, Psychometric Monographs **7**, 84 (1952).
[39] F. Samejima, Estimation of latent ability using a response pattern of graded scores, Psychometrika Monograph Suppl. **34**, 100 (1969).
[40] G. N. Masters, A rasch model for partial credit scoring, Psychometrika **47**, 149 (1982).
[41] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.14.020116 for (i) An R program that calculates the effect size, confidence interval and correlation coefficient from students' responses to a question. (ii) Data from students' responses to the 37 questions to generate Figures 4, 5 and 7. (iii) The R program which made it possible to verify the formula of the confidence interval. This program simulates fictitious student responses.