Force Concept Inventory-based multiple-choice test for investigating students' representational consistency

Pasi Nieminen, Antti Savinainen, and Jouni Viiri
Department of Teacher Education, University of Jyväskylä, Jyväskylä FIN-40014, Finland
(Received 14 April 2010; published 25 August 2010)

This study investigates students' ability to interpret multiple representations consistently (i.e., representational consistency) in the context of the force concept. For this purpose we developed the Representational Variant of the Force Concept Inventory (R-FCI), which makes use of nine items from the 1995 version of the Force Concept Inventory (FCI). These original FCI items were redesigned using various representations (such as motion map, vectorial and graphical), yielding 27 multiple-choice items concerning four central concepts underpinning the force concept: Newton's first, second, and third laws, and gravitation. We provide some evidence for the validity and reliability of the R-FCI; this analysis is limited to the student population of one Finnish high school. The students took the R-FCI at the beginning and at the end of their first high school physics course. We found that students' ($n=168$) representational consistency (whether scientifically correct or not) varied considerably depending on the concept. On average, representational consistency and scientifically correct understanding increased during the instruction, although in the post-test only a few students performed consistently both in terms of representations and scientifically correct understanding. We also compared students' ($n=87$) results of the R-FCI and the FCI, and found that they correlated quite well.

DOI: 10.1103/PhysRevSTPER.6.020109 PACS number(s): 01.40.−d, 45.20.D−

I. INTRODUCTION

The role of multiple representations in learning is an important topic in the field of educational research [1,2]. Multiple representations (e.g., text, diagram, graph and equation) are often required for the understanding of scientific concepts and for problem solving. Multiple representations have many functions in learning, which Ainsworth [3–5] divides into the three parts:

(1) *To complement other representations*. Representations may differ either in the information each expresses or in the processes each supports. A single representation may be insufficient to carry all the information about the domain or be too complicated for learners to interpret if it does.

(2) *To constrain other representations*. For instance, graphs can be used to guide the interpretation of equations.

(3) *To construct a more complete understanding*. As when students integrate information from more than one representation.

Even though using multiple representations in teaching has great potential benefits, it can also jeopardize the learning process due to an increased cognitive load [7]. There are a number of cognitive tasks that students have to perform to cope successfully with multiple representations: they must learn the format and operators of each representation, understand the relation between the representation and the domain it represents, and understand how the representations relate to each other [8].

The importance of multiple representations has also been reported in physics education research. Van Heuvelen and Zou [9] offer several reasons why multiple representations are useful in physics education: they foster students' understanding of physics problems, build a bridge between verbal and mathematical representations, and help students develop images that give meaning to mathematical symbols. These researchers also argue that one important goal of physics education is helping students to learn to construct multiple representations of physical processes, and to learn to move in any direction between these representations. Furthermore, it has been pointed out that in order to thoroughly understand a physics concept, the ability to recognize and manipulate that concept in a variety of representations is essential [10].

There are also studies concerning multiple representations in problem solving [9,11–14]. Other studies show that the representational format in which the problem is posed affects student performance [15–17]. This effect has also been observed when computer-animated and static (paper and pencil) versions of the same problem were administered [18]. Both the context and the representation affect students' responses: the student might be able to apply a concept in a familiar context using a certain representation but fail when the context or the representation is changed [19].

Several research-based multiple-choice tests have been developed for evaluating students' conceptual understanding in the domain of introductory mechanics, the most widely used being perhaps the Force Concept Inventory (FCI) [20–22]. The FCI addresses several representations in a variety of contexts but it does not provide a systematic evaluation of students' ability to use multiple representations when the context is fixed. The existing tests like the FCI are limited in that they do not permit comprehensive evaluation of students' skills in using multiple representations. This deficiency led us to develop a multiple-choice test—the Representational Variant of the Force Concept Inventory (R-FCI) [23]—to evaluate students' representational consistency, i.e., their ability to use different representations consistently (scientifically correctly or incorrectly) between isomorphic (with the context and content as similar as possible) items.

In this paper we present the rationale and structure of the R-FCI: the test is based on the 1995 version of Force Concept Inventory [21]. First-year Finnish high school students' ($n=168$) pre- and post-test data of the R-FCI are analyzed

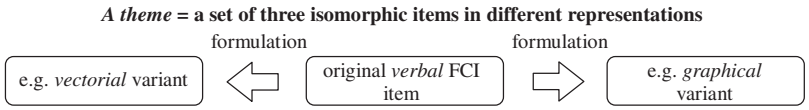


FIG. 1. A theme is a set of three isomorphic items (differing only in their representations).

from the perspectives of evaluating students' representational consistency, assessing students' learning gain of the force concept, and studying the effect of a representational format for students' performance in the isomorphic items. The post-test data are also used for calculating five statistical indices [24] from classical test theory (see Sec. III E). To investigate the validity of the R-FCI in this student population, we analyzed students' ($n=104$) written justifications for their multiple-choice answers. It is worth noting, however, that these validity and statistical index analyses are limited, since they are based on the data from one Finnish high school taught by one teacher (author AS). Finally, in order to study the suitability of the R-FCI for evaluating students' understanding of the force concept, we compared students' ($n=87$) results of the R-FCI and the FCI.

II. METHODS

A. Structure of the R-FCI

An earlier version of the test (previously called the Representation Test) was developed in 2006 [17], consisting of 21 items concerning gravitation and Newton's third law. The test was then improved and expanded until in 2007 the final version consisted of 27 items concerning gravitation and Newton's first, second and third laws.

The R-FCI is based on nine items taken from the 1995 version of the FCI: [21] items number 1, 4, 13, 17, 22, 24, 26, 28, and 30. The original verbal multiple-choice alternatives of the FCI items were redesigned using various representations. The purpose was to form isomorphic variants, keeping the physical concept and context of the items as similar as possible. For each of the nine FCI items, two new isomorphic variants were formulated in different representations. We use the term theme for the set of three isomorphic items that consist of an original FCI item and two isomorphic variants (see Fig. 1). Themes are named according to an original FCI item. Theme 4 (T4), for example, refers to item 4 in the FCI. There are altogether nine themes in the R-FCI, so the test contains 27 items in total. Table VIII in Appendix A shows the themes of the R-FCI, a concept and a context of a theme that was dealt with, and representations in which items of a theme were posed.

All the original FCI items (themes) were not included in the R-FCI because the test would have become too long. Items were selected on the grounds of suitability for the formulation of various representations. In addition, we wanted the test to cover the essential dimensions of the force concept, all Newton's laws, and gravitation. As Table VIII shows, five different representational formats were used in the R-FCI. Representational formats of a certain theme were selected on the basis of suitability, as some representations are more natural than others for a particular context. For instance, in Newton's third law a vectorial representation depicts very accurately the essence of the law, whereas a motion map would not be appropriate. Figure 2 presents corresponding multiple-choice alternatives of theme 4 (T4) that are depicted via different representations. All items of T4 include identical verbal description of the question to be answered, with alternatives. The question is not presented here to preserve the confidentiality of the original FCI items.

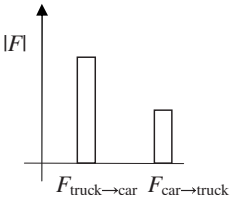
B. Participants and data collection

The participants of this study consisted of four groups of Finnish high school students: Phys1 2007 ($n=79$), Phys1 2008 ($n=64$), Pre-IB 2008 ($n=25$), and Phys2 2006 ($n=56$) (see Table I). Both Phys1 groups consisted of regular first-year students, and the Pre-IB group consisted of first-year students preparing themselves for the International Baccalaureate program. All the groups except the Pre-IB group were taught in different sections with 25–33 students per section: for instance, the Phys1 2007 and Phys1 2008 groups were each taught in three sections.

The first-year students (aged 16, $n=168$ altogether) were taking their first, compulsory, high school physics course which included a general introduction to physics, elementary kinematics and Newton's laws. The Pre-IB students studied in English using an American textbook [25], whereas all the others studied in Finnish using a Finnish textbook [26]. Despite having different textbooks, all the students had many common exercises addressing the use of multiple representations in kinematics and Newton's laws.

The Phys2 2006 group consisted of second-year Finnish high school students (aged 17, $n=56$) who had chosen to study physics beyond the compulsory course. The course dealt with kinematics and Newton's laws, and involved a lot of problem solving. These students had already had three physics courses before taking the R-FCI: the first was the one briefly described above, and the other two dealt with mechanical energy, thermophysics, and waves. Each course

Item 2, alternative (d).



Item 11, alternative (a).

The truck exerts a greater amount of force on the car than the car exerts on the truck.

Item 20, alternative (c).

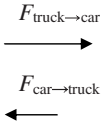


FIG. 2. Corresponding multiple-choice alternatives of theme 4 (T4) in the R-FCI. The representational formats of the alternatives are a bar chart (item 2), verbal (item 11) and vectorial (item 20). All three items include an identical original FCI question in the verbal form. The questions of bar chart and vectorial items include explanation of notations such as $F_{\text{truck}\rightarrow\text{car}}$.

TABLE I. Groups and students ($n=224$) who took the R-FCI as a pre- and post-test.

Group	Year	Number of students	Test version
Phys1	2007	79	Final (27 items)
Phys1	2008	64	Final
Pre-IB	2008	25	Final
Phys2	2006	56	Earlier (21 items)

involved 30 h of teaching time and was naturally delivered in Finnish.

We collected multiple-choice data using the R-FCI before and after teaching. We analyzed only the pre- and post-test scores of students who had (a) taken *both* the pre- and post-tests, and (b) answered *all* questions on the pre- and post-tests. The Pre-IB group had their pretest in Finnish, as their English was not strong enough at the beginning of the course. However, their post-test was in English. All other groups took the R-FCI in Finnish. All the first-year students answered the final version of the R-FCI (27 test items), whereas the second-year students (Phys2 2006) answered the earlier version of the R-FCI (21 test items). In addition to the doing the multiple-choice test, Phys1 2007 and Phys2 2006 groups were required to write down their justification for choosing their answer in the individual test item of the post-R-FCI test. These data were collected for validation of the test. Written justifications from the Phys2 group ($n=56$, the earlier version of the test) were used to investigate the validity of 14 common items in the earlier and final versions of the test; the written data for the remaining 13 items were collected in Phys1 in 2007 ($n=48$, the final version). So we gathered validation data from 104 students altogether. Moreover, the Phys2 group data were used only for validation of the test items; all other analyses were done using the final version data of the R-FCI,

In order to find out how suitable the R-FCI is for assessing students' understanding of the force concept, some FCI data were also collected and the results of the R-FCI and FCI were compared. The students ($n=87$) that took both tests were from Phys1 2008 and Pre-IB 2008.

All the groups were taught by one of the authors (AS), using interactive-engagement teaching methods with various representations; he has used these methods for many years (for details, see [27]). Furthermore, he made use of a specific representation (the Symbolic Representation of Interaction, SRI) to help students to perceive forces as interactions. [28] The SRI serves as a visual tool showing all the interacting objects and the nature of the interactions between them; it is similar to the system scheme used in the Modeling approach [29].

C. Data analysis

1. Analyzing R-FCI data

Next we describe three different analyses of the R-FCI data (see Fig. 3):

(A) The first analysis—arrow A in Fig. 3 and Sec. III A—made use of the theme structure of the test: this enables the evaluation of students' representational consistency by examining students' answers within a certain theme (see Fig. 2). Students exhibited representational consistency when all the answers in a given theme were consistently correct or consistently incorrect. Furthermore, students exhibited scientific consistency when all the answers in a given theme were correct in terms of both physics and representations. In this analysis, scientific consistency is considered a subconcept or a special case of representational consistency.

(B) In the second analysis—arrow B in Fig. 3 and Sec. III B—raw scores of the test were exploited. The R-FCI was administered before and after instruction. This made it possible to evaluate the average normalized gain [30,31], which was used as a rough measure of the development of students' understanding due to instruction. This is a common method in physics education research for utilizing data of multiple-choice tests.

(C) In the third analysis—arrow C in Fig. 3 and Sec. III C—the effect of the representational format on understanding the force concept was investigated: in other words, how the representation in which an item was posed affected students' performance on isomorphic items. This analysis was based on comparing averaged scores of isomorphic test items. This kind of examination was used by Meltzer [15] as well as Kohl and Finkelstein [16].

2. Categorization of consistency

We studied each theme separately and the data were analyzed from the perspective of the students' ability to use multiple representations consistently both in cases of representational consistency and scientific consistency, as explained above. For both representational and scientific con-

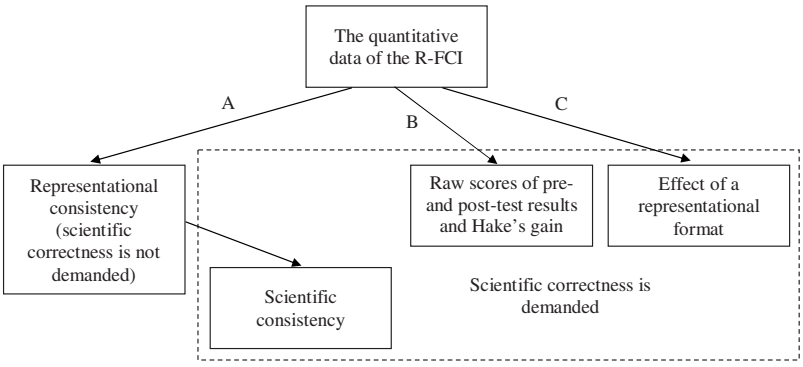


FIG. 3. Different ways for analyzing the R-FCI data.

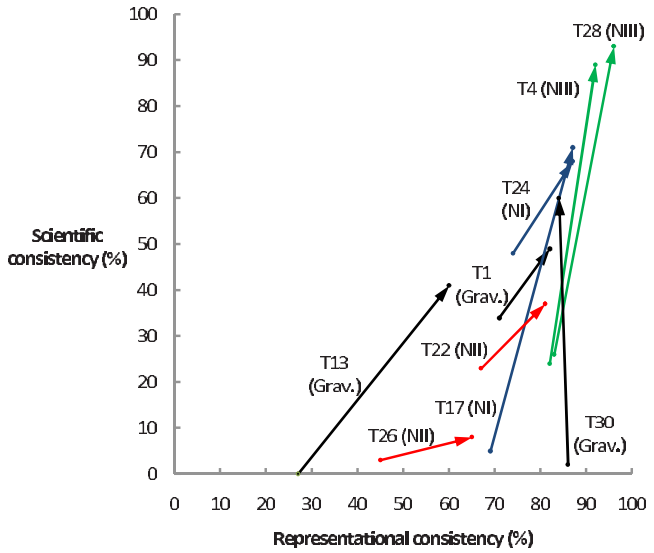


FIG. 4. (Color) Students' progress in representational and scientific consistency between pre- and post-tests. The starting point of an arrow shows pretest results and the head shows post-test results.

sistency, students' answers in a given theme were graded in the following way:

(i) Two points, if they had chosen corresponding alternatives in all three items of the theme.

(ii) One point, if they had chosen corresponding alternatives in two of the three items of the theme.

(iii) Zero points, if no corresponding alternatives in the items of the theme were selected (see Appendix B for examples of alternative options and designated points).

In order to evaluate students' representational and scientific consistency in the whole test, the average points for all the themes were calculated. This meant that a student's points for nine themes (see Table VIII in Appendix A) were added together and divided by nine, so the average was also between zero and two points. On the basis of the average points, students' representational and scientific consistency was categorized into three levels:

(i) Level I: an average of 1.7 (85% of the maximum) or higher indicates that thinking was consistent.

(ii) Level II: an average between 1.2 and 1.7 (60%–85% of the maximum) indicates that thinking was moderately consistent.

(iii) Level III: an average below 1.2 indicates that thinking was inconsistent.

The categorization rules are arbitrary, but they are similar to those used with the FCI. An FCI score of 60% is regarded as being the 'entry threshold' to Newtonian physics, and 85% as the "mastery threshold" [32].

III. RESULTS

A. Consistency of students' thinking

Students' representational and scientific consistency in each theme was studied and graded from zero to two points, as described above (see Table X in Appendix C; Figs. 4 and 5 are based on these numbers). Figure 4 shows students'

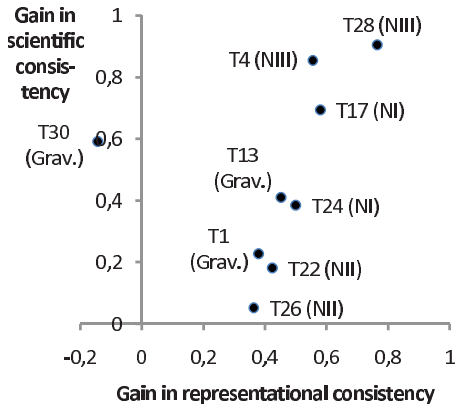


FIG. 5. (Color) Average normalized gains in representational and scientific consistency of themes.

progress in representational and scientific consistency between pre- and post-tests as measured by average points of consistency. The starting point of an arrow represents pretest results and the head represents post-test results. The percentages of average points vary considerably, depending on the theme. Students performed excellently in the post-test (arrow heads) in the context of Newton's third law (themes 4 and 28) in terms of representational and scientific consistency. On the other hand, Newton's second law (themes 22 and 26) seemed to be very difficult in terms of scientific consistency. However, in the post-test, especially in theme 26, the average representational consistency was very high (65%) compared with the scientific consistency (8%).

Newton's first law (themes 24 and 17) was handled better than Newton's second law. It is interesting that the post-test results of the themes were almost the same, although the contexts and representations of the themes were quite different. The pretest results of the themes concerning gravitation (themes 1, 13, and 30) varied considerably, but the post-test results (arrow heads), especially for themes 1 and 30, were more similar.

The directions of the arrows (Fig. 4) of T17, T4, T28, and T30 imply that students did better in scientific consistency than in gaining representational consistency. The *average* change in scientific consistency (39%) is higher than the *average* change in representational consistency (14%) (average points of all themes are shown in Table X in Appendix C). However, it should be noted that the pretest results for representational consistency were much higher than those for scientific consistency. For this reason, the average normalized gains for the points for scientific and representational consistency were calculated for all themes (see Fig. 5).

For example, in theme 28 the change in the points for scientific consistency is 67%, whereas the change in the points for representational consistency is only 13% (see Fig. 4). However, the difference in the average normalized gains is not so large: the average normalized gain in scientific consistency is 0.91, and the gain in representational consistency is 0.76. Theme 30 is quite interesting because the average normalized gain in scientific consistency is 0.59 whereas the average normalized gain in representational consistency is −0.14: students improved greatly in their scientific consistency, but the change in representational consistency was ac-

TABLE II. Levels of representational and scientific consistency ($n=168$).

		I (%) Consistent	II (%) Moderately Consistent	III (%) Inconsistent
Levels of representational consistency	Pretest	11	65	24
	Post-test	42	49	9
Levels of scientific consistency	Pretest	0	1	99
	Post-test	11	35	54

tually negative. However, the *averages* of the average normalized gains in representational consistency (0.43) and in scientific consistency (0.48) are quite similar. In themes 4, 28, 17, and 30 the average normalized gain in scientific consistency is higher than the average normalized gain in representational consistency, whereas in themes 26, 22, 1, 24, and 13 the situation is reversed.

Finally, in very simplified way, students were categorized to levels of representational and scientific consistency based on the average of points of consistency as previously described (see Categorization of consistency). In the post-test 42% of students could use representations consistently (Table II). However not many students (11%) thought consistently in terms of scientific consistency. This shows that mastering multiple representations does not guarantee the correct scientific understanding of physics concepts although it certainly is a prerequisite for that.

B. Raw scores

Table III shows the pre- and post-test raw score results and Hake's average normalized gain of 168 first-year students that took the final version of the R-FCI. The results are quite similar between the groups. Only the difference between the pretest scores of the Phys1 2007 and Pre-IB groups is nearly statistically significant (Mann-Whitney U test, $z=1.92$, $p=0.055$). This might be at least partially due to the fact that the Pre-IB students were specially selected for the International Baccalaureate program. However, there are no statistically significant differences in the post-test scores or average normalized gains.

In order to discover how suitable the R-FCI is for assessing students' understanding of the force concept, the R-FCI and FCI results of Phys1 2008 and Pre-IB students ($n=87$) were compared. As Table IV shows, the pre- and post-test scores are quite similar. The correlation coefficient between the scores of the pretests is 0.78, and that between the scores of the post-tests is 0.86. The correlations are high indicating a strong relationship between the FCI and R-FCI. It should be noted that the tests include the nine common items, which increases the correlations. If the common items are excluded from the FCI, the correlation coefficients between the 21 FCI items and the R-FCI are 0.60 for the pretests and 0.77 for the post-tests. The correlation coefficients are lower, but still fairly high.

Furthermore, the R-FCI contains isomorphic items (the same item occur three times), which may magnify the correlations in the analysis. If only the nine verbal R-FCI items (the original FCI items) are included in the analysis, the correlation coefficients between the nine R-FCI items and the 21 FCI items are 0.50 for the pretests and 0.74 for the post-tests. In this case the coefficient for the pretests is only moderate, but for the post-test it is still quite high.

Consequently, there is a strong relationship between the scores of the tests. A clean performance in the R-FCI predicts success in the FCI. The R-FCI can be considered to be quite a good tool for assessing students' understanding of the force concept, even though it does not include all the dimensions of the force concept that the FCI covers (for a discussion on the dimensions and representations of the FCI, see [33]). The average normalized gain of the R-FCI is higher than that of the FCI. This may be due to the fact that the R-FCI does not include all the items of the FCI.

TABLE III. Percentages of pre- and post-R-FCI raw score results and students' average normalized gain. Standard errors are in parentheses.

Group	Number of students	Pretest (%)	Post-test (%)	Gain
Phys1 2007	79	20(2)	61(3)	0.51
Phys1 2008	64	23(2)	60(2)	0.48
Pre-IB	25	28(3)	62(4)	0.47
All	168	22(1)	61(2)	0.49

C. Effect of a representational format

The R-FCI makes it possible to examine the effect of a representational format on students' performance in a certain context, i.e., the difference in correct answers between two isomorphic items of a certain theme. Figure 6 shows the percentages of correct answers in the themes with statistically significant differences between representations. For ex-

TABLE IV. Students' ($n=87$) pre- and post-test raw scores and average normalized gains of the R-FCI and FCI. Standard errors are in parentheses.

	R-FCI	FCI
Pretest (%)	24(2)	31(2)
Post-test (%)	61(2)	58(2)
Gain	0.48	0.40

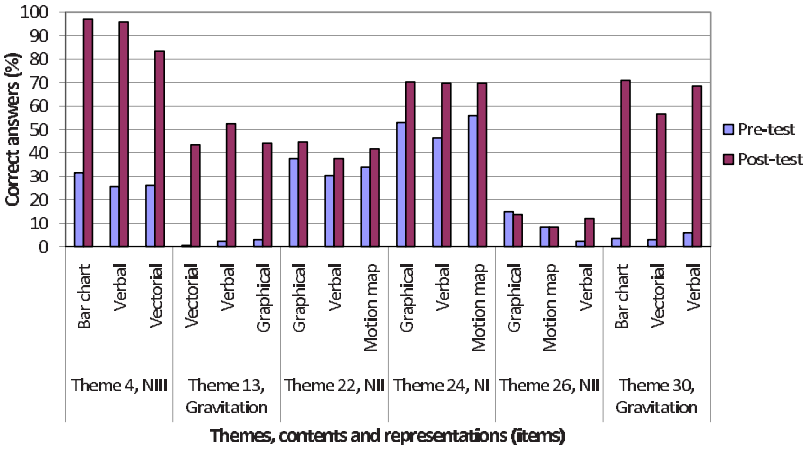


FIG. 6. (Color) The percentages of correct answers in the themes with statistically significant differences between representations.

ample, for theme 4 in the post-test (Newton's III law), the percentages of correct answers were 97% in the bar chart item, 96% in the verbal and 83% in the vectorial. When McNemar's tests were conducted, students were found to have performed better in the bar chart, $p<0.001$, and in the verbal, $p<0.001$, than in the vectorial item.

Table V shows all the statistically significant differences ($p<0.05$) when the correct answers of two representations of a theme were compared using McNemar's test. In the pretest, as in the post-test, there were statistically significant differences in six comparisons.

D. Validity

The items of the R-FCI are based on the nine items of the FCI. The reliability and validity of the FCI are well documented (for a review, see [34]). The physical contents (concepts and contexts) of the R-FCI items are almost the same as those of the original FCI items. Four of the R-FCI items are exactly the same as the original (including the representation). The others contain the same content depicted in different representation, and possibly some additional information that makes it possible to use different representations. For example, an original verbal item may not describe the directions of forces, but these might be depicted in the vectorial variant item.

We were interested in finding out how well the students ($n=104$) could justify their answers. For this purpose, students' written answers for each multiple choice in the post-test were examined by one of authors (PN). The criteria for correct explanations are shown in Table XI in Appendix D. Some explanations (themes 1, 4, 13, 28 and 30) had also been analyzed previously mainly by another author (AS), and the results [17] of these analyses were very consistent with those of author PN.

Table VI shows that 92% of the correct answers were accompanied by correct explanations, and 5% had partially correct explanations. Hence, the number of clear false positives is very small (3% of all correct answers). The number of false negatives is 7% of all incorrect answers. One possible reason for some false negatives might be that some students made mistakes in writing down the answers on the answer sheets; this seems likely in some cases where the verbal explanation was perfect and did not match the chosen answer at all.

TABLE V. Statistically significant differences ($p<0.05$) between correct answers of items (representations) in themes (McNemar's test). Abbreviations: BC=bar chart, Ver=verbal, Vec=vectorial, G=graphical, and MM=motion map.

Theme		Compared representations	p-value
T4	Pretest	BC vs Ver	0.021
		BC vs Vec	0.049
	Post-test	BC vs Vec	<0.001
		Ver vs Vec	<0.001
T13	Post-test	Vec vs Ver	<0.001
		Ver vs G	0.022
T22	Post-test	G vs Ver	0.017
T24	Pretest	G vs Ver	0.035
		Ver vs MM	0.006
T26	Pretest	G vs Ver	<0.001
		MM vsVer	0.013
T30	Post-test	BC vs Vec	<0.001
		Ver vs Vec	0.001

E. Statistical indices

Classical test theory provides different measures to evaluate multiple-choice tests and their items. Five measures, which were used in this study, have been often used in

TABLE VI. A cross tabulation of 104 students' written explanations and correct or incorrect classified multiple-choice answers.

	Correct answer (%)	Incorrect answer (%)
Correct explanation	92	7
Partially correct explanation	5	6
Incorrect explanation	3	87

TABLE VII. Evaluation of the R-FCI.

Evaluation measure	Values of the R-FCI	Desired values
Difficulty index (P)	Average of 0.61	0.3–0.9
Discrimination index (D)	Average of 0.30	$\geq$0.30
Point biserial coefficient (r_{pbi})	Average of 0.44	$\geq$0.20
Reliability index (r_{test})	0.87	$\geq$0.70
Ferguson's delta (δ)	0.97	$\geq$0.90

science education research [24]. Three of them were for item analysis: item difficulty level (P), discrimination index (D) and point biserial coefficient (r_{pbi}). Two measures were for test analysis: Kuder-Richardson reliability index (r_{test}) and Ferguson's delta (δ).

In order to calculate the measures, post-test data from students ($n=168$ altogether) in Phys1 (years 2007 and 2008) and Pre-IB were used. These students had taken the latest version of the R-FCI (see Table I). The values of reliable measures for the R-FCI are gathered in Table VII. This paper gives only a brief outline of the meaning of these measures. More detailed information and definitions of the measures can be found in Ding and Beichner [24].

1. Item difficulty index

The item difficulty index (P) indicates the difficulty of a certain test item. The value of the difficulty index varies between 0 and 1, with 0.5 being the best value. A widely used range for acceptable values is from 0.3 to 0.9 [35].

The difficulty index (P) values for each item of the R-FCI are shown in Table VIII in Appendix A. The values of P vary between 0.08 and 0.97, with most of the items having 0.4–0.7. Only three items were below 0.3 and five items were above 0.9. The averaged difficulty index is 0.61, which is exactly in the middle of the acceptable range of 0.3 to 0.9.

Three items (3, 12 and 21) which had item difficulty index values below 0.3 were not necessarily unsatisfactory test items: they are very difficult for high schools students in the first physics course. The items were three representational variants of theme 26 concerning Newton's II law. Five items (2, 8, 11, 17, and 26) which had item difficulty index values above 0.9 were not necessarily unsatisfactory either. The items were variants of T4 and T28, both concerning Newton's III law in the very similar context of collisions. Presumably, students learned this concept very well because the interaction was emphasized in teaching as briefly described in Sec. II B.

2. Item discrimination index

The item discrimination index (D) is a measure of the discriminatory power of an item. It indicates how well an item differentiates between high-achieving and low-achieving students. The simplest and most often used system to categorize students into high- and low-achieving groups is to divide them in two equal-sized groups based on the median of the students' total score. In our data, the median of the post-test total score was 16. Altogether, 168 students were divided into two groups of 84. The total score in the low group was below 16, and above 16 in high group. Eleven students had the median score of 16. They were randomly divided into two groups so that the size of both groups was 84. The values of the item discrimination index were calculated on the basis of this division.

The item discrimination index ranges from -1 to $+1$, where -1 is the worst and $+1$ the best value. If D were -1, everyone in the low group would have given correct responses while everyone in high group would have given incorrect responses. If D were $+1$, the situation would be reversed, and the discriminatory power of the item would be the best possible. The value of D must be positive or the item does not really operate in the correct way in the test. Generally, values of $D\geq 0.3$ have been considered satisfactory [36].

The item discrimination indexes of the R-FCI are shown in Table VIII in Appendix A. The discrimination index values of items ranged from 0.06 to 0.65. The values of 18 items were above 0.3, with most of the values (15) being 0.3–0.5. Hence, the majority of items of the R-FCI had quite satisfactory discriminatory power. The lowest discrimination indices occur for the items with either the highest (themes 4 and 28) or lowest (T26), which is most extreme, difficulty indices. The averaged discrimination index was 0.30, which was also in the satisfactory range.

3. Point biserial coefficient

The point biserial coefficient indicates how consistently an item measures students' performance in relation to the whole test. The desirable value for the point biserial coefficient is $r_{pbi}\geq 0.2$ [37]. The values of r_{pbi} are shown in Table VIII in Appendix A. They are above 0.2 except for one item, which supports the notion that almost all the items of the R-FCI are reliable and consistent. The average point biserial coefficient for the R-FCI is 0.44, which also supports this.

4. Kuder-Richardson reliability index

KR-20 (r_{test}) is an often used measure of internal consistency when test items are dichotomous (i.e., correct or incorrect) as in the R-FCI. [38] If a test has good internal consistency, different test items measure the same characteristic, and there are high correlations between individual test items.

The values of r_{test} range from 0 to 1. A widely used criterion for a reliable group measurement is $r_{test}\geq 0.7$. If a test is meant to be a measurement of individuals, the reliability index should be higher than 0.8 [39]. The reliability index of the R-FCI was 0.87, hence it could be considered as reliable for measuring student groups and single students alike.

5. Ferguson's delta

Ferguson's delta (δ) is a measure of the discriminatory power of a test. It takes into account how broadly students' total scores are distributed over the possible range. If a test

has a good discriminatory power, the distribution of total scores is wide ranging.

The values of delta range from 0 to 1. If δ is 0, all students score the same. If δ is 1, the distribution of scores is rectangular. If the delta value is higher than 0.9, a test is considered to have good discriminatory power [40]. Ferguson's delta for the R-FCI was 0.97, so the test had good discriminatory power.

IV. DISCUSSION AND CONCLUSIONS

In this study, our main goal was to develop a quantitative test for evaluating students' representational consistency. We have presented the structure and design of the R-FCI and have examined validity and reliability aspects.

Designing a valid and reliable multiple-choice tests of higher-order learning [41] is a demanding, multiphased, and time-consuming task [42–44]. We wanted the R-FCI to be valid and reliable, so the FCI was an excellent starting point. We chose nine items which were suitable for the formulation of various representations, and covered the basics of the Newtonian force concept. The R-FCI contains 27 items, and it is easy and rapid to use in a classroom. Students usually complete the test in about half an hour. For validation, we examined students' written explanations justifying their multiple-choice answers, and found good compatibility: analysis shows that 92% of the correct answers are accompanied by correct explanations. For statistical indices of merit, we used five measures for item and test analysis—see Sec. III E. The results (see Table VII) indicate that the R-FCI has sufficient discriminatory power, and is a reliable instrument for measuring single students and groups. It is important to note that the validation and reliability analysis was carried out using students in one Finnish high school. Furthermore, the students were taught by one teacher (author AS) using interactive-engagement teaching methods and multiple representations. Since it is reasonable to suppose that the validity and reliability measures might be affected by the student population and the teaching methods, we recognize that this paper provides only limited evidence of these attributes of the test. We will gather more data from different institutes in the future to investigate the general validity and reliability of the test.

We wanted to show how the results of the R-FCI could be analyzed to provide quite detailed information about student use of representations. The main purpose of the R-FCI is to evaluate students' representational consistency, meaning the students' ability to interpret multiple representations consistently, whether scientifically correctly or not. As a subconcept of representational consistency, scientific consistency can also be evaluated. In that case, the student's representational consistency and scientific understanding are evaluated. The answers of 168 high school students show that representational and scientific consistency depend to a large extent on the theme (concept and context). On average, 11% of the students were representationally consistent in the pretest, compared with 42% in the post-test. This can be considered to be acceptable progress after the first obligatory physics course. On the other hand, their scientific consistency was quite low: none of the students reached a consistent level in the pretest, and only 11% were consistent in the post-test. However, scientific consistency increased during the course, which indicates better understanding of the force concept. It is worth noting that students' ($n=168$) raw score averages improved quite a lot (from 22% to 61%), as indicated by the average normalized gain (0.49). These data suggest that scientific consistency in the use of representations is quite a demanding skill which is not directly predictable from the raw score average.

Previous research has shown that the ability to use multiple representations is an essential tool for doing physics [10]. Our results do not conflict with this, but they suggest that the ability to interpret multiple representations is a necessary but not sufficient condition for the correct scientific understanding of physics concepts.

We were interested in what the raw scores of the R-FCI imply concerning understanding of the force concept, so we compared students' ($n=87$) R-FCI and FCI results (see Table IV). The correlations between the pre-($r=0.78$) and post-tests ($r=0.86$) are strong. When the scores of nine verbal items of the R-FCI and 21 items of the FCI (no common items) are compared, the correlation is moderate for pretests items ($r=0.50$) and quite high for post-tests items ($r=0.74$). If the FCI is kept as the point of comparison, the results show that the R-FCI assesses quite well students' understanding of the force concept, although it does not include all the aspects of this concept.

The R-FCI makes it possible to study the effect of the representational format on students' performance when the context is fixed. Our results (see Sec. III C) support those of previous research [15–17] showing the effect of representational format on students' performance.

As pointed out above, one limitation of this study– and also a reason for future research—is that the data were collected from only one school and from the courses of one teacher. Hence, it would be fruitful to collect data from different teachers, schools and levels. Furthermore, it would be useful to discover how a student's ability to interpret multiple representations as measured by the R-FCI is related to their performance in open-ended multiple representation problems of the force concept when the construction of representations is demanded. We conclude that the R-FCI is a very promising and versatile tool for evaluating students' representational consistency and understanding of the force concept.

ACKNOWLEDGMENTS

This work has been supported by a grant from the Rector of the University of Jyväskylä and the Academy of Finland (Project No. 132316). We would like to thank Charles Henderson for commenting on this paper and David E. Meltzer for feedback concerning the R-FCI. We also thank Vivian Michael Paganuzzi for his invaluable assistance in revising the language of this paper.

APPENDIX A

Content of items and statistical indices (Table VIII).

TABLE VIII. Items, themes, concepts and context of the R-FCI and the results of item analysis.

Item	Theme	Concept	Context	Representation	Difficulty index	Discrimination index	Point biserial coefficient
1	T1	Gravitation	Falling balls	Verbal	0.52	0.37	0.53
2	T4	Newton III	Collision of cars	Bar chart	0.97	0.06	0.25
3	T26	Newton II	A woman pushes a box	Graphical	0.14	0.15	0.41
4	T13	Gravitation	A steel ball is thrown vertically upwards	Vectorial	0.43	0.65	0.69
5	T17	Newton I	An elevator	Verbal	0.72	0.32	0.44
6	T22	Newton II	A spaceship	Graphical	0.45	0.42	0.49
7	T24	Newton I	A spaceship	Graphical	0.70	0.45	0.57
8	T28	Newton III	Students sitting on office chairs push each other of	Verbal	0.95	0.08	0.30
9	T30	Gravitation	A tennis ball passes through the air after being struck	Bar chart	0.71	0.42	0.55
10	T1	Gravitation	Falling balls	Motion map	0.50	0.40	0.53
11	T4	Newton III	Collision of cars	Verbal	0.96	0.08	0.38
12	T26	Newton II	See item 3	Motion map	0.08	0.10	0.27
13	T13	Gravitation	See item 4i	Verbal	0.52	0.64	0.69
14	T17	Newton I	An elevator	Vectorial	0.76	0.32	0.50
15	T22	Newton II	A spaceship	Verbal	0.38	0.51	0.63
16	T24	Newton I	A spaceship	Verbal	0.70	0.44	0.57
17	T28	Newton III	See item 8	Bar chart	0.93	0.10	0.33
18	T30	Gravitation	See item 9	Vectorial	0.57	0.39	0.46
19	T1	Gravitation	Falling balls	Bar chart	0.52	0.39	0.53
20	T4	Newton III	Collision of cars	Vectorial	0.83	0.05	0.15
21	T26	Newton II	See item 3	Verbal	0.12	0.19	0.43
22	T13	Gravitation	See item 4i	Graphical	0.44	0.48	0.53
23	T17	Newton I	An elevator	Bar chart	0.77	0.37	0.53
24	T22	Newton II	A spaceship	Motion map	0.42	0.43	0.54
25	T24	Newton I	A spaceship	Motion map	0.70	0.46	0.58
26	T28	Newton III	See item 8	Vectorial	0.94	0.07	0.28
27	T30	Gravitation	See item 9	Verbal	0.68	0.44	0.52

APPENDIX B

Alternatives for items related to theme 4 and examples regarding the grading system for consistency (Table IX, Figs. 7–9).

TABLE IX. Examples regarding the grading system for consistency for theme 4.

Exemplar selection			Points	
Item 2	Item 11	Item 20	Representational consistency	Scientific consistency
a	e	a	2	2
a	e	d	1	1
a	c	d	0	0
b	d	d	2	0
d	a	b	1	0

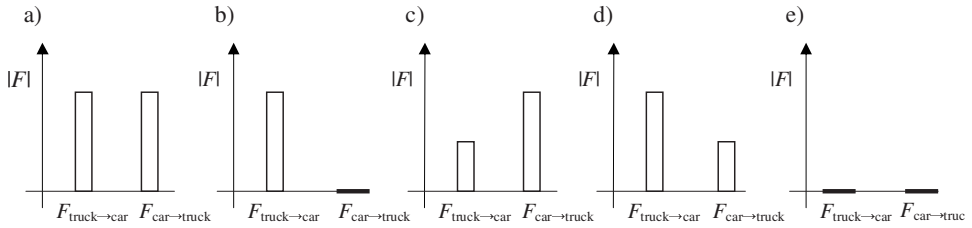


FIG. 7. Alternatives for item 2.

a) the truck exerts a greater amount of force on the car than the car exerts on the truck.
b) the car exerts a greater amount of force on the truck than the truck exerts on the car.
c) neither exerts a force on the other, the car gets smashed simply because it gets in the way of the truck.
d) the truck exerts a force on the car but the car does not exert a force on the truck.
e) the truck exerts the same amount of force on the car as the car exerts on the truck.

FIG. 8. Alternatives for item 11.

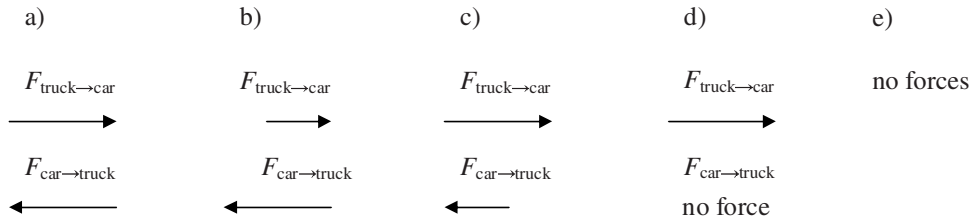


FIG. 9. Alternatives for item 20.

APPENDIX C

Average points for consistency in themes (Table X).

TABLE X. Students' ($n=168$) percentage of average points for representational and scientific consistency in themes.

	Representational consistency		Scientific consistency	
Theme	Pretest (%)	Post-test (%)	Pretest (%)	Post-test (%)
T1	71	82	34	49
T4	82	92	24	89
T13	27	60	0	41
T17	69	87	5	71
T22	67	81	23	37
T24	74	87	48	68
T26	45	65	3	8
T28	83	96	26	93
T30	86	84	2	60
All	67	81	18	57

APPENDIX D

Validation criteria (Table XI).

TABLE XI. Validation criteria.

Theme	Criteria for the correct explanation in a given theme
T1	Acceleration due to gravity is independent of the mass (or weight) of an object. Hence, both objects have the same acceleration.
T4	Forces arising from the same interaction have equal magnitudes and opposite directions OR mentioning Newton's third law.
T13	Gravitational force is the only force acting OR there is no "hit force" after the hit.
T17	The net force acting on the elevator is zero (Newton's first law) OR the object has no acceleration so the net force is zero (Newton's second law).
T22	The net force is not zero so the rocket is accelerating (Newton's second law).
T24	No forces are acting on the rocket. Hence, it has a constant velocity (Newton's first law).
T26	A constant (net) force causes constant acceleration OR A nonzero net force causes an acceleration.
T28	Forces arising from the same interaction have equal magnitudes and opposite directions OR mentioning Newton's third law.
T30	Gravitational force and air-resistance are acting. There is no "hit force."

[1] *Learning with Multiple Representations*, edited by M. W. van Someren, P. Reimann, H. P. A. Boshuizen, and T. de Jong (Pergamon, New York, 1998).

[2] A. Lesgold, Multiple Representations and Their Implications for Learning, in Ref. [1], pp. 307–319.

[3] S. E. Ainsworth, The functions of multiple representations, Comput. Educ. **33**, 131 (1999).

[4] S. E. Ainsworth, DeFT: A conceptual framework for considering learning with multiple representations, Learn. Instr. **16**, 183 (2006).

[5] S. E. Ainsworth, The educational value of multiple representations when learning complex scientific concepts, in Ref. [6], pp. 191–208; available at ⟨http://www.psychology.nottingham.ac.uk/staff/sea/Ainsworth_Gilbert.pdf⟩.

[6] *Visualization: Theory and Practice in Science Education*, edited by J. K. Gilbert, M. Reiner, and M. Nakhleh (Springer, New York, 2008).

[7] T. de Jong, S. Ainsworth, M. Dobson, A. van der Hulst, J. Levonen, P. Reimann, J-A. Sime, M. W. van Someren, H. Spada, and J. Swaak, Acquiring Knowledge in Science and Mathematics: The Use of Multiple Representations in Technology-Based Learning Environments, in Ref. [1], p. 34.

[8] S. Ainsworth, P. Bibby and D. Wood, Analyzing the Costs and Benefits of Multi-Representational Learning Environments, in Ref. [1], pp. 123–125.

[9] A. Van Heuvelen and X. Zou, Multiple representations of work-energy processes, Am. J. Phys. **69**, 184 (2001).

[10] D. Hestenes, Modeling methodology for physics teachers, in *The Changing Role of Physics Departments in Modern Universities: Proceedings of the International Conference on Undergraduate Physics Education, College Park, 1996*, AIP Conference Proceedings No. 399 edited by E. Redish and J. Rigden (AIP, New York, 1997) p. 935; available at ⟨http://modeling.asu.edu./r&e/ModelingMeth-jul98.pdf⟩.

[11] E. Scanlon, How Beginning Students Use Graphs of Motion, in Ref. [1], pp. 67–86.

[12] E. R. Savelsbergh, T. de Jong and M. G. M. Ferguson-Hessler, Competence-Related Differences in Problem Representations: A study in Physics Problem Solving, in Ref. [1], pp. 263–282.

[13] P. Kohl and N. Finkelstein, Effects of representation on students solving physics problems: A fine-grained characterization, Phys. Rev. ST Phys. Educ. Res. **2**, 010106 (2006).

[14] D. Rosengrant, A. Van Heuvelen, and E. Etkina, Do students use and understand free-body diagrams? Phys. Rev. ST Phys. Educ. Res. **5**, 010108 (2009).

[15] D. E. Meltzer, Relation between students' problem-solving performance and representational format, Am. J. Phys. **73**, 463 (2005)).

[16] P. B. Kohl and N. D. Finkelstein, Student representational competence and self-assessment when solving physics problems, Phys. Rev. ST Phys. Educ. Res. **1**, 010104 (2005).

[17] A. Savinainen, P. Nieminen, J. Viiri, J. Korkea-aho, and A. Talikka, *Proceedings of the Physics Education Research Conference, Greensboro, 2007*, AIP Conference Proceedings No. 951, edited by L. Hsu, C. Henderson, and L. McCullough (AIP, New York, 2007), p. 176.

[18] M. Dancy and R. Beichner, Impact of animation on assessment of conceptual understanding in physics, Phys. Rev. ST Phys. Educ. Res. **2**, 010104 (2006).

[19] A. Savinainen and J. Viiri, *Proceedings of the Physics Education Research Conference, Madison, 2003*, AIP Conference Proceedings No. 720, edited by J. Marx, S. Franklin, and K. Cummings (AIP, New York, 2004), p. 77; available at http://kotisivu.dnainternet.net/savant/representations_perc_2003.pdf⟩.

[20] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992). Tables I and II, re-

vised for the 1995 version (Ref. [21]), are available at ⟨http://modeling.asu.edu/R&E/Research.html⟩, directly below the first reference under "Articles about the FCI."

[21] I. Halloun, R. R. Hake, E. P. Mosca, and D. Hestenes, Force Concept Inventory, (Revised 1995); available (password protected) at ⟨http://modeling.asu.edu/R&E/Research.html⟩, scroll down to "Evaluation Instruments." Currently available in 19 languages: Arabic, Chinese, Czech, English, Finnish, French, French (Canadian), German, Greek, Italian, Japanese, Malaysian, Persian, Portuguese, Russian, Spanish, Slovak, Swedish, and Turkish.

[22] D. Hestenes, *Modelling in Physics and Physics Education, Proceedings of GIREP Conference 2006, Amsterdam*, edited by E. van den Berg, T. Ellermeijer, and O. Slooten (2006), p. 34; available at ⟨http://modeling.asu.edu/R&E/Research.html⟩ where it is stated that: "Pages 16–22 are very important in explaining why the FCI is so successful in assessing student concept understanding.

[23] Instructors and researchers can obtain The Representational Variant of the FCI by e-mailing Pasi Nieminen (pasi.k.nieminen@jyu.fi) or Antti Savinainen (antti.savinainen@kuopio.fi).

[24] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).

[25] D. Giancoli, *Physics—Principles with Applications*, 5th ed. (Prentice-Hall International, Englewood Cliffs, NJ, 1998).

[26] J. Hatakka, H. Saari, J. Sirviö, J. Viiri, and S. Yrjänäinen, *Physica 1* (WSOY, Porvoo, 2004).

[27] A. Savinainen and P. Scott, Using the Force Concept Inventory to monitor student learning and to plan teaching, Phys. Educ. **37**, 53 (2002)⟩.

[28] A. Savinainen, P. Scott, and J. Viiri, Using a bridging representation and social interactions to foster conceptual change: Designing and evaluating an instructional sequence for Newton's third law, Sci. Educ. **89**, 175 (2005).

[29] L. Turner, System schemas, Phys. Teach. **41**, 404 (2003).

[30] 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).

[31] R. R. Hake, Interactive-engagement methods in introductory mechanics courses 1998, unpublished; available at ⟨http://www.physics.indiana.edu/~sdi/IEM-2b.pdf⟩.

[32] 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).

[33] A. Savinainen and J. Viiri, The Force Concept Inventory as a Measure of Students Conceptual Coherence, Int. J. Sci. Math. Educ. **6**, 719 (2008).

[34] A. Savinainen and P. Scott, The Force Concept Inventory: a tool for monitoring student learning, Phys. Educ. **37**, 45 (2002)⟩.

[35] R. Doran, *Basic Measurement and Evaluation of Science Instruction* (NSTA, Washington, DC, 1980), p. 97: available at ⟨http://tinyurl.com/c8nl58⟩. Pages 23–24, 56, 77, 81, 85, 118, and 121 are unavailable due to copyright restrictions.

[36] See Ref. [35], p. 99.

[37] P. Kline, *A Handbook of Test Construction: Introduction to Psychometric Design* (Methuen, London, 1986), p. 143.

[38] See Ref. [37], p. 124.

[39] See Ref. [35], p. 104.

[40] See Ref. [37], p. 144.

[41] *Systems for State Science Assessment*, edited by M. R. Wilson and M. W. Bertenthal (Nat. Acad. Press, Washington, DC, 2005), p. 94; available at ⟨http://www.nap.edu/catalog.php?record_id=11312⟩.

[42] I. Halloun and D. Hestenes, The initial knowledge state of college physics students, Am. J. Phys. **53**, 1043 (1985)⟩. The print version contains the "Mechanics Diagnostic" test, precursor to the "Force Concept Inventory" (Refs. [20,21]).

[43] G. J. Aubrecht and J. D. Aubrecht, Constructing Objective Tests, Am. J. Phys. **51**, 613 (1983).

[44] R. J. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).