Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning

Pasi Nieminen, Antti Savinainen, and Jouni Viiri
Department of Teacher Education, University of Jyväskylä, Jyväskylä FIN-400014, Finland
(Received 12 December 2011; published 16 May 2012)

Previous physics education research has raised the question of ''hidden variables'' behind students' success in learning certain concepts. In the context of the force concept, it has been suggested that students' reasoning ability is one such variable. Strong positive correlations between students' preinstruction scores for reasoning ability (measured by Lawson's Classroom Test of Scientific Reasoning) and their learning of forces [measured by the Force Concept Inventory (FCI)] have been reported in high school and university introductory courses. However, there is no published research concerning the relation between students' ability to interpret multiple representations consistently (i.e., representational consistency) and their learning of forces. To investigate this, we collected 131 high school students' pre- and post-test data of the Representational Variant of the Force Concept Inventory (for representational consistency) and the FCI. The students' Lawson pretest data were also collected. We found that the preinstruction level of students' representational consistency correlated strongly with student learning gain of forces. The correlation (0.51) was almost equal to the correlation between Lawson prescore and learning gain of forces (0.52). Our results support earlier findings which suggest that scientific reasoning ability is a hidden variable behind the learning of forces. In addition, we suggest that students' representational consistency may also be such a factor, and that this should be recognized in physics teaching.

DOI: 10.1103/PhysRevSTPER.8.010123 PACS numbers: 01.40.−d

I. INTRODUCTION

Assessing students' conceptual understanding has been a popular issue in physics education research (for a review, see [1] and references therein). In this field, perhaps the most widely used assessment instrument is the Force Concept Inventory (FCI) [2], intended for evaluating students' conceptual understanding of force. An important aspect of the research is evaluating the change in students' conceptual understanding during instruction. There are various ways of gauging the change, but one popular measure in physics education research is the average normalized learning gain $\langle g \rangle$ [3], which is defined as the ratio of the actual gain to the maximum possible gain:

$$\langle g \rangle = \frac{(\text{postscore}\%) - (\text{prescore}\%)}{100\% - (\text{prescore}\%)}.$$

The average normalized learning gain is used for measuring the change in a class of students (i.e., pre- and postscores are class averages), but the formula above has also been used for evaluating individual student's learning gain (see, for example, [4]). In the latter case, G is called a single student normalized gain, and the pre- and postscores in the formula are those of a single student.

The normalized gain is a useful measure as it allows the comparison of results with different preinstruction scores (a possible relation between prescore and gain is discussed later). For example, a normalized gain of 0.5 can be achieved with different combinations of test scores, e.g., 60% in the pre- and 80% in the post-test, or 80% in the pre- and 90% in the post-test. Hence, it has been used for comparing test results of student groups and thus the effectiveness of different teaching methods.

Hake [3] analyzed extensive FCI data from 62 introductory physics courses ($n = 6542$) and showed that the average normalized learning gains were higher in interactive-engagement (IE) courses (0.48 ± 0.14; mean $\pm$ standard deviation) than in traditional courses (0.23 ± 0.04). There is no reason to doubt that normalized learning gain may depend on the instructional method used, but it has been suggested [5,6] that differences between the gains of student groups may not be due simply to instructional methods. Various hidden variables such as general intelligence, reasoning ability, and study habits may also influence the size of the learning gain a certain student population can achieve.

In Hake's study [3], no significant correlation was found between FCI prescores and average normalized FCI gain. However, Coletta and Phillips [6] reported a correlation ($r = 0.63$) between the class prescore and class average normalized gain in 38 college and university interactive-engagement classes. They also reported a significant positive correlation between the FCI prescore and single student normalized FCI gain in three of four university

courses where IE methods were used ($r = 0.33$, $n = 285$; $r = 0.30$, $n = 96$; $r = 0.15$, $n = 1648$).

Coletta, Phillips, and Steinert [4,7] and Coletta and Phillips [6] suggest that a student group's scientific reasoning level may explain why there is or is not a correlation between FCI prescore (FCI_{pre}) and single student normalized FCI gain (G_{FCI}) in some groups. Students with the strongest reasoning abilities may get both high FCI_{pre} and high G_{FCI}. Such students achieve higher G_{FCI} in high school, so they have high FCI_{pre} in university, and because of their high reasoning ability they also achieve high G_{FCI}. This hypothesis was supported by the finding [7] in 98 university students that the Lawson prescore (L_{pre}) and the FCI_{pre} correlated ($r = 0.53$), and that the L_{pre} and the G_{FCI} also correlated ($r = 0.51$). In this student group the correlation between FCI_{pre} and G_{FCI} was positive ($r = 0.33$). On the other hand, Coletta and Phillips have proposed that perhaps these correlations do not exist among the high-level reasoners who would score very high on the Lawson test. They reported no correlation between FCI_{pre} and L_{pre} ($r = 0.005$) nor between FCI_{pre} and G_{FCI} ($r = 0.01$) among the best reasoners of 65 university students ($n = 16$; top quartile of Lawson scores) [6]. They considered that this could explain why the correlation between FCI_{pre} and G_{FCI} did not exist in one of the four universities studied (Harvard University), whose students they supposed to be such high-level reasoners.

Coletta and Phillips [7] reported a correlation between L_{pre} and G_{FCI} also among high school students ($r = 0.53$, $n = 199$). Such a correlation has also been found in many replication studies [8]. Coletta, Phillips, and Steinert have argued that achieving a high FCI gain can be easier in classes where the level of the students' scientific reasoning is also high. They have also created a program for identifying students who have low scientific reasoning ability which can also enhance their reasoning in order to help them to learn physics [8].

Previous research has shown that expert scientists are able to fluently use multiple representations when they are thinking and sharing ideas [9,10], and it is argued that one important goal of a physics education is to guide students to expertlike use of multiple representations for successful problem solving and a good conceptual understanding of physics [11,12]. Even the representational format (e.g., graph, vector, or motion map) in which a problem is posed can affect student performance [13–16]. Physics education research has shown that an instructional approach emphasizing multiple representations is helpful for students' use of multiple representations when the approach is strongly or weakly directed [17]. In the chemistry education context as well, it has been reported that students' learning from multiple representations can be supported by directive and nondirective help depending on their prior knowledge [18].

It is reasonable to assume that the ability to use multiple representations could play some role in students' conceptual gain in a physics course. Hence, our first aim was to clarify the relation between students' representational consistency and G_{FCI}. By representational consistency we mean students' ability to interpret various representations (e.g., graphs, vectors, and motion maps) between isomorphic items in which content and context are essentially identical. For this we use the Representational Variant of the Force Concept Inventory (R-FCI) [16]. Our second aim was to investigate the relations between the students' FCI results and Lawson prescores. This was motivated by the interesting findings on the relations between FCI results and Lawson prescores among university and high school physics students in the U.S. [6,7]; specifically, we wanted to find out whether or not these findings can be replicated in a Finnish high school setting. Figure 1 summarizes the correlations investigated in this paper. We posed the following research questions:

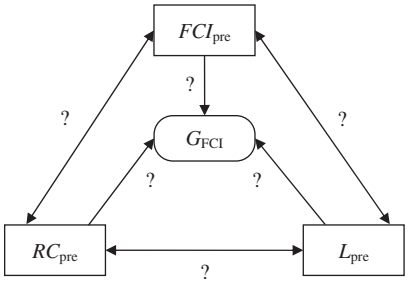

FIG. 1. The correlations between single student normalized FCI gain (G_{FCI}) and the three pretest variables: representational consistency on the R-FCI pretest (RC_{pre}), the FCI prescore (FCI_{pre}), and the Lawson prescore (L_{pre}).

(1) Is there a relation between the preinstruction level of students' representational consistency (RC_{pre}) and single student normalized FCI gain (G_{FCI})?
(2) To what extent can we confirm earlier findings concerning the relation between FCI prescore (FCI_{pre}) and Lawson prescore (L_{pre}) and their relation to G_{FCI}?

The motivation to study representations in the context of force was due to two main reasons. Firstly, our research group has done research on the teaching and learning of the force concept for over ten years. Hence, we have special expertise in this particular domain. Secondly, students have some ideas regarding the force concept even before any formal schooling (unlike, for example, regarding special relativity). This is particularly relevant in our study as we were investigating the understanding of students taking their first, mandatory high school course on physics.

II. METHODS

A. Research instruments

1. Force Concept Inventory

The Force Concept Inventory [2,19] is a multiple-choice test for assessing students' understanding of the force

concept. It is probably the most widely used instrument for evaluating the effectiveness of instruction in physics education research [20]. It has gone through a lengthy process of validation and its reliability has been well established (for a review, see [21,22]). The 1995 version contains 30 items that cover the most basic concepts in Newtonian physics. Each item has five alternatives: one correct Newtonian alternative and four incorrect common sense alternatives. Most of the items are presented verbally, but some items also contain information in pictorial format.

2. Representational Variant of the Force Concept Inventory

(a) Description of the structure.—We have previously presented the structure, validation, and purpose of the R-FCI [16], which is based on nine items taken from the 1995 version of the FCI [19]. The original, verbal multiple-choice alternatives of the FCI items were redesigned using various representations (bar charts, graphs, vectors, motion maps). 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 consisting of an original FCI item and two isomorphic variants. Figure 2 presents corresponding multiple-choice alternatives of a theme depicted via different representations. There are nine themes in the R-FCI, so the test contains 27 items in total. The themes deal with Newton's laws and gravitation. For a more detailed description of the R-FCI, see our previous article [16].

(b) Analysis of R-FCI results.—The R-FCI score gives information about students' conceptual understanding of the force concept. We have found a strong correlation between R-FCI and FCI scores, although the R-FCI does not include all the dimensions of the force concept that the FCI covers [16]. Furthermore, the R-FCI results carry information about students' representational consistency, i.e., their ability to use different representations consistently between isomorphic items. To reach a deeper understanding of this, consistency analyses were conducted.

Representational consistency does not necessarily require scientific correctness in terms of physics. When exhibiting representational consistency, a student may answer all the items in a certain theme scientifically correctly. On the other hand, all the answers for the theme can be scientifically incorrect, and still the alternatives of the items correspond with regard to the representations (see Fig. 2 for an example of a scientifically incorrect but representationally consistent answer pattern in a theme). Thus, only the ability to interpret multiple representations is considered in the concept of representational consistency.

To determine the students' representational consistency their answers for a given theme were given points in the following way:

- two points, if they had chosen corresponding alternatives in all three items of the theme
- one point, if they had chosen corresponding alternatives in two of the three items of the theme
- zero points, if no corresponding alternatives in the items of the theme were selected

In this paper we do not use information about consistency in single themes as we did in our previous study [16]. In contrast, we consider the average consistency in all themes. Thus, all numbers relating to the representational consistency presented in this study are percentages of maximal representational consistency of all themes.

The consistency analysis was solely based on quantitative data, that is, students' multiple-choice answers. These were typed in a spreadsheet which was used to implement the analysis according to coded categorization rules. Hence, there was no significant researcher effect on the consistency analysis and thus no requirement for an inter-rater reliability analysis.

3. Classroom test of scientific reasoning

Lawson's Classroom Test of Scientific Reasoning [23], or the Lawson test, is designed to assess the students' level of formal reasoning. The version [24] used in this study contains 24 multiple-choice items concerning the conservation of mass and volume, proportional reasoning, control of variables, probabilistic reasoning, correlational reasoning, and hypothetico-deductive reasoning (see Table IX in [25]). The validity of the original test version [23] has been established by several studies (see references in [26]).

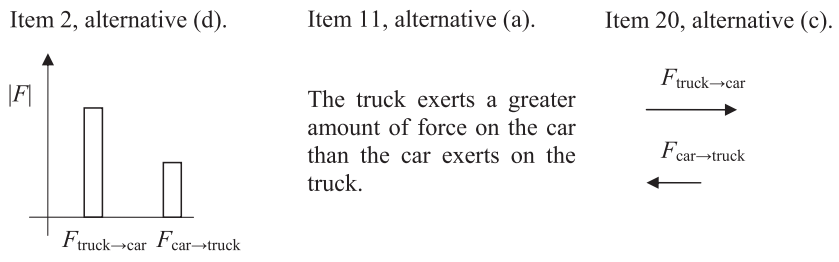

FIG. 2. Corresponding multiple-choice alternatives of a theme. The representational formats of the alternatives are a bar chart (item 2), verbal (item 11), and vectorial (item 20). All three items include an identical, original FCI question in verbal form. The questions with the bar chart and vectorial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

TABLE I. Participants.

Group (year)	n
Phys1a (2008)	31
Phys1b (2008)	31
Pre-IB (2008)	21
Phys1 (2010)	25
Pre-IB (2010)	23
Total	131

B. Participants and data collection

Five groups of Finnish first-year high school students ($n = 131$, aged 16) participated in this study (Table I). The Phys1 groups consisted of regular students, and the Pre-IB groups consisted of students preparing for the International Baccalaureate program.

The students 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 [27], whereas all the others studied in Finnish using a Finnish textbook [28]. Despite having different textbooks, the students had many common exercises addressing the use of multiple representations in kinematics and Newton's laws.

All participants took the R-FCI and FCI before and after the courses, but the Lawson test only before the courses. The Phys1 groups took all tests in Finnish. The Pre-IB group took their pretests in Finnish because their English was not good enough at the beginning of the course; however, as all teaching took place using the English language, their post-tests were in English. This may cause a concern about the effect of language on students' performance. To look for evidence of this possible effect, we compared the single student normalized FCI and R-FCI gain between Pre-IB and Phys1 groups: we did not find statistically significant differences (described in more detail below). In this regard, Pre-IB students' learning was very similar to that of Phys1 students, despite the change of language in the post-tests.

All the groups were taught by one of the authors (A. S.), using interactive-engagement teaching methods with various representations; this author has used these methods for many years (for details, see [22]). Furthermore, the teaching approach had a strong focus on treating forces as interactions; this approach has been very successful in fostering students' understanding of Newton's third law [29].

We did not separate Pre-IB and Phys1 students in the data analysis. Despite the described differences between the Pre-IB and Phys1 courses, the students had the same teacher, were exposed to the same instructional methods, and they were all participating in their first high school physics course. Certainly, it was possible that there were some differences between Pre-IB and regular students' academic skills (e.g., language skills) given that Pre-IB students selected to study under the International Baccalaureate program using the medium of English, which is not their native language. However, we did not find statistically significant differences between the student groups in the preinstruction results (L_{pre}, FCI_{pre}, $R\text{-}FCI_{\mathrm{pre}}$, RC_{pre}), or with regard to the single student normalized FCI or R-FCI gain when analysis of variance (ANOVA) and Kruskal-Wallis tests were conducted. Hence, for the purposes of this study we consider all students as one group.

III. RESULTS

A. Results for the whole group of students

The results of the different tests are given in Table II. The average normalized FCI gain (0.38) was in the ''medium-g region'' (between 0.3 and 0.7 [3]). The pretest results of the R-FCI revealed a big difference between the score (scientifically correct answers) and the representational consistency: despite the rather low pretest score (23%), students exhibited some representational consistency (64%). The R-FCI prescore was statistically significantly lower than the FCI prescore when the Wilcoxon signed-rank test was conducted ($z = 6.34$, $p < 0.001$). In contrast, the postscore and single student normalized gain (0.50 for R-FCI and 0.40 for FCI) were higher for the R-FCI than for the FCI. The differences were statistically significant for the postscores ($z = 4.36$, $p < 0.039$) and the single student normalized gains ($z = 7.21$, $p < 0.001$). One possible reason for this may be that the items of the R-FCI used various representational formats, which can be difficult for students to handle at the beginning of their first high school course. The R-FCI gain indicates an increase in the conceptual understanding of forces and in representational consistency. The difference between the postscores could indicate that the FCI was more difficult for the students, which, in turn, might be due to the greater number and difficulty of items in the FCI.

For calculating correlations between different variables, Spearman's rank correlation coefficient (ρ) was used, because many of the variables studied did not distribute normally. Figure 3 shows correlations between G_{FCI}, R-FCI pretest representational consistency (RC_{pre}), FCI prescore (FCI_{pre}), and Lawson prescore (L_{pre}).

TABLE II. Students' ($n = 131$) results in different tests. Means and average normalized gains for test scores and representational consistency of the R-FCI. Standard error of the mean is in parentheses.

	FCI score	R-FCI score	Representational consistency	Lawson score
Pretest (%)	29 (1)	23 (1)	64 (1)	61 (2)
Post-test (%)	56 (2)	61 (2)	82 (1)	⋯
Gain	0.38	0.49	0.50	⋯

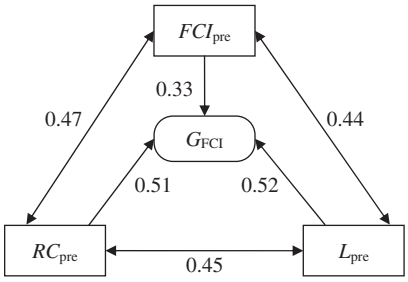

FIG. 3. Spearman's rank correlation between single student normalized FCI gain (G_{FCI}) and the three pretest variables for all the students ($n = 131$): representational consistency on the R-FCI (RC_{pre}), the FCI score (FCI_{pre}), and the Lawson test score (L_{pre}). All correlations are statistically significant ($p < 0.001$).

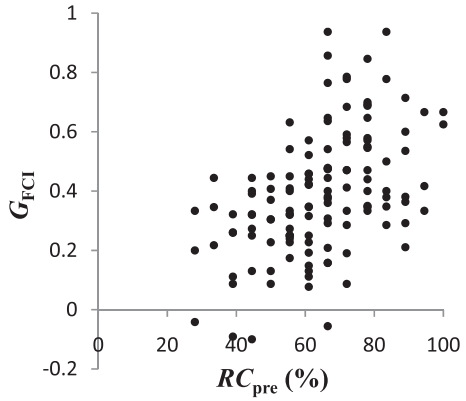

FIG. 4. Scatter plot for the students' ($n = 131$) representational consistency on the R-FCI pretest (RC_{pre}) and the single student normalized FCI gain (G_{FCI}). Spearman's rank correlation is 0.51 ($p < 0.001$).

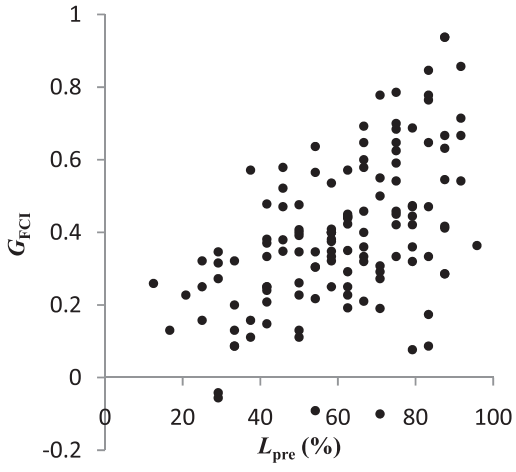

FIG. 5. Scatter plot for the students' ($n = 131$) Lawson prescore (L_{pre}) and the single student normalized FCI gain (G_{FCI}). Spearman's rank correlation is 0.52 ($p < 0.001$).

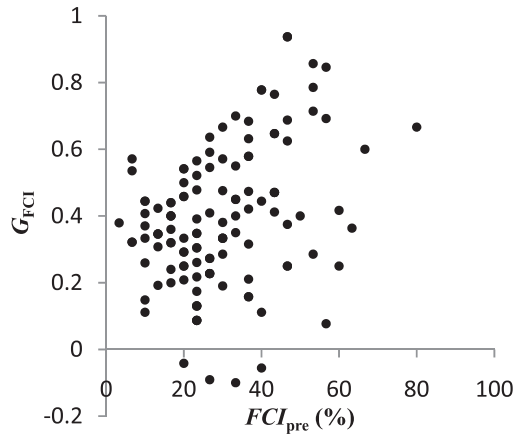

FIG. 6. Scatter plot for the students' ($n = 131$) FCI prescores (FCI_{pre}) and the single student normalized FCI gain (G_{FCI}). Spearman's rank correlation is 0.33 ($p < 0.001$).

Figures 4–6 show scatter plots for the correlations between different pretest variables and G_{FCI}. There was a positive correlation ($\rho = 0.33$, $p < 0.001$) between the FCI_{pre} and G_{FCI}, but it was clearly weaker than the correlation between G_{FCI} and RC_{pre} ($\rho = 0.51$, $p < 0.001$) or the correlation between G_{FCI} and the L_{pre} ($\rho = 0.52$, $p < 0.001$).

It should be noted that the R-FCI representational consistency and the R-FCI score are very different measures. We found that the R-FCI prescore ($R\text{-}FCI_{pre}$) correlated only weakly with G_{FCI} ($\rho = 0.23$, $p = 0.008$), whereas the correlation of RC_{pre} and G_{FCI} was 0.51. There was also a strong correlation between $R\text{-}FCI_{pre}$ and FCI_{pre} ($\rho = 0.79$, $p < 0.001$), which indicates that the different tests were quite accurately measuring the same construct, i.e., the understanding of the force concept. In contrast, the correlation between RC_{pre} and FCI_{pre} was not so high ($\rho = 0.47$, $p < 0.001$); it was almost the same as the correlation between RC_{pre} and L_{pre} ($\rho = 0.45$, $p < 0.001$) and that between FCI_{pre} and L_{pre} ($\rho = 0.44$, $p < 0.001$).

We found some interesting results concerning single student gain on representational consistency. In calculating this gain, two of the 131 students had to be excluded because their pretest representational consistency was 100%, and in such a case the calculation of normalized gain is impossible because the divisor would be zero (see the equation for normalized gain in the Introduction). There was no correlation between the pretest representational consistency and single student normalized gain on representational consistency ($\rho = -0.026$, $p = 0.77$, $n = 129$), indicating that the students had learned to interpret multiple representations regardless of their preinstruction level of representational consistency. This gain also correlated very weakly with the $R\text{-}FCI_{pre}$ ($\rho = 0.11$, $p = 0.20$, $n = 129$) and the FCI_{pre} ($\rho = 0.18$, $p = 0.041$, $n = 129$). Moreover, there was a weak positive correlation between

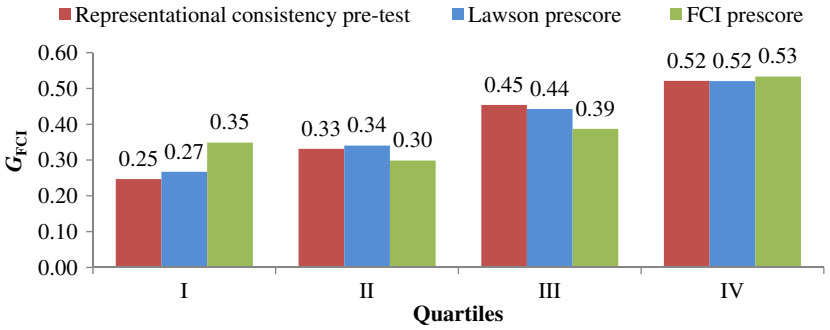

FIG. 7 (color online). Average of single student normalized FCI gain (G_{FCI}) in quartiles of representational consistency on the R-FCI pretest, Lawson prescores, and FCI prescores.

the L_{pre} and representational consistency gain ($\rho = 0.28$, $p = 0.001$, $n = 129$). The correlation between representational consistency gain and G_{FCI} was strong and positive ($\rho = 0.44$, $p < 0.001$, $n = 129$).

B. Results in the subgroups

Figure 7 shows the G_{FCI} averages in different quartiles. Quartiles were constructed in such a way that students were divided into four equal-sized groups according to a certain variable, for example, their RC_{pre}. As regards the RC_{pre} and L_{pre} quartiles, we found that the G_{FCI} average increased from the lowest to the highest quartile. In addition, in each of the four quartiles, the G_{FCI} averages for representational consistency and Lawson score within a given quartile were nearly equal to each other. In contrast, when the FCI_{pre} quartiles were considered, the G_{FCI} average was even higher in the first quartile than in the second. It can be seen from Fig. 7 that the quartile distributions are consistent with the correlations in Figs. 4–6: representational consistency and the Lawson score correlated more strongly with the G_{FCI} than did the FCI score.

The analysis of correlations in the quartiles was problematic because of the small range of values of the variables studied in some quartiles. For example, when the RC_{pre} quartiles were considered, it was difficult to calculate the correlation between RC_{pre} and G_{FCI} in a certain quartile because the RC_{pre} may have had only two values in the quartile. Therefore (with one exception shown below), instead of quartiles we studied correlations when students were placed into the top and bottom half according to their L_{pre} and RC_{pre} .

As explained in Sec. III A, there was a positive correlation ($\rho = 0.33$, $p < 0.001$) between FCI_{pre} and G_{FCI}. When students were placed into the top (T) or bottom (B) half according to their Lawson prescore (see Fig. 8), we found that this correlation did not exist in the bottom ($\rho = -0.017$, $p = 0.90$) but did in the top half ($\rho = 0.43$, $p < 0.001$). Moreover, the correlation between FCI_{pre} and L_{pre} did not exist in the lower half ($\rho = 0.028$, $p = 0.83$), but was strong in the top half ($\rho = 0.50$, $p < 0.001$).

Because our results seemed to contradict the earlier results [6] regarding the students in the highest Lawson quartile discussed in our Introduction, we also studied these correlations in the highest Lawson quartile ($n = 30$): the correlation between FCI_{pre} and L_{pre} was positive but nonsignificant ($\rho = 0.34$, $p = 0.069$), as was the correlation between FCI_{pre} and G_{FCI} ($\rho = 0.33$, $p = 0.072$).

These correlations were very similar when the division was done according to the pretest representational consistency (see Fig. 9): the correlation between FCI_{pre} and G_{FCI} was not statistically significant and even negative in the bottom half (B, $\rho = -0.22$, $p = 0.091$), but strong and positive in the top half (T, $\rho = 0.45$, $p < 0.001$). FCI_{pre} and L_{pre} did not correlate among students in the bottom half ($\rho = 0.15$, $p = 0.25$), but the correlation was strong in the top half ($\rho = 0.46$, $p < 0.001$).

When the Lawson division was considered (see Fig. 8), L_{pre} correlated with G_{FCI} in the bottom ($\rho = 0.46$, $p < 0.001$) and top half ($\rho = 0.30$, $p = 0.011$), although the correlation was stronger in the bottom half. Likewise, the correlation between RC_{pre} and G_{FCI} was stronger in the bottom ($\rho = 0.49$, $p < 0.001$) than in the top half ($\rho = 0.34$, $p = 0.003$). In contrast, RC_{pre} correlated with FCI_{pre} more strongly in the top ($\rho = 0.48$, $p < 0.001$) than in the bottom half ($\rho = 0.33$, $p = 0.010$). Also, the

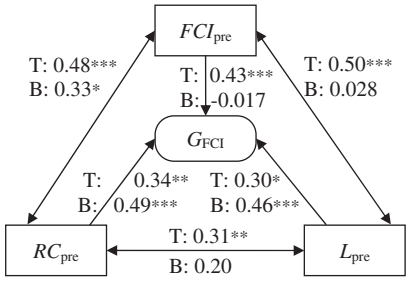

FIG. 8. Spearman's rank correlation between single student normalized FCI gain (G_{FCI}) and the three pretest variables: representational consistency on the R-FCI (RC_{pre}), the FCI score (FCI_{pre}), and the Lawson score (L_{pre}). Students were placed into the top (T, $n = 71$) or bottom (B, $n = 60$) half according to their L_{pre}. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

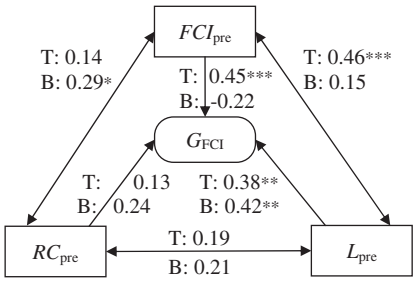

FIG. 9. Spearman's rank correlation between single student normalized FCI gain (G_{FCI}) and the three pretest variables: representational consistency on the R-FCI (RC_{pre}), the FCI score (FCI_{pre}), and the Lawson score (L_{pre}). Students were placed into the top (T, $n = 69$) or bottom (B, $n = 62$) half according to their RC_{pre}. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

correlation between RC_{pre} and L_{pre} was stronger in the top ($\rho = 0.31$, $p = 0.009$) than in the bottom half ($\rho = 0.20$, $p = 0.12$).

When students were placed into two groups according to their pretest representational consistency (Fig. 9), there was a correlation between L_{pre} and G_{FCI} in the bottom ($\rho = 0.42$, $p = 0.001$) and top quartiles ($\rho = 0.38$, $p = 0.001$). Correlations between RC_{pre} and other variables were quite weak. For example, the correlation between RC_{pre} and G_{FCI} was weak and statistically nonsignificant in both the bottom ($\rho = 0.24$, $p = 0.063$) and top half ($\rho = 0.13$, $p = 0.28$), while this correlation was strong among all students ($\rho = 0.51$, $p < 0.001$).

C. Reliability index

For internal consistency we calculated values of the Kuder-Richardson formula 20 (KR-20) for all the tests used (Table III). For a reliable group measurement, the KR-20 should be higher than 0.7, and for an individual measurement it should be over 0.8 [30].

We used the tests as an individual measurement because single student results were used. All test values were over 0.8 except for that of the FCI pretest (0.75). However, we accepted this value because it was near 0.8. In addition, the post-test value of the FCI was over 0.8. Because a reliability index is always sample dependent, it is possible that the FCI was quite difficult for the students at the beginning of their first physics course, and this produced the value under 0.8 for the pretest.

TABLE III. KR-20 values for different tests ($n = 131$).

Test	KR-20
R-FCI pretest	0.83
R-FCI post-test	0.87
FCI pretest	0.75
FCI post-test	0.83
Lawson pretest	0.81

IV. DISCUSSION

Our first research question was to investigate the correlation between the R-FCI pretest representational consistency (RC_{pre}) and single student normalized FCI gain (G_{FCI}). The second research question was to examine the relations between the FCI and Lawson test results to confirm earlier findings [6,7]. In addition to the whole group of students, we also studied these relations among subgroups in order to discover whether the students' preinstruction level of representational consistency or scientific reasoning had an effect on the existence or absence of some relations.

We found that students' RC_{pre} correlated strongly with G_{FCI} ($\rho = 0.51$, $p < 0.001$), which was bigger than the correlation between FCI prescore (FCI_{pre}) and G_{FCI} ($\rho = 0.33$, $p < 0.001$), but almost the same as the correlation between Lawson prescore (L_{pre}) and G_{FCI} ($\rho = 0.52$, $p < 0.001$). When students were placed into the top and bottom half according to their RC_{pre} , the correlation between RC_{pre} and G_{FCI} disappeared in the subgroups. In that regard, the correlation seemed to be a property of the whole student group. Likewise, this correlation existed in both the bottom ($\rho = 0.49$) and the top half ($\rho = 0.34$) when students were split according to the L_{pre}, although the correlation was slightly weaker among the top-half reasoners.

Interestingly, we found no correlation between students' pretest representational consistency and representational consistency gain ($\rho = -0.026$), indicating that students can learn to interpret multiple representations regardless of their preinstruction level of representational consistency. Furthermore, students' preinstruction score on the Lawson test correlated weakly ($\rho = 0.28$) with representational consistency gain.

We are not aware of previous reports concerning the relation between the ability to interpret multiple representations and the learning gain of a certain concept. We found a strong positive correlation between students' preinstruction level of representational consistency and their learning of forces. We cannot say that the relation is certainly causal. However, causality is not impossible, because an understanding of representations is required for the adequate use of scientific concepts. It is of course possible that there are also other influential factors, such as general intelligence and spatial ability which explain the ability to interpret multiple representations.

Coletta, Phillips, and Steinert [7] reported a strong positive correlation ($\rho = 0.53$) between the FCI and Lawson prescores among the 98 American university students they examined. They assumed that the students with high reasoning abilities had achieved higher learning gains in high school, so they would have high pretest scores in university. This would explain the correlation between the FCI and Lawson prescores in university. In this study, this correlation also existed among students in their first high school course ($\rho = 0.44$, $p < 0.001$), but it was weaker than that found in the aforementioned study. This seems

reasonable, because the students in our study are unlikely to have achieved much conceptual understanding of force during their lower secondary school education. Their prescore on the FCI varied between 3% and 80%, and the average was 29%. It was slightly higher than the probabilistic score produced by guessing, which in this case would have been 20%. In addition, our high school data showed that the correlation between L_{pre} and FCI_{pre} was higher among better reasoners and among the more the representationally consistent students: the correlation was 0.50 in the top and 0.028 in the bottom half (L_{pre} split) and 0.46 in the top and 0.15 in the bottom half (RC_{pre} split).

We found a positive correlation ($\rho = 0.33$, $p < 0.001$) among all students between FCI prescore and G_{FCI}. This was quite the same as Coletta and Phillips had reported [6] concerning two of four university courses where IE methods were used ($r = 0.33$, $n = 285$; $r = 0.30$, $n = 96$; $r = 0.15$, $n = 1648$). The correlation was not found among students of Harvard University ($r = 0.037$, $n = 670$). Coletta and Phillips assumed that many of the Harvard University students had achieved a high level of scientific reasoning and would have scored very high on the Lawson test for that reason. They found that among 65 students from Loyola Marymount University, as regards the students ($n = 16$) who scored highest on the Lawson test (top quartile), there was no correlation between FCI and Lawson prescores ($r = 0.005$), nor between FCI prescore and G_{FCI} ($r = 0.01$). Among the top quartile in our data ($n = 30$), these correlations existed ($\rho = 0.34$, $p = 0.069$; $\rho = 0.33$, $p = 0.072$, respectively), but were not statistically significant. It must be noted that the participants in our study were first-year high school students, whereas those in the study by Coletta and Phillips were attending university. In our data Lawson prescore (85%) and G_{FCI} (0.52) in the top quartile were lower than was the case in the top quartile of the study by Coletta and Phillips (93% and 0.59, respectively). There is a possibility that the scientific reasoning of the top quartile students in our data was not strong enough, so that these correlations would not have existed in their case. Anyway, in our high school data, the correlation between FCI_{pre} and G_{FCI} was stronger among top-half students when the students were placed into the top and bottom half according to the L_{pre} and RC_{pre} (see Figs. 8 and 9).

Coletta, Phillips, and Steinert [7] reported a strong correlation between students' preinstructional level of scientific reasoning ability and the single student normalized FCI gain among 98 university students ($r = 0.51$) and 199 high school students ($r = 0.53$). They have also reported that such a correlation has been found in many replication studies [8]. Further, they [7,8] have created a program for identifying students who have low scientific reasoning ability, and which can be used to enhance their reasoning in order to help them to learn physics. We were able to confirm the correlation between students' scientific reasoning ability and G_{FCI} in our data ($\rho = 0.52$, $p < 0.001$). Hence, we are convinced that weak physics students might particularly benefit from the explicit teaching of scientific reasoning skills.

V. VALIDITY AND LIMITATIONS

The data of this study were collected with quantitative multiple-choice tests that were straightforward to take, administer, and score without researcher bias. The reliability and validity of the study are affected by the reliability and validity of the test instruments. We discuss the validity of the tests in Sec. II A, and we consider them valid for high school students. For reliability, which is a prerequisite for any validity, we calculated KR-20 values, and these were acceptable for all the tests used in this study (Table III).

External validity (generalizability) is the major limitation of this study. The results cannot be generalized even to the population of all first-year high school students in Finland, because the data were collected in a particular high school and from students taking courses with a particular teacher.

VI. IMPLICATIONS

Our results concerning the strong relationship between students' representational consistency and their learning of forces are well in line with those of previous studies [11–14] supporting that careful consideration of multiple representations is important for learning and understanding physics concepts. One way to increase knowledge about multiple representations in physics teaching among Finnish high school teachers would be to offer resources for teaching multiple representations, such as research-based materials and practices. There is a clear need for the aforementioned resources as physics textbooks in Finland do not often include many multiple representation exercises [31]. Furthermore, textbooks tend to have a central role in Finnish high school physics teaching. Another potentially effective field in which to highlight the importance of multiple representations could be in the training of preservice physics teachers.

Earlier research has shown that an instructional approach emphasizing multiple representations can be helpful to university students in their use of multiple representations [17,18]. Our study cannot fully take part in the discussion on instructional approach as only one teaching method was used and without comparison groups. However, in our other study [31], Finnish high school students ($n = 28$) answered open-ended, paper-and-pencil questions which we had designed to emphasize the use of multiple representations in the context of forces. The results lend some support that students' understanding of the force concept and multiple representations was increased.

The data of the present study were collected in one Finnish high school where the interactive-engagement (IE) teaching method and multiple representations were used. In the future, attempts should be made to replicate the results with different groups of students. Further studies should investigate what kind of correlation exists between preinstruction representational consistency and G_{FCI} in high school when IE methods are *not* used, as well as when multiple representations are *not* used. Research is also needed to clarify whether a correlation between representational consistency on the R-FCI pretest and G_{FCI} exists at the university level, where students probably have the competence to interpret the standard formats of representations used in the R-FCI.

ACKNOWLEDGMENTS

This study was supported by the Academy of Finland (Project No. 132316).

[1] J. Docktor and J. Mestre, *A Synthesis of Discipline-Based Education Research in Physics* (National Research Council, Board on Science Education, Washington, DC, 2010) [http://www7.nationalacademies.org/bose/DBER_Docktor_October_Paper.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] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Interpreting force concept inventory scores: Normalized gain and SAT scores, Phys. Rev. ST Phys. Educ. Res. **3**, 010106 (2007).

[5] D. Meltzer, The relationship between mathematics preparation and conceptual learning gains in physics: A possible ''hidden variable'' in diagnostic pretest scores, Am. J. Phys. **70**, 1259 (2002).

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

[7] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Why you should measure your students' reasoning ability, Phys. Teach. **45**, 235 (2007).

[8] V. P. Coletta, J. A. Phillips, and J. Steinert, FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects, AIP Conf. Proc. **1413**, 23 (2012).

[9] R. B. Kozma, The material features of multiple representations and their cognitive and social affordances for science understanding, Learn. Instr. **13**, 205 (2003).

[10] P. B. Kohl and N. D. Finkelstein, Patterns of multiple representation use by experts and novices during physics problem solving, Phys. Rev. ST Phys. Educ. Res. **4**, 010111 (2008).

[11] D. Hestenes, Modeling methodology for physics teachers, AIP Conf. Proc. **399**, 935 (1997).

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

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

[14] 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).

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

[16] P. Nieminen, A. Savinainen, and J. Viiri, Force Concept Inventory-based multiple-choice test for investigating students' representational consistency, Phys. Rev. ST Phys. Educ. Res. **6**, 020109 (2010).

[17] P. B. Kohl, D. Rosengrant, and N. D. Finkelstein, Strongly and weakly directed approaches to teaching multiple representation use in physics, Phys. Rev. ST Phys. Educ. Res. **3**, 010108 (2007).

[18] T. Seufert, Supporting coherence formation in learning from multiple representations, Learn. Instr. **13**, 227 (2003).

[19] I. Halloun, R. R. Hake, E. P. Mosca, and D. Hestenes, Force Concept Inventory, http://modeling.asu.edu/R&E/Research.html (password protected), revised 1995.

[20] 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).

[21] A. Savinainen and J. Viiri, The Force Concept Inventory as a measure of students conceptual coherence, Int. J. Sci. Math. Educ. **6**, 719 (2008).

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

[23] A. E. Lawson, The development and validation of a classroom test of formal reasoning, J. Res. Sci. Teach. **15**, 11 (1978).

[24] A. E. Lawson, Classroom test of scientific reasoning, http://www.ncsu.edu/per/TestInfo.html, revised 2000.

[25] L. Bao, K. Fang, T. Cai, J. Wang, L. Yang, L. Cui, J. Han, L. Ding, and J. Luo, Learning of content knowledge and development of scientific reasoning ability: A cross culture comparison, Am. J. Phys. **77**, 1118 (2009).

[26] A. E. Lawson, D. L. Banks, and M. Logvin, Self-efficacy, reasoning ability, and achievement in college biology, J. Res. Sci. Teach. **44**, 706 (2007).

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

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

[29] 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).

[30] R. Doran, *Basic Measurement and Evaluation of Science Instruction* (NSTA, Washington, DC, 1980).

[31] P. Nieminen, A. Savinainen, N. Nurkka, and J. Viiri, An intervention for using multiple representations of mechanics in upper secondary school courses, in *Proceedings of the ESERA 2011 Conference, Lyon, 2011*, edited by C. Bruguiere, A. Tiberghien, and P. Clement, http://lsg.ucy.ac.cy/esera/e_book/base/strand3.html, p. 140.