JYVÄSKYLÄ STUDIES IN EDUCATION, PSYCHOLOGY AND SOCIAL RESEARCH
470

Pasi Nieminen

Representational Consistency and the Learning of Forces in Upper Secondary School Physics

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN EDUCATION, PSYCHOLOGY AND SOCIAL RESEARCH 470

Pasi Nieminen

Representational Consistency and the Learning of Forces in Upper Secondary School Physics

Esitetään Jyväskylän yliopiston kasvatustieteiden tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen auditoriossa 2
toukokuun 31. päivänä 2013 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Education of the University of Jyväskylä,
in building Agora, Auditorium 2, on May 31, 2013 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2013

Representational Consistency and the Learning of Forces in Upper Secondary School Physics

JYVÄSKYLÄ STUDIES IN EDUCATION, PSYCHOLOGY AND SOCIAL RESEARCH 470

Pasi Nieminen

Representational Consistency and the Learning of Forces in Upper Secondary School Physics

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2013

Editors
Timo Saloviita
Department of Teacher Education, University of Jyväskylä
Pekka Olsbo,
Publishing Unit, University Library of Jyväskylä

URN:ISBN:978-951-39-5217-4
ISBN 978-951-39-5217-4 (PDF)

ISBN 978-951-39-5216-7 (nid.)
ISSN 0075-4625

Jyväskylä University Printing House, Jyväskylä 2013

ABSTRACT

Nieminen, Pasi
Representational consistency and the learning of forces in upper secondary school physics
Jyväskylä: University of Jyväskylä, 2013, 58 p.
(Jyväskylä Studies in Education, Psychology and Social Research
ISSN 0075-4625; 470)
ISBN 978-951-39-5216-7 (nid.)
ISBN 978-951-39-5217-4 (PDF)

This dissertation focused on multiple representations in of the force concept. Sociocultural views of learning consider that learning takes place in a cultural context via social processes wherein language plays a central role. In addition to talk and text, the language of physics includes a diverse set of other representations, such as graphs, vectors, and equations. To learn physical language, and thus to solve physical problems successfully, students must become competent in multiple representations. This means that when solving a problem, students must be able to interpret and construct different representations, identify their similarities and distinctions, and move between these representations. The benefits of multiple representations have been known in physics education research, but there has not been an appropriate research instrument to study students' representational consistency, i.e., their ability to interpret multiple representations. One contribution of this dissertation was to produce such instrument. The designed quantitative multiple choice test was administered to upper secondary school students (n = 322) in the Sub-studies of the dissertation. Data was also collected using other quantitative measures, open-ended exercises, teacher interviews, and video recordings. It was found that students' (n = 133) pre-instructional representational consistency was related to their conceptual learning of forces. In addition, observed gender differences in learning of forces were diminished when pre-instructional representational consistency and scientific reasoning were controlled in statistical analysis. Further, an intervention study for multiple representations was conducted in two upper secondary mechanics courses, and it was found that students' (n = 28) learning of forces seemed to increase compared to students who did not participate in the intervention (n = 22). However, this difference was not statistically significant.

The findings show that students' representational skills are related to their learning. Hence, multiple representations should be included in physics lessons so that students learn to interpret and construct different representations and move between them. The important role of multiple representations should be emphasised in teacher training. The research instrument designed in this dissertation is a methodological tool for the community of physics researchers and instructors. Dozens of researchers around the world have already requested a copy of the test. From a theoretical perspective, the findings of this dissertation are in tune with sociocultural views of learning, wherein language (including physical representations) plays a central role. In addition, learning theories concentrating on individual views are relevant in explaining the findings.

Keywords: physics teaching and learning, multiple representations, representational consistency, force concept

Author's address Pasi Nieminen
Department of Teacher Education
University of Jyväskylä, 40014
pasi.k.nieminen@jyu.fi

Supervisors Professor Jouni Viiri
Department of Teacher Education
University of Jyväskylä

Docent Antti Savinainen
Department of Teacher Education
University of Jyväskylä

Reviewers Docent Kalle Juuti
University of Helsinki

Associate Professor Charles Henderson
Western Michigan University

Opponents Docent Kalle Juuti

PREFACE

I acknowledge and thank the many individuals who have helped me with this dissertation or who have in many other ways supported and enriched my life during this long-term project. It is very clear that without the help of my supervisors, Professor Jouni Viiri and Docent Antti Savinainen, this work would not have even started. During the research process, I have always benefited from the excellent guidance of my supervisors. Professor Viiri is an important developer of research practices in our science and mathematics team (InClass), but also in the wider Faculty of Education. Cooperation in InClass has greatly advanced during the last few years, which has helped my work considerably. Thus, I thank all my colleagues in our InClass team and my department.

I am grateful my dissertation reviewers, Docent Kalle Juuti from University of Helsinki and Associate Professor Charles Henderson from Western Michigan University, whose remarks led to significant improvements. Further, I thank Docent Juuti acting as opponent in the dissertation defense. I also acknowledge the Academy of Finland, the Finnish Cultural Foundation, former University Rector Aino Sallinen, our department, and the Magnus Ehrnrooth Foundation for funding of this dissertation.

Educational research must be done in real-world education contexts. Hence, I thank the teachers and students who participated in the Sub-studies of this dissertation. Additional thanks go to the community of Lehtisaari lower secondary school, which was my workplace before and partly during the dissertation process. Further, I thank the students in our university whom I have had an opportunity to teach or supervise. These teaching experiences have given me many ideas and perspectives for my educational research.

Finally, I am very grateful to my parents and siblings for their care and support. I know I can always trust you – I appreciate this greatly. Thank you to my friends for the late night discussions and the music. My deepest thanks to my love Elina. I am also grateful to her family.

Jyväskylä 15.4.2013
Pasi Nieminen

AUTHOR'S CONTRIBUTION

Design of the studies and methods. The author participated in the planning of the Sub-studies. The author had main role in the design of the R-FCI test and the intervention material used in Sub-study IV.

Data collection. The author participated in the data collection. Different classroom tests in the Sub-studies were administered to students by teachers of courses. The author interviewed teachers and organized videotaping in Sub-study IV. The author videotaped a couple of lessons, and the rest were videotaped by the teachers (fixed camera).

Data analyses. The author had main role in the data analyses. The co-author Niina Nurkka analysed the video data in Sub-study IV based on the discussions with the author.

Findings and writing. All the authors participated in the writing of the articles, but the author (Nieminen) had the main role. Antti Savinainen wrote the descriptions of student groups and instruction, especially in the articles I–III.

LIST OF ORIGINAL PUBLICATIONS

I Nieminen, P., Savinainen, A., & Viiri, J. (2010). Force Concept Inventory-based multiple-choice test for investigating students' representational consistency. *Physical Review Special Topics – Physics Education Research 6*(2), 020109.

II Nieminen, P., Savinainen, A., & Viiri, J. (2012). Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning. *Physical Review Special Topics – Physics Education Research 8*(1), 010123.

III Nieminen, P., Savinainen, A., & Viiri, J. (in press). Gender differences in learning of the force concept, representational consistency, scientific reasoning. *International Journal of Science and Mathematics Education*.

IV Nieminen, P., Savinainen, A., Nurkka, N., & Viiri, J. (2012). An intervention for using multiple representations of mechanics in upper secondary school courses. In C. Bruguière, A. Tiberghien & P. Clément (Eds.), *E-Book Proceedings of the ESERA 2011 Conference: Science learning and Citizenship.* Part 3 (co-eds. Marisa Michelini and Reinders Duit), (pp. 140-147) Lyon, France: European Science Education Research Association. ISBN: 978-9963-700-44-8

FIGURES

TABLES

CONTENTS

1 INTRODUCTION

This dissertation focused on multiple representations in learning of the force concept in upper secondary school level. The use of multiple representations is an essential part of physical language, because it uses a diverse set of representations (e.g., graphs, vectors and equations) that are utilised in inter-personal communication, intra-personal reasoning, and various media such as books and computer applications. Hence, students should be familiarised with this language so that they will be able to understand physical concepts and solve physical problems (Van Heuvelen & Zou 2001, Mercer et al. 2004, Lemke 1990). Likewise, the concept of force is an important concept in classical physics, and it is a central concept in secondary school physics. However, it has been shown that students' conceptions of forces can differ outstandingly with the scientific force concept, and this difference is not easy to change (Halloun & Hestenes 1985a, Chi 2008). Thus, students' conceptual understanding and learning of forces is widely studied and appropriate tools for evaluating that understanding has been developed, such as Force Concept Inventory (FCI) (Hestenes, Wells & Swackhamer 1992) and Force and Motion Conceptual Evaluation (FMCE) (Thornton & Sokoloff 1998).

TABLE 1 Description of the Sub-studies of the dissertation.

Sub-study	Aim
I	To describe the purpose, design and validity of the R-FCI
II	To study the relationship between representational consistency, conceptual understanding of forces and scientific reasoning
III	To study gender difference in learning of the force concept when students' pre-instructional representational consistency and scientific reasoning is covariated
IV	To evaluate the effect of an intervention (focus on the use of multiple representations) on students' learning of the force concept

This dissertation consists of four Sub-studies (Table 1). A central contribution of the dissertation to earlier research is developing and validating of a conceptual

test for evaluating students' ability to interpret multiple representations (i.e., *representational consistency*) in the context of forces. Previous conceptual tests of physics were not adequate for that. The designed test – Representational Variant of the Force Concept Inventory (R-FCI) – was presented in Sub-study I. Sub-study II showed the relationship between students' representational consistency and their learning of the force concept. This can be considered quantitative evidence for benefits of multiple representations for learning.

Gender differences in physics have been studied widely. Generally males outperform females in participation, attitudes and achievement in physics (see a review in Section 2.4). In Sub-study III we reported big differences favouring males in the learning of the force concept, representational consistency and scientific reasoning. However, we found that the difference in learning the force concept did not exist when students' learning was covariated with pre-instructional representational consistency and scientific reasoning. This indicated that gender difference in learning was not produced by an instruction used, but was related to students' abilities before the instruction.

Sub-study IV presented an intervention for emphasising the use of multiple representations in upper secondary mechanics courses. The implementation of the intervention was quite light, as teachers did not received any extra training and the intervention exercises were used as homework, whereupon the teaching time used for the intervention was very limited (4 – 7 % of the total classroom time depending on a school). The results were promising as it seemed that the light intervention increased students' learning of the force concept and their representational consistency compared to students who did not participate in the intervention. However, the difference between the groups was not statistically significant. Thus, replication studies would be needed for confirm (or refute) the effect of intervention.

This dissertation has a following structure. In Chapter two, its' theoretical foundations are discussed, such as, individual and social views of learning and relevance of multiple representations in physics learning. Research aims are presented in Chapter three. Chapter four covers methodological issues including data collection and analyses, research designs in Sub-studies and statistical methods. Chapter five summarises the results of the Sub-studies, and finally, general discussion (Chapter 6) concludes the dissertation.

2 THEORETICAL FRAMEWORK

This chapter focuses on constructivist views of learning – individual and sociocultural – which are both useful in studying learning with multiple representations. It is followed by definitions of the concepts of representation and multiple representations, and a review the research on learning with multiple representations. This is followed by a discussion of the force concept and previous research on learning this concept. Reported gender differences in physics education are then presented. Finally, design and evaluation of teaching interventions in science education are discussed.

2.1 Individual and social views of learning

In the early twentieth century, research on learning was dominated by behaviourism (Woolfolk 2007, Tynjälä 1999, Kauppila 2007). Behaviourists focused on background factors that affect learning achievement, but they were not interested in students' mental processes while learning. Learning was understood via stimulus and response, in which a learner has a quite passive role as a responder to external stimuli. By the 1950s and 1960s, a shift toward constructivist views of learning took place. Constructivism covers many different views, but all of them see learning as an active and creative construction process in contrast to passive reception. The main difference between these views is in the following issue: who is the constructor of knowledge – the individual, group, or community? Thus, different constructive views can be classified into *individual* or *social* constructivism.

Learning physics with multiple representations (discussed in Section 2.2.2) can be viewed as an individual and/or social process. Hence, it can be examined, at least, from viewpoints of *cognitive constructivism* and *sociocultural approaches*, as Tynjälä (1999) has described them. These have also been referred to as *individual* and *social views* of learning (Driver et al. 1994, Leach & Scott 2003).

Cognitive constructivism is strongly influenced by Jean Piaget (1896–1980) and is mainly interested in individuals' biological and psychological learning mechanisms (Woolfolk 2007, Tynjälä 1999). Some relevant issues are, for example, how individuals process information in sensory, working, and long-term memory and how they improve their mental representations and models (i.e., learn), even without the aid of other individuals. In contrast, the sociocultural approach, largely based on the work of Lev Vygotsky (1896–1934), is not so much interested in the content of an individual's mind, but how knowledge is constructed in society, e.g., in a physics class as a social setting. This view considers that learning and knowledge construction are social phenomena that should be examined in their social, cultural, and historical frameworks (Tynjälä 1999, Kauppila 2007).

One important aspect relating to learning with multiple representations is Vygotsky's idea of describing language as a *cultural* and *psychological tool*. Vygotsky expanded Marxist thought that human ability to use physical tools is important to the development of societies (Tynjälä 1999, Mercer 2000). Vygotsky thought that, similar to how humans have created physical tools for the physical world, they have created a tool, language (i.e., symbols and semiotic systems), for helping their thinking and social behaviour. Different representational formats that are created and used in physics are also part of this language. Language is used as a cultural tool when knowledge is shared and jointly developed, and, it is used as a psychological tool in individual reasoning (Mercer 2000).

Critical analysis of constructivist theories is beyond the scope of this dissertation. Although the division into individual and social views has been presented before, both views recognise, more or less, the roles of individual learning and social interaction in learning (Woolfolk 2007, Tynjälä 1999). I prefer the view that both individual and social views are useful for understanding science learning (Driver et al. 1994, Leach & Scott 2003). On the other hand, student mental representations or models are not studied in this dissertation, although it is accepted that individual learning is related to changes in these mental structures. Further, this dissertation does not focus on social interaction among individuals. However, the importance of social language from a sociocultural approach is adopted: the ability to understand and use multiple representations is part of the social language of physics.

2.2 Representations in physics education

2.2.1 Internal and external representations

This section describes the term *representation* and its sub-concepts *internal* and *external representation*. Although only external representation is studied in this dissertation, the division into external and internal representation is briefly discussed as it is commonly used in the research literature. When the term repre-

sentation is used in this dissertation (outside of this section), it always refers to external representation.

A representation is "an object or an event that stands for something else" (Schnotz et al. 2010, p. 18). For example, a symbol ":)" has a meaning of a smile in modern-day written communication. An example of an event as representation could be the Eucharist in Christian tradition. Further, a representation can be an external or an internal in relation to the human mind. External representations exist in the physical world, e.g., the symbol ":)", the event Eucharist, or a picture painted by a child. In contrast, internal (mental) representations are the knowledge and structure in memory in different forms, such as propositions, productions, schemas and neural networks (Zhang 1997). A person's internal representations cannot be directly studied but they need to be studied via external representations that he or she expresses.

External representations are always present in communication when any kind of understanding has been constructed or shared. Interpersonal communications coordinate oral language, gestures, and visual representations (like graphs and pictures) to transmit the message clearly between the sender and receiver (e.g., teacher-students, student-teacher, or student-student). Likewise, external representations are used in various media, such as books or computer applications and networks. It has been stated that learning of science means that students need to learn to talk science (Mercer et al. 2004, Lemke 1990) and learn to fluently handle various modes of disciplinary discourse in which external representations have an essential role (Airey & Linder 2009). From this viewpoint, external representations are an intrinsic part of the language of science.

External representations can serve as cognitive tools for thinking (Schnotz et al. 2010, Zhang 1997). For example, instead of processing everything internally, a person can create an external representation to reduce the cognitive load. On the other hand, Zhang (1997) stresses that the external representations are more than inputs and stimuli to the internal mind or just memory aids: they can guide, constrain and even determine cognitive behaviour. Indeed, it is clear that the modern society would not have evolved to its present state without powerful external representations such as writing and numerals (see review in Zhang 1997).

2.2.2 Multiple representations and their benefits for learning

This section reviews research concerning multiple representations. The Sub-studies I–III focused on students' representational consistency, i.e., their ability to interpret standard formats of representations. Sub-study IV included the evaluation of students' ability to interpret and construct standard formats. Further, the intervention emphasising multiple representations was examined in this Sub-study.

The term *multiple representations* refers to circumstances where various representations are used for learning a concept or solving a problem instead of, for example, only verbal or mathematical representations. It has been shown

that expert scientists are able to fluently use multiple representations and move between representations when they are thinking and sharing ideas (Kozma 2003, Kohl & Finkelstein 2008). There is a lot of research stating the importance of multiple representations in science learning in terms of successful problem solving and a good conceptual understanding. According to Ainsworth (2006), the functions of multiple representations in learning can be devided into three parts. First, multiple representations can complement each other because they 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 so. A second function of multiple representations is to help students develop a better understanding of a domain by using one representation to constrain their interpretation of a second one. For instance, graphs can be used to constrain the interpretation of equations. Thirdly, multiple representations can support the construction of deeper understanding when students integrate information from more than one representation.

Likewise, Van Heuvelen and Zou (2001) justify 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 mathematical symbols meaning. 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.

Multiple representations have been studied from many perspectives. Many researchers have been studied students' ability to interpret and construct standard or canonical formats of representations (e.g., graphs, vectors, and bar charts) and students' ability to select the appropriate format in different situations (Meltzer 2005, Kohl & Finkelstein 2005, Acevedo Nistal et al. 2010). Students' ability to create their own representations has also been studied (diSessa 2004). Physics education research has shown that an instructional approach emphasising multiple representations is helpful for students' use of multiple representations when the approach is strongly or weakly directed (Kohl, Rosengrant & Finkelstein 2007). In the chemistry education context as well, it has been reported that students' learning from multiple representations can be supported by directive and non-directive help, depending on their prior knowledge (Seufert 2003).

2.3 Learning of the force concept

2.3.1 Scientific formulation of the force concept

All Sub-studies of this dissertation include evaluation of students' understanding of the force concept, i.e., Newton's laws and related kinematics. Related kinematics means that in order to understand Newton's laws, students need to

understand many kinematics concepts, such as position, velocity and acceleration. Scientific formulation of Newton's laws differs slightly in literature. The next formulation was translated from Physica 1 (Hatakka et al. 2004), which is the textbook used in Sub-studies I–III:

- Newton's I law: An object continues its rectilinear motion with constant velocity or stays at rest, if it does not interact with other objects.
- Newton's II law: An external force $\overline{F}$ that acts on an object produces acceleration $\overline{a}$ on the object.
- There is relation between the force acting on the object, its mass, and its acceleration
  $\overline{F} = m\overline{a}$.
- Newton's III law: Force and opposite force are created from an interaction between two objects. The force and opposite force are equal in magnitudes but opposite in directions, and they act on different objects.

Formulation of Newton's laws varies between different textbooks. Physica 1 starts the formulation from the concept of *interaction* (between objects). Then it explains that interaction between two objects generates two forces and defines the force as magnitude of interaction. Hence, the formulation of Newton's III law in Physica 1 includes the concept of interaction, which is not often presented in the formulation of Newton's III law. For example, Giancoli (2005, p. 78) presents Newton's III law as follows: "Whenever one object exerts a force on a second object, the second object exerts an equal force in the opposite direction on the first".

After Newton's III law, Physica 1 covers the I and II laws. It must be noted that formulation of Newton's I law often includes the concept of *net force*, which is not included in Physica's formulation (see above). For example, Giancoli (2005, p. 74) presents Newton's I law as follows: "Every object continues in its state of rest, or of uniform velocity in a straight line, as long as no net force acts on it". In contrast, Physica 1 presents net force separately after the three Newton's laws have been concerned. Under the title "unchangeable motion" it is stated that, "Object stays at rest or continues its rectilinear motion, if net force acting on it is zero" (Hatakka et al. 2004, p. 89). After that Physica 1 underlines that the clause is not same as Newton's I law because the object interacts with other objects.

2.3.2 Students' conceptual understanding and learning of the force concept

There is a large body of research concerning students' conceptions of physics that shows that students' concepts can be considerably inconsistent with corresponding scientific concepts (Halloun & Hestenes 1985a, Hestenes, Wells & Swackhamer 1992, diSessa 1993). These conceptions have been referred by various terms like preconceptions, misconceptions, alternative conceptions or phenomenological primitives (p-prims). Further, it has been stated that instruction

that does not take these conceptions into account is probably ineffective in terms of learning (Halloun & Hestenes 1985b).

Students' conceptions about forces have been widely studied. The FCI includes various alternative student concepts, such as the "dominance principle" (Hestenes et al. 1992). It means, for example, that during an interaction of two bodies, the greater mass exerts the greater force. The taxonomy of the misconceptions included the FCI is presented in Hestenes et al. (1992) and in Hestenes & Jackson (for the revised 1995 version of the FCI). Because the R-FCI items are based on nine selected FCI items (1, 4, 13, 17, 22, 24, 26, 28 and 30), the taxonomy is partly suitable for the R-FCI too.

The notions of *conceptual understanding* and *learning* of the force concept have been frequently used in this dissertation. Conceptual understanding of the force concept has been used for referring to a relationship between a student's conception of the force and corresponding scientific concept in a moment of measuring. Measuring has been performed using the FCI or R-FCI (the measures are described in Section 4.1). These tests provide an opportunity to select a scientific view (correct understanding) or various alternative non-scientific views (incorrect understanding). Further, learning of the force concept has been used to denote the possible change in conceptual understanding that has been measured by FCI or R-FCI after and before an instruction.

2.4 Gender differences in physics education

Sub-study III focused on gender differences in upper secondary school students' learning of forces, representational consistency and scientific reasoning when interactive engagement (IE) teaching method was used. Hence, only gender differences concerning these issues are shortly described here. A slightly wider review of gender differences is in the article of Sub-study III.

Gender differences favouring males in participation, attitudes and achievement in physics have been widely reported. For example, male students in high schools and universities achieve higher learning gains on forces and kinematics than their female peers (Lorenzo, Crouch & Mazur 2006, Coletta, Phillips & Steinert 2012, Pollock, Finkelstein & Kost 2007) when learning is measured using FCI and Force and Motion Conceptual Evaluation (FMCE) (Thornton & Sokoloff 1998). There is evidence that learning gains on force concept are associated with students' scientific reasoning ability (Coletta, Phillips & Steinert 2007) and that there are gender differences reported where males had both higher scientific reasoning ability and learning gain of the force concept (Coletta, Phillips & Steinert 2012). The factors that may influence women's participation, attitude, and achievement in physics are complex and interlaced, such as sociocultural forces or some cognitive abilities (Halpern et al. 2007). Physics education researchers and teachers are probably able to affect only a limited number of these factors; therefore, it is important to identify factors that

can be addressed in physics teaching and that can affect learning, which may reduce gender differences in physics.

Lorenzo et al. (2006) review seven teaching strategies that can help to narrow gender differences, i.e., integration of everyday experiences and interests, assessment and use of students' prior knowledge, interactive environments enhancing cooperation and communication, activities fostering students' understanding, activities decreasing competitiveness, alternation between group discussion and structured teaching, and diverse and frequent assessment practices with feedback. Interactive engagement (IE) teaching methods include some of these characteristics, as IE methods have been developed for improving teaching, holding students' attention, and increasing teacher-student and student-student interaction (Hake 1998). Indeed, Lorenzo et al. (2006) stated that gender difference in FCI gain was reduced when IE method was used. However, Pollock et al. (2007) provided evidence that IE method cannot explain the reduction alone, but the instructor effect is significant also.

2.5 Design and evaluation of teaching interventions

Mehéut and Psillos (2004) review studies concerning *teaching-learning sequence* (TLS), which is a generally used term in the science education research community referring to topic-oriented sequences instead of long-term curricula. These sequences are founded on research knowledge about, e.g., students' conceptions, features of the specific scientific domain and educational context. TLS studies also include evaluation of student learning. Another name, which is used in studies of teaching-learning practices, is *teaching intervention*. Millar, Leach, Osborne and Ratcliffe (2006) have defined teaching interventions using the concepts *research evidence-informed* and *research evidence-based*. Evidence-informed refers to a *design* of an intervention. An intervention can be considered as research evidence-informed if the design is based on ideas and findings from research. In contrast, evidence-based refers to *learning outcomes* of an intervention. A teaching intervention is research evidence-based when its positive effect on students' learning outcomes has been supported by evidence. This evidence is gathered via systematic data collection and analysis that are open to public scrutiny and thus, they are possibly replicable. Hence, one intervention can be research evidence-informed without evidence about learning outcomes. On the other hand, another intervention can be research evidence-based without influences of research on design (Millar et al. 2006). One can argue that for an intervention or a TLS, which is valuable for researchers and teachers of the science education community, research should be both evidence-informed and -based.

In Sub-study IV we used the term an intervention instead of a sequence, because we felt it fit better with the characteristics of our study. For example, our intervention was not strictly restricted in terms of timetable and ways of implementing. Otherwise, our intervention has a lot of common with TLS and

teaching interventions as they were defined before. Our intervention was research evidence-informed, which is discussed in Section 4.3.2. Further, students' learning during the intervention was evaluated in order to call it research evidence-based.

In intervention studies, design and implementation (i.e., teaching) have been often conducted by one person (teacher-researcher) or at least co-operation has been close between designers and teachers. Leach, Scott, Ametller, Hind and Lewis (2006) studied the transferability of short teaching interventions and their learning outcomes. They used the term *baseline teacher* for teachers who took part of the design of an intervention. Students' learning outcomes were studied in classes of baseline teachers. Then, interventions were given, with minimal training, to *transfer teachers* who did not participate in the design phase. Students' learning in classes of transfer teachers was studied in order to examine transferability of learning outcomes between baseline and transfer classes. These interventions were evidence-informed, because they were carefully based on research knowledge and designed in co-operation between researchers and baseline teachers in a real school context. Studying learning outcomes in classes of transfer teachers, researchers aim to show that these interventions were also research-based.

We adopted the term transfer teacher in Sub-study IV, although the baseline data in our study was collected in a different way than in the studies of Leach and others (2006). However, our intention was similar: our intervention was founded on research knowledge, and it was piloted in a school. Further, we wanted to study the effect of our intervention in classes of transfer teachers (i.e., teachers who did not participate in the design phase).

3 RESEARCH AIMS

The main aim of this work was to examine students' ability to interpret and construct multiple representations and the relation of abilities to learning of the force concept.

1. The aim of Sub-study I was to design a test for evaluating students' representational consistency and collect data for the validation of the test.
2. The aim of Sub-study II was to study relations between students' learning of the force concept, representational consistency and scientific reasoning.
3. The aim of Sub-study III was to study gender differences in learning of the force concept, representational consistency and scientific reasoning. In addition, the purpose was to study gender difference in learning of the force concept when pre-instructional level of representational consistency and scientific reasoning were used as covariates.
4. The aim of Sub-study IV was to design and conduct an intervention that stresses the use of multiple representations in upper secondary mechanics course. The effect of an intervention on students' learning of forces was evaluated using pre- and post-testing in intervention and baseline groups.

4 RESEARCH METHODS

Section 4.1 introduces the constructs for students' understandings and abilities, and the instruments that were used to measure these. Section 4.2 describes the procedures for analysing data collected by these instruments. Research designs including participants are presented in Section 4.3. Section 4.4 introduces the normalised learning gain as a measure for learning. Finally, the used statistical analyses are covered in Section 4.5.

4.1 Measured constructs and used instruments

Three multiple-choice tests were used for evaluating students' understandings and abilities (Table 2). Students' understanding of the force concept was documented by Representational Variant of the Force Concept Inventory (R-FCI) (Nieminen, Savinainen & Viiri 2010) and by Force Concept Inventory (FCI) (Halloun et al. 1995). The FCI covers wider the force concept than R-FCI does, but the correlation between FCI and R-FCI scores is very strong (post-tests, r = .86, n = 87) (Nieminen, Savinainen & Viiri 2010) indicating that the tests are measuring the same construct. The R-FCI has been used for documenting students' representational consistency (RC), that is, their ability to interpret multiple representations consistently. RC score is a different measure from R-FCI score (total of correct answers) although both measures are calculated from the same instrument. Definition for RC score is given in Section 4.2.2. Students' scientific reasoning ability was documented by Classroom Test of Scientific Reasoning (CTSR) (Lawson 2000).

TABLE 2 Measured constructs by multiple-choice tests in the studies

Construct	Measure	Instrument	Sub-study
Conceptual understanding of force	R-FCI score	R-FCI	I – IV
Conceptual understanding of force	FCI score	FCI	I, II
Representational consistency (RC)	RC score	R-FCI	I – IV
Scientific reasoning ability	CTSR score	CTSR	II, III

Note. The scoring rules for the different measures are given in Section 4.2.

4.1.1 Force Concept Inventory

The Force Concept Inventory (FCI) (Hestenes, Wells & Swackhamer 1992) is a widely used test for assessing students' understanding of the force concept (Hestenes & Halloun 1995). The FCI has gone through a lengthy process of validation and its reliability has been well established (Savinainen & Viiri 2008). The 1995 version (Halloun et al. 1995) contains 30 multiple-choice items with five response alternatives (one correct Newtonian alternative and four incorrect common-sense alternatives) that cover the most basic concepts in Newtonian physics. Most of the items are presented in verbal representation, but some items contain visual information.

There is some research regarding the possible gender bias in the FCI items. McCullough (2004) reported significant gender differences in certain FCI items when the contexts of items were modified from stereotypical male ones (e.g., cannonball or hockey) to female ones (e.g., shopping or cooking). Female respondents did not perform better in female contexts on average, but male respondents' performance decreased in these contexts. On the other hand, Osborn Popp, Meltzer, and Megowan-Romanowicz (2011) analysed the FCI data of 4775 high school students using differential item functioning analysis to detect possible gender bias in the items. These authors found three possibly biased items, two favouring males and one favouring females. However, they concluded that the FCI items were not systematically biased in favour of male respondents. These findings also support the validity of the R-FCI as a research instrument for gender analysis, as FCI items are used for the R-FCI. It is noteworthy that the three possibly biased FCI items are not included in the R-FCI.

However, the article of Osborn Popp et al. (2011) is based on a conference presentation. It is not published in a peer-reviewed journal which would increase the validity of the study. Hence, discussion of gender bias of the FCI is probably not finalized.

4.1.2 Representational Variant of Force Concept Inventory

The design and application of the Representational Variant of Force Concept Inventory (R-FCI) was an important aim in my thesis. The purpose, design and validity of this test were the topics of Sub-study I. However, as the test was also one research instrument in all the studies, the structure of the test and the consistency analysis are presented here in the Research Methods section. The valid-

ity of the R-FCI was established in Sub-study I and it will be discussed in Section 5.1.2 and 6.3.1.

The R-FCI is based on nine verbal items of the Force Concept Inventory (Halloun et al. 1995) concerning Newton's laws and gravitation in different contexts. Two new isomorphic variants (physical concept and context as similar as possible) were formulated with different representations (graphs, vectors, motion maps or bar charts) for each of the nine FCI items, which resulted in a triplet of items. The term *theme* is used for these triplets of isomorphic items that consist of an original FCI item and two isomorphic variants (Table 3). Altogether there are nine themes in the R-FCI, totalling 27 items (9x3). Themes are named according to an original FCI item. Theme 4 (T4), for example, refers to item 4 in the FCI. Figure 1 presents Theme 4, which deals with Newton's third law (forces exerted during a collision). All three items presented include an identical verbal question, which is not presented here to preserve the confidentiality of the original FCI items. Performance on the themes illustrated students' representational consistency (see Section 4.2.2).

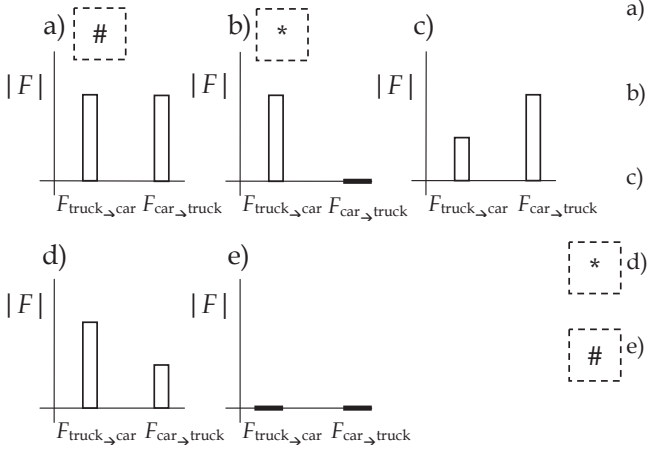
TABLE 3 Themes of the Representational Variant of Force Concept Inventory (R-FCI) in terms of concept and representation of items

Theme	Concept	Verbal	Graphical	Vectorial	Motion map	Bar chart
T17	NI	x		x		x
T24	NI	x	x		x	
T22	NII	x	x		x	
T26	NII	x	x		x	
T4	NIII	x		x		x
T28	NIII	x		x		x
T1	gravitation	x			x	x
T13	gravitation	x	x	x		
T30	gravitation	x		x		x

4.1.3 Classroom Test of Scientific Reasoning

The Classroom Test of Scientific Reasoning (CTSR) (Lawson 1978) was designed to assess a student's formal reasoning. The version used in this study contains 24 multiple-choice items concerning conservation of mass and volume, proportional reasoning, control of variables, probabilistic reasoning, correlational reasoning, and hypothetico-deductive reasoning (Lawson 2000). Bao et al. (2009) provided a detailed description and classification of the items. Lawson, Banks, and Lovgin (2007) found .79 KR-20 reliability coefficient with a sample of 459 college students; they reported that the validity of the original test version had been established by several studies.

Bar chart item



Two representationally consistent answer patterns:
#: scientifically correct item
***: scientifically incorrect item**

Verbal item (an original FCI item)

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.

Vectorial item

Let us denote the force exerted by the truck on the compact car as $F_{truck \to car}$ and the force exerted by the compact car on the truck as $F_{car \to truck}$. Which of the following alternatives best describes the average forces exerted on the truck and the compact car during the collision?

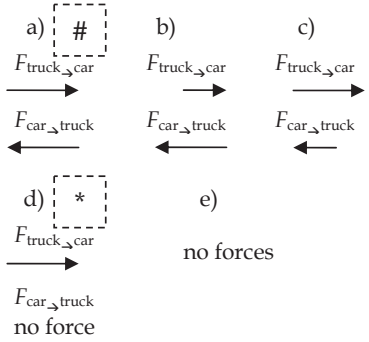


FIGURE 1 The theme 4 of the R-FCI and two representationally consistent answer patterns.

4.2 Analysis of test data

4.2.1 R-FCI, FCI and CTSR scores

All used tests consist of multiple-choice items, which can be answered either scientifically correctly or incorrectly. A usual way to use a test data is to calculate test score or "raw score", which is the number of correctly answered items. All used tests were scored this way. R-FCI and FCI scores evaluate the level of students' conceptual understanding of forces and CTSR scores students' scientific reasoning ability. The R-FCI can be used also for evaluating students' representational consistency, as the next section describes.

4.2.2 Representational consistency score

In the Sub-studies the representational consistency (RC) is defined to be a measure for the consistent interpretation of multiple representations, regardless of whether students' answers are scientifically correct or not. Figure 1 shows two answer patterns that are fully representationally consistent. In one pattern, all selected items are scientifically correct; in another pattern, all are scientifically incorrect (common-sense alternatives selected in each case within the theme). In both patterns, the selected alternatives of the theme correspond in terms of representations, regardless of scientific correctness. To determine students' RC, their answers in a given theme were scored in the following way:

- 2 points, if they selected corresponding alternatives in all three items of the theme
- 1 point, if they selected corresponding alternatives in two of the three items of the theme
- 0 points, if no corresponding alternatives in the items of the theme were selected.

There are five possible answer patterns for a theme that are fully consistent (2 points): one in which all items are scientifically correct and four in which items are answered scientifically incorrectly. There are many more patterns that are partially consistent (1 point) or non-consistent (0 point). Table 4 shows explicit examples for this scoring rubric for five student response patterns regarding the theme presented in Figure 1.

The RC score used in our analysis is determined as the sum of the scores of the individual themes; thus, the maximum score is 18 points (nine themes, 2 points maximum from each theme). The terminology has evolved between the first published article (Sub-study I) and the latest one (Sub-study III). Terms used in Sub-study I are explained in Section 5.1.2.

A spreadsheet was developed for consistency analysis of R-FCI data; students' multiple-choice response patterns were scored using the coded categorisation rules. Hence, there was no researcher error and, thus, no requirement for an inter-rater reliability concerns.

TABLE 4 Examples regarding the grading system for consistency for theme 4 (see Figure 1).

Exemplar selection			Representational consistency points
Bar chart item	Verbal item	Vectorial item	
a	e	a	2
a	e	e	1
a	b	e	0
b	d	d	2
b	d	e	1

4.3 Research designs in Sub-studies

Table 5 shows an overview of the aim, design and total number of participants in different Sub-studies. Sub-studies I–III used roughly speaking the same design, so they are considered together in Section 4.3.1. The design of Sub-study IV is described in Section 4.3.2.

TABLE 5 Overview of research aims, designs and total number of participants in the Sub-studies.

Sub-study	Aim	Design	$n_{students}$	$n_{teachers}$
I	To describe the purpose, design and validity of the R-FCI	One-group pre-test–post-test design	224	1
II	To study the relationship between representational consistency, conceptual understanding of forces and scientific reasoning	One-group pre-test–post-test design	131	1
III	To study gender difference in learning of the force concept when students' pre-instructional representational consistency and scientific reasoning is covariated	One-group pre-test–post-test design	131	1
IV	To evaluate the effect of an intervention (focus on the use of multiple representations) on students' learning of the force concept	Quasi-experimental pre-test–post-test design	50	2

4.3.1 Designs in Sub-studies I–III

Research design in Sub-studies I – III was a one-group pre-test–post-test design (Gall, Gall & Borg 2003), which means that all participants were considered as one group, and random sampling or control groups were not used. Students were tested using different multiple-choice tests (two to three tests per study) so that each test was administered to students before and after instruction. The exception was Sub-study I, in which one-group consideration concerned only 168 students of the whole sample (n = 224): the data of 56 students were used

for validation purposes only. This is explained later on. The research undertaken in the Sub-studies was approved by the school principal. Students took tests as part of the course and they were told that their answers will be handled confidentially.

Instruction

One-group design was appropriate, because all the students (except the aforementioned 56 students) in Sub-studies I, II, and III were studying their first physics course in upper secondary school and they were taught by same teacher. In addition, statistically significant differences in students' pre-instructional abilities or understanding (R-FCI, RC, FCI or CTRS score) were not found (Nieminen, Savinainen & Viiri in press, Nieminen, Savinainen & Viiri 2012b). The interactive engagement (IE) teaching method was used with emphasis on multiple representations. The first upper secondary course included many topics covering physics generally. Thus the course did not concentrate on the force concept (which was the subject concept in the studies), albeit the force concept was carefully taught. The instruction is described in the article of Sub-study III. As discussed in Section 2.4, IE methods have been developed for improving teaching, holding students' attention, and increasing teacher-student and student-student interaction.

Participants

The participants in Sub-study II and III were the same students (n = 131; 45 males and 86 females). They took R-FCI and FCI pre- and post-tests, and CTRS pre-test. Sub-study II used data of all the three tests, and Sub-study III the data of R-FCI and CTSR. In Sub-study I 168 students (males 63 and females 105) took the R-FCI test. Some of the students also took the FCI (n = 87), and 83 of them took all the R-FCI, FCI and CTRS. These 83 students were included also in Sub-study II and III.

In contrast to other students in Sub-studies I–III, Sub-study I included also a group of 56 students who were second-year upper secondary students. However, all students had the same teacher. The data of the group of 56 students was used only for validation for the R-FCI (see Section 5.1.2). Thus excluding this validation data, the results presented in Sub-study I collected a quite homogenous group of first-year students.

4.3.2 Design in Sub-study IV

Sub-study IV was a quasi-experimental pre-test–post-test design, which means that random sampling was not used, but there were treatment and control groups. These were called intervention and baseline groups, respectively. The research aim was to study the effect of a *research evidence-informed intervention* on students' learning of the force concept when students were taught by *transfer teachers* (see definition of concepts in Section 2.5). The intervention emphasised

the use of multiple representations in the context of force and kinematics. Our intervention was a research evidence-informed intervention because it was founded on knowledge about students' learning of multiple representations and the force concept. The research group (Pasi Nieminen, Antti Savinainen, and Jouni Viiri) designed the intervention and Antti Savinainen piloted it in an upper secondary school course in 2008. After the piloting, we made some minor clarifying improvements to the intervention material. Our intention was to study the effect of the intervention on students' learning of forces in classes of transfer teachers.

The intervention material

The intervention material consists of seven exercises (three to six sub-items per exercise). These open-ended, paper-and-pencil exercises emphasise the use of multiple representations and guide movement between the representations in different contexts that address the force concept. The Appendix shows an intervention exercise and a model answer written. All exercises are freely available on the Internet (Nieminen, Savinainen & Viiri 2012a).

Participants and data collection

The intervention was implemented in two upper secondary schools in fall of 2009. The most important criterion for selection of teachers was experience and willingness to participate. The teachers were found through personal contacts of the research group. Both selected teachers had over 10 years' experience in teaching upper secondary school physics. Permission to participate in the study was sought from the principals of the schools. In addition, parents were informed about the study: They were told that student answers would be handled confidentially and that only teachers' actions would be video recorded. The schools are referred as School 1 and School 2, and the teachers as Teacher 1 and Teacher 2, respectively. The students (aged 17) were taking their fourth physics course (mechanics, total duration ca. 20 hours).

Prior to the beginning of the courses, we gave the intervention material and model answers for the exercises. We did not want to interfere too much in the teachers' plans for the course, hence we only made suggestions about when to use a certain exercise during the course in relation to the content of the course. We did not provide instructions on how to use the material.

In the intervention courses, all lessons were video recorded and the teachers were interviewed after the course. The students' answers for the intervention exercises were collected before the teacher gave the correct answers. We did not, however, get all of the students' answers for all of the seven exercises. The exercises were given at different points during the course and some students were not in school on certain days or did not complete certain exercises for one reason or another. In the analysis of the intervention exercises, we have included students (n = 21) who completed four to seven of the exercises.

To evaluate the effect of the intervention on students' conceptual understanding of forces and representational consistency, some baseline was needed. For this purpose, the R-FCI pre- and post-test was administered in the intervention (n = 28) and baseline (n = 22) courses. Intervention and baseline courses were the same mechanics courses taught by the same teacher in different academic years. The main differences were that the intervention was not used in the baseline course and, of course, the different year-groups contained different students. Table 6 indicates the types of data collected in the baseline and intervention courses.

TABLE 6 Data collected in the two schools of Sub-study IV.

Course	Year	Data			
		R-FCI (pre-post)	Videos (all lessons)	Teacher interviews	Students' answers for the intervention exercises
Baseline	2008	x	-	-	-
Intervention	2009	x	x	x	x

4.4 Measuring of learning in pre-post-design

One important aspect in educational research is how to assess student learning during a period (e.g., a course or an intervention). The simplest way in a quantitative pre-post-design, is calculate the difference between post- and pre-test score (actual gain). However, this is not very usable: for example, if a student achieves an almost perfect score in pre-test, he or she cannot achieve very high absolute gain. One measure, which solves the described problem, is the average normalised learning gain (Hake 1998). It is defined as the ratio of the actual gain to the maximum possible gain:

$$G = \frac{\text{Post-test \%} - \text{Pre-test \%}}{100\ \% - \text{Pre-test \%}}$$

The average normalised learning gain is often used for measuring the change in a class of students (i.e., pre-test and post-test scores are class averages), but the formula above can be used for evaluating individual students' learning gain. In the latter case, G is called a single student normalised gain, and the pre-test and post-test scores in the formula refer to a single student. The single student gain values were used in these studies because individual student performances were needed for statistical analysis.

4.5 Statistical analyses

Table 7 shows statistical methods that were used for studying statistical significance and effect size in comparisons between and within student groups, and dependence between different variables. Methods for testing statistical significance are generally used in educational research and they are presented in common method books (e.g., Gall, Gall & Borg 2003, Robson 2002, Metsämuuronen 2009).

Statistical significance is considerably influenced by sample size: When sample size is large, quite small difference between student groups can be statistically significant, while its practical or theoretical significance can be only small (Gall, Gall & Borg 2003). Hence, American Psychological Association (2010) recommends that besides statistical significance, an effect size should be presented too. An effect size is "an estimate of the magnitude of a difference, a relationship, or other effect in the population represented by a sample" (Gall, Gall & Borg 2003, p. 624). The main approaches to measuring an effect size are "evaluate the proportion of variance explained" and "calculating standardized measures of differences in statistics" (Robson 2002, p. 402). Cohen's d (1988), which was used in Sub-studies III and IV, belongs in the last-mentioned group. It was calculated as the ratio of mean difference and pooled standard deviation of two independent samples (Metsämuuronen 2009, p. 472).

In Sub-study IV effect sizes were also calculated for related samples, that is, for the change between pre- and post-test results within a student group. Herein Cohen's d is not an appropriate effect size measure by reason of a correlation between pre- and post-test results. Cohen's d must be divided by the term that includes the correlation coefficient (see equation 8 in, Morris & DeShon 2002). Now calculated values (for related measures) are comparable to Cohen's d values (for independent groups). These effect sizes were not presented in the Article IV, but they have been added in the last paragraph of Section 5.4.2.

The use of parametric statistical methods requires certain assumptions (Gall, Gall & Borg 2003). A variable needs to be measured in interval scale, it should be approximately normally distributed and variances of comparison groups are about equal. If these assumptions are violated, non-parametric methods should be used. Non-parametric statistics can be used in any case, but when assumptions for parametric methods are valid, these methods should be used, as they can be more powerful compared to non-parametric methods.

However, research methods do not give strict guidelines for the evaluation of normality of a distribution, but it is case-specific. In addition, many parametric methods are quite robust for the violation of normality assumption. Spearman's ρ (non-parametric) was used for correlation analyses in Sub-study II, because some distributions seemed to be a little skewed. However, the same analyses were conducted with parametric Pearson's r in Sub-study III (the same data in both studies). The reason was that parametric methods were used in

Sub-study III, especially parametric ANCOVA, when some variables were re-estimated and they were found to be normal enough for parametric analyses.

TABLE 7 Different statistical methods used in the Sub-studies.

Method	Purpose	Sub-study
Mann-Whitney *U*-test	Non-parametric test for comparison between two independent samples. For ordinal or interval scales.	I, III
Wilcoxon signed-rank test	Non-parametric test for comparison between two paired samples. For ordinal or interval scales.	II
McNemar's test	Non-parametric test for comparison between two paired samples. For nominal and dichotomous data.	I
Kruskall–Wallis test	Non-parametric test for comparison between two or more independent samples. For ordinal or interval scales.	II
Independent samples t-test	Parametric test for comparison between two independent samples.	III
One-way analysis of variance (ANOVA)	Parametric test for comparison between more than two independent samples.	II, III
Analysis of covariance (ANCOVA)	For comparing means of a dependent variable between two of more student groups when the effect of other variable or variables (covariates) has been controlled.	
Cohen's *d*	An effect size that is defined as the standardised difference between two means	III, IV
Spearman's rank correlation coefficient (ρ)	Non-parametric measure for statistical dependence between two variables. For ordinal or interval scales.	I, II
Pearson *r*	The product-moment correlation coefficient for linear dependence between two interval variables that are approximately normally distributed	III

Five statistical indices were used to evaluate test reliability in Sub-study I (Table 8). KR-20 was used also in Sub-study III. These indices are closely described in Sub-study I.

TABLE 8 Five statistical indices used for test reliability in Sub-studies I and II (KR-20 only)

Measure	Purpose
Item difficulty index (*P*)	Difficulty of a test item
Item discrimination index (*D*)	Discriminatory power of a test item
Point biserial coefficient (r_{pbi})	Indicates how consistently an item measures students' performance in relation to the whole test
Kuder-Richardson reliability index (KR-20)	A measure for internal consistency of a test when items are dichotomous (e.g., correct or incorrect)
Ferguson's delta	Discriminatory power of a test

5 SUMMARY OF RESULTS

5.1 Sub-study I: Force Concept Inventory-based multiple-choice test for investigating students' representational consistency

5.1.1 Aims

Sub-study I presents the structure of the R-FCI, the R-FCI pre- and post-test results, and reliability data for sample of 168 upper secondary school students. The designed R-FCI test was used as pre- and post-test instrument also in Sub-studies II–IV. Hence, the R-FCI is presented in research methods (Section 4.1.2).

5.1.2 Results

Validity and Reliability

One purpose of Sub-study I was to establish validity of the R-FCI. The starting point for the design of the R-FCI was the FCI, which have been considered to be a valid test for measuring students' conceptual understanding of forces and related kinematics (Savinainen & Scott 2002). The R-FCI is based on nine of the thirty 30 FCI items, and therefore do not cover all Newtonian dimensions that are included in the FCI. However, the strong positive correlations were found between R-FCI and FCI pre-test (r = .78) and post-test (r = .86) scores among the subsample of students (n = 87) who took both R-FCI and FCI tests. These correlations indicate that the tests are measuring the same construct. Further, the correlations can be considered as evidence about construct validity of the R-FCI.

We studied correspondence of students' (n = 104) multiple-choice answers on R-FCI and written justifications why they chose a certain multiple-choice alternative. We found good compatibility: 92% of the correct answers were accompanied by correct explanations. However, this was done for four of the nine R-FCI themes. The same analysis was done earlier for other five themes (Savinainen et al. 2007). Five different statistical indices (see Section 4.5) were calcu-

lated for the reliability of the R-FCI. Table VII in the Article I show the results that indicate that the R-FCI has sufficient discriminatory power, and it is a reliable instrument for measuring single students and groups.

Representational consistency and the effect of the representational format

This study presents results concerning students' representational and scientific consistency. Representational consistency is a measure for the ability to interpret multiple representations consistently, regardless of scientific correctness (see Section 4.2.2). Scientific consistency is a sub-concept of representational consistency. When a student exhibits full scientific consistency in a certain theme he or she answers scientifically correctly all three items this theme (see pattern # in Figure 1). Students were placed into three classes (consistent, moderately consistent and inconsistent) according to their representational consistency score (Table 9). The classification has been explained in the section "Categorization of consistency" in the Article I.

TABLE 9 Levels of representational and scientific consistency (n=168).

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

Firstly, one statement in the "Discussion and Conclusion" section must be re-examined and criticised. As Table 9 shows, it was found that students exhibited low levels in scientific consistency: none of the students reached a "scientific consistent" level in the pre-test and in the post-test only 11% did so. In contrast, it was found that the means of students' R-FCI score (the sum of correct answers; measures conceptual understanding of forces) were 24% in pre- and 61% post-test (Table IV in the Article I). Thus, the R-FCI scores tended to be higher than the numbers concerning scientific consistency. Hence, it was concluded: "These data suggest that scientific consistency in the use of representations is quite a demanding skill that is not directly predictable from the raw score [R-FCI score] average." This conclusion is a little flawed. The reason lies behind different measures: R-FCI scores were presented by mean and scientific consistency was not. Instead, the mean of students' scientific consistency score was 18% in pre- and 58% in post-test (can be referred from bottom line in Table X in the Article I) that were quite near to the R-FCI scores (24% and 61%, respectively). Also, scatter plots (Figures 2 and 3) for scientific consistency and R-FCI score illustrate the strong relation of these measures: the Spearman's rho correlation coefficient for scientific consistency score and R-FCI score is .96 ($p < .001$, $n = 168$) in pre-test and .98 ($p < .001$, $n = 168$) in post-test. These analyses indi-

cate that the conclusion was not carefully reflected. However, scientific consistency and R-FCI score are not the same measure, and it is true that scientific consistency is more demanding than R-FCI score, but the difference is not so big as it was presented.

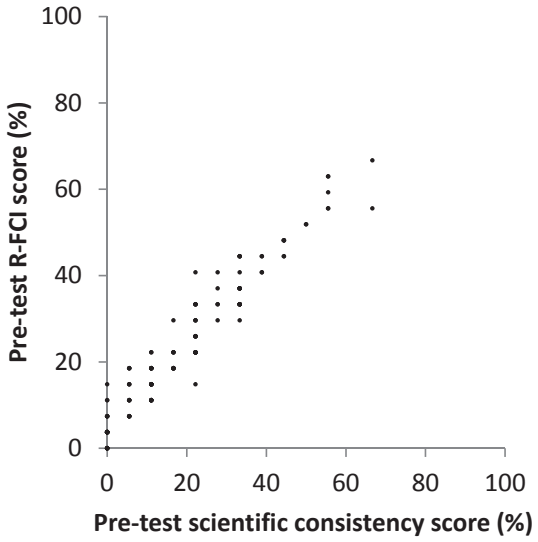


FIGURE 2 Scatter for plot students' (n = 168) scientific consistency and R-FCI scores in pre-test.

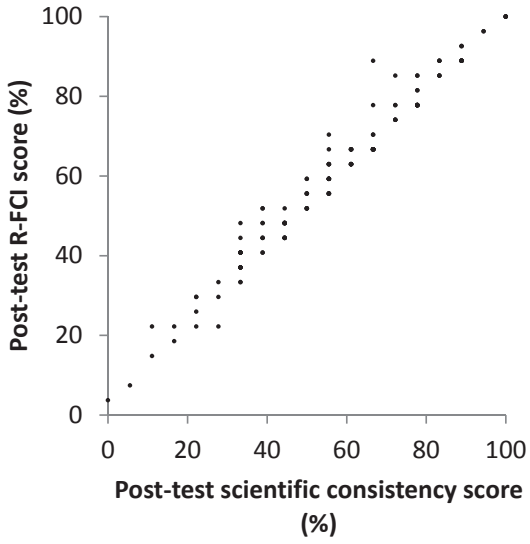


FIGURE 3 Scatter plot for students' (n = 168) scientific consistency and R-FCI scores in post-test.

Instead, the following correct and important statement was expressed in Discussion and Conclusion: "…the ability to interpret multiple representations is a necessary but not sufficient condition for the correct scientific understanding of physics concepts." This was supported by evidence from data. As Table 9 shows, students exhibited clearly more representational consistency that scientific consistency. In addition, the mean of representational consistency score was 67% in the pre-test and 81% in the post-test (can be referred from bottom line in Table X in the Article I), which were clearly higher than corresponding figures of R-FCI score (24% and 61%, respectively). Thus, representational consistency is mostly clearly higher than the R-FCI score. Students are able to interpret multiple representations even if they do not necessarily understand the physics behind representations.

The concept of representational consistency and scoring system for that (presented in Section 4.2.2) were considered very usable, and they were used in the following Sub-studies II-IV. The classification in levels of consistency (Table 9) was not used in the following studies. However, such kind of classification can be useful when performance of student groups is wanted to illustrate.

In the article we also reported that students' performance was influenced by a representational format of an item: i.e. rates of scientifically correct answers between items of a theme (representational formats) varied. This was in line with earlier findings among university students in the United States (Meltzer 2005, Kohl & Finkelstein 2005).

5.1.3 Conclusion

In Discussion and Conclusion it was mentioned as a limitation that data was collected in only one school and from students of one teacher. In addition, the question of the relationship of multiple-choice items of R-FCI to open-ended multiple representation problems that demand the construction of representations was also posed. Sub-study IV reports R-FCI data collected from two other schools. Further, the intervention material used in these schools comprised the aforementioned open-ended problems.

Sub-study I was valuable for two main reasons. The existing multiple-choice tests were limited in that they did not provide a systematic evaluation of students' representational consistency. The study presented a valid and reliable test for this lack. We think the other valuable finding was that students were able to exhibit representational consistency without correct scientific understanding. We were not aware of studies in physics education research that were systematically concerned with the use of multiple representations from viewpoint of non-scientific consistency. Meltzer also (2005) examined consistency of students' errors, but he did not follow individual students' answer patterns similarly as we did in our study.

5.2 Sub-study II: Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning

5.2.1 Aims

Previous physics education research has suggested that there are different variables like general intelligence, reasoning ability and study habits that may influence students' success in learning physics concepts (Meltzer 2002). For example, a positive correlation between students' scientific reasoning ability and normalised learning gain on FCI has been reported (Coletta & Phillips 2005). In this Sub-study the first aim was to investigate the possible correlation between students' pre-instructional representational consistency on R-FCI and their single student normalised learning gain on FCI. The second aim was to examine the relationship between FCI and Lawson test to confirm earlier findings in the U.S. (Coletta & Phillips 2005, Coletta, Phillips & Steinert 2007). The pre- and post-test data (n = 131) were collected from five student groups in their first upper secondary school physics course taught by one teacher. Students were taken their first obligatory physics course.

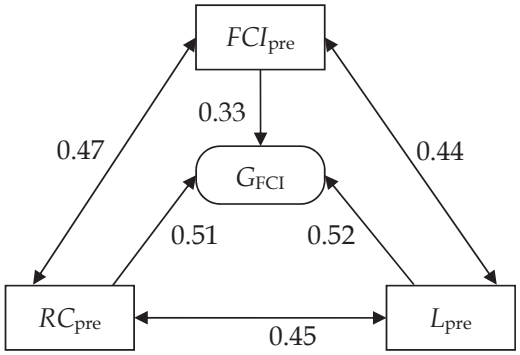


FIGURE 4 Spearman's rank correlation between single student normalised FCI gain (G_{FCI}) and the three pre-test variables for all the students (n = 131): representational consistency on the R-FCI (RC_{pre}), the FCI score (FCI_{pre}), and the Lawson test score (L_{pre}). All correlations are statistically significant ($p < 0.001$).

5.2.2 Results

Figure 4 summarises the findings concerning different correlations studied. (It must be noted that the figure illustrates only bivariate correlations, not results of structural equation modelling, for example.) We found that the Spearman's rank correlation between representational consistency on R-FCI pre-test and G_{FCI} (ρ = .51) was almost the same as the correlation between the Lawson pre-test score and G_{FCI} (ρ = .52). The correlation between FCI pre-score and G_{FCI}

was only .33. One interesting detail was that there was no correlation (ρ = -.026) between pre-instructional representational consistency and normalised gain on representational consistency, which indicates that students learned to interpret multiple representations regardless of their pre-instructional representational consistency.

5.2.3 Conclusion

As described in Section 2.2.2, the importance of the representational competence has recognised widely, but we were not aware of studies indicating straight quantitative evidence between this competence and learning gain. The major significance of the study was the finding of the correlation between representational consistency and learning gain of forces. Our results suggest that representational consistency may be one hidden variable behind learning of forces. As the data was collected only in one upper secondary school, replications in other schools and levels would be valuable.

5.3 Sub-study III: Gender differences in learning of the force concept, representational consistency, and scientific reasoning

5.3.1 Aims

As discussed in Section 2.4 females are underrepresented in physics-related education and workforce. Furthermore, gender differences favouring males in physics achievement have been reported, such as in FCI and FMCE evaluate understanding of the force concept. In Sub-study III we wanted to focus on gender differences in students' (n = 131, 45 males and 86 females) learning of the force concept, representational consistency and scientific reasoning. Learning of the force concept was evaluated also when students' pre-instructional level of representational consistency and scientific reasoning was controlled in ANCOVA. Students were studying their first physics course in upper secondary school. Interactive engagement (IE) teaching method was used. Students' learning of the force concept was documented using the single student normalised learning gain on R-FCI score ($G_{\text{R-FCI}}$). Representational consistency (RC) was evaluated using the RC score from the R-FCI test and scientific reasoning by Classroom Test of Scientific Reasoning (CTSR).

5.3.2 Results

There were statistically significant gender differences favouring males in R-FCI and RC pre- and post-test score, and CTSR pre-test score. Effect sizes for these comparisons were between .79 and 1.07, when an effect size over .8 can be considered large (Cohen 1988). $G_{\text{R-FCI}}$ was .6 for males and .45 for females. The dif-

ference was statistically significant and effect size was .71. These results were in line with previous research that has reported gender differences in conceptual understanding of forces favouring males (Osborn Popp, Meltzer & Megowan-Romanowicz 2011), even after an effective IE course (Coletta, Phillips & Steinert 2012, Pollock, Finkelstein & Kost 2007). However, these studies did not use pre-test scores as covariates to discern the treatment effect.

We conducted analysis of covariance (ANOVA) for $G_{R\text{-}FCI}$ using RC and CTSR pre-test scores as covariates and gender as a fixed factor: gender difference in $G_{R\text{-}FCI}$ was not statistically significant, $F(1, 127) = .956$, $p = .330$. Adjusted $G_{R\text{-}FCI}$ was .53 for males and .49 for females. This indicates that the gender difference in the unadjusted gains is due to pre-instructional differences, and teaching method seemed to be gender neutral. Our result is in line with an earlier study (Kost, Pollock & Finkelstein 2009) that found that gender difference in achievement was largely accounted for when students' prior knowledge in mathematics and physics and their incoming attitudes and beliefs were taken into account. However, they used a different set of pre-instructional variables than we did.

We also found that gender differences varied depending on a sub-concept of the force. Single student normalised gain on Newton's third law items (R-FCI) were excellent for both males and females, and there was no significant gender difference. We think excellent learning on Newton's third law items were due to the use of certain representation – interaction diagram. Instead, gender difference was significant, favouring males in single student normalised gain on item groups of other sub-concepts of the force (Newton's first and second laws, and gravitation). We could not conduct ANCOVA for sub-concepts of the force, because the data did not fulfil the assumptions required. It is possible that some of the gender differences in sub-concepts would be non-significant if representational consistency and scientific reasoning had been covariated.

5.3.3 Conclusion

An important goal of science education research is to find effective, gender-neutral teaching strategies to enhance students' conceptual understanding on target domain. In this context, gender neutrality refers to the goal that physics instruction should not favour males or females. Sub-study III was not able to show that gender differences diminished or eliminated due to IE method which happened in one university study in the U.S. (Lorenzo, Crouch & Mazur 2006). On the other hand, when students' pre-instructional level of representational consistency and scientific reasoning was controlled, the gender difference in learning of forces did not exist. This indicates that teaching was gender neutral.

In addition, a positive and statistically significant correlation ($r = .47$, $p < .001$) between RC and learning gain supported an assumption that students' representational competence has a relationship with their learning. This correlation justified using RC as a covariant in gender analysis, in which RC was revealed to be an effective factor. Some earlier studies have shown that emphasising multiple representations in teaching can be helpful to students in their use

of multiple representations (Kohl, Rosengrant & Finkelstein 2007, Seufert 2003). We maintain that multiple representation approaches can be helpful for the learning of both genders. Sub-study IV tries to find evidence that the multiple representation approach can enhance students' representational competence and learning, but gender differences are not discussed as the number of students, especially females, was small.

5.4 Sub-study IV: An intervention for using multiple representations of mechanics in upper secondary school courses

5.4.1 Aims

The aim was to study the effect of an intervention on students' learning of forces, when students were taught by transfer teachers. The notion of transfer teacher means here that teachers were experienced in-service teachers who did not belong to the research group that designed the intervention. In other words, we were interested in how the research-evidence informed teaching intervention affected students' learning when it was implemented by transfer teachers without any extra training. Often teachers in intervention studies act also in the role of researchers or they are part of a research team. Teachers like this have an opportunity to improve their skills and knowledge about teaching over a long time, which makes it harder to evaluate the effect of pure teaching material or certain strategies on student learning. There are fewer intervention studies conducted with transfer teachers, as defined before, when orientation for an intervention is just a brief.

The intervention of Sub-study IV emphasised the use of multiple representations and included the intervention material and short instructions for teachers. The description of the intervention, data and participants is given in Section 4.3.2.

5.4.2 Results

Classroom activities and exercises

We analysed from videos how the teacher used teaching time (Figure 5). The time used for the intervention was very limited in both schools: 7% in School 1 and 4% in School 2. The reason was that both teachers used the intervention exercises as homework: they gave an intervention exercise to students as homework and then presented correct answers at the beginning of the next lesson.

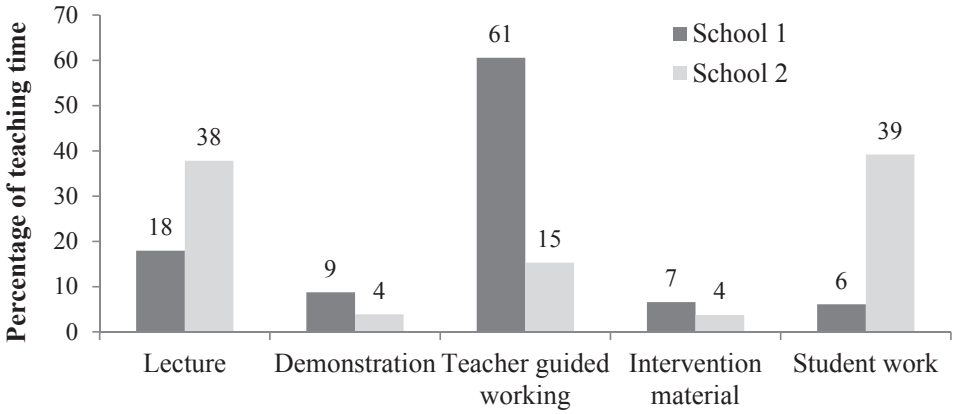


FIGURE 5 Teaching time used for different activities.

In both schools the verbal interaction between a teacher and students was quite minor, and mostly a teacher was presenting. Teacher 2 gave much more time to students to work alone (or with a peer) without affecting their working.

School 1 used a different textbook (Lehto et al. 2006) than School 2 (Hatakka et al. 2005). Each textbook exercise used for class- or homework was categorised from two viewpoints: what was the amount of representational formats of an exercise and was an exercise quantitative or qualitative. This analysis did not concentrate on which representations were demanded for solving the exercises, but only which were presented on the textbook page. In a quantitative exercise, calculations were needed in the solution when in a qualitative exercise these were not needed. A certain quantitative exercise may contain some qualitative characteristics, but a qualitative exercise does not contain any calculations.

TABLE 10 Classification of textbook exercises in School 1 (number of exercises; n = 58) and School 2 (n = 62).

School			Mathematics needed in solution Yes (quantitative)	No (qualitative)	Total
1	Number of representational formats	One	67	10	77
		Two	14	9	23
	Total		81	19	100
2	Number of representational formats	One	39	11	50
		Two	24*	26	50
	Total		63	37	100

Notes. All figures shown are percentages. *One exercise (2%) included three representational formats.

The number of performed textbook exercises was quite equal in the schools (Table 10). The majority of exercises were quantitative with one format in both schools. If an exercise included only one representational format, that was always a verbal one. However, exercises used in School 2 included more repre-

sentational formats than in School 2 and they were more often qualitative. Exercises in School 2 also included greater diversity in different formats that were verbal, graphical, pictorial, vectorial, table and motion map. Vectors and motion maps were missing in School 1.

Representational consistency and conceptual understanding of forces

We found that single student normalised learning gain on R-FCI was about double in the intervention groups compared to the baseline groups, but the difference was not statistically significant in School 1 (p = .18), School 2 (p = .17) or in the combined data of the both schools (p = .077). For the mentioned differences, medium effect sizes (Cohen's d) were found (.58, .50, and .56, respectively). It is possible that the differences were non-significant due to the small number of students. Our results cautiously indicate that students' learning of forces was higher in intervention groups.

There were no differences between the intervention and baseline students' representational consistency in R-FCI pre-test. In the post-test, representational consistency seemed to be higher among intervention groups, but the differences were not statistically significant. Instead, the change in students' representational consistency was statistically significant, with moderate effect size among intervention students (data of the schools combined; Wilcoxon signed-rank test; z = 2.80, p = .005, d = .62). This change was non-significant with small effect among baseline students (z = 1.13, p = .26, d = .27). The change in representational consistency was bigger in School 2 than School 1.

5.4.3 Conclusion

Despite the lightness of the implementation, the results suggest that the intervention increased students' understanding of forces (medium effect sizes were found in normalised gain of the R-FCI scores). However, differences between intervention and baseline groups were not statistical significant; this will be discussed in Section 6.3.2. In addition, the change in representational consistency was greater in intervention students than in baseline students, especially in School 2. The differences between schools can be influenced for many reasons, which cannot be reached by the research design of this study. One reason can lie in the different textbook exercises: in School 2 these included more representational formats than in School 1. Other reason can be students' general academic abilities, because School 2 had much better success in national matriculation results: in recent years School 2 has belonged into top 20 % of Finnish upper secondary schools whereas School 1 has been below the middle level.

The results of this intervention support the assumption that the use of multiple external representations benefits student learning. The promising results with this light intervention may well suggest that research-informed teaching material could be effective in supporting student learning even when no extra training is provided to transfer teachers.

Validity

The intervention was used only in these two courses and with a small number of students; thus, the sample is not representative. Further, albeit effect sizes for comparison of conceptual learning among intervention and baseline groups were substantial, the differences were not statistically significant. Hence, the results cannot be generalized to larger student populations. Replication studies in other courses and schools would be of value.

According to the teachers, they did not change their teaching between baseline and intervention courses, excluding the intervention material, which was credible as teachers were experienced and taught the same course for many years. Data also showed that multiple representation exercises are quite uncommon in textbooks. Hence, it is possible that the intervention exercises can have an effect on students' representational consistency and conceptual understanding. However, it should be noted that intervention exercises are not designed just for the use of multiple representations, but they are also based on knowledge about students' conceptions of forces. Both of these aspects can affect students' learning, not just multiple representations.

The intervention described in this Sub-study can be called research evidence-informed (see Section 2.5) and one can cautiously consider that it is also research evidence-based. Millar et al. (2006, p, 10) wrote that a practice can be called research evidence-informed if it is implemented and evaluated. In addition, evidence for intended outcomes should be systematically collected, analysed and available for public inspection. This research may basically satisfy these conditions, but the evidence has been published in a peer-reviewed conference proceedings. Probably the best way for public inspection would be to publish in a peer-reviewed journal, as they mostly have more intensive review process and then, better credibility than conference proceedings have.

6 GENERAL DISCUSSION

6.1 Main findings and their relation to previous research

This first contribution of this dissertation was to design the R-FCI test, as existing conceptual tests were limited in their capability to evaluate students' ability to interpret multiple representations. It must be noted that the idea for designing the R-FCI was influenced by studies of Savinainen (2004) and Savinainen and Viiri (2008). They have studied students' conceptual understanding using the notion of conceptual coherence, which is divided into three parts. One of the parts is representational coherence, which has the same meaning as representational consistency in this dissertation. Other parts are not discussed in this dissertation. Further, the representational consistency and conceptual understanding has been deliberately considered as separated measures, although representational consistency (or coherence) could be considered to be part of conceptual understanding, as Savinainen and Viiri have done (2008). Here, the intention was to focus only in multiple representations and consider the representational consistency as a background factor for conceptual understanding and learning.

Sub-study I showed that when students answer isomorphic multiple-choice items, a representational format in which a multiple-choice item is posed can have an effect on students' conceptual understanding. This was in line with the findings of Meltzer (2005) and Kohl and Finkelstein (2005). Further, in Sub-study I the notion of representational consistency was defined so that it does not demand scientific correctness, and it was illustrated how the R-FCI can be used for measuring that ability. It was found that although a student exhibits representational consistency in relation to a physical concept, it does not necessarily mean that he/she understands that concept correctly in scientific means. However, it was shown in Sub-study II that pre-instructional representational consistency correlates (ρ = .51) with learning of the force concept (i.e., normalised gain on the FCI). The correlation was quite equal as the correlation (ρ = .52) between students' pre-instructional level of scientific reasoning ability and their

learning of the force concept. This latter correlation was found in earlier studies (Coletta & Phillips 2005, Coletta, Phillips & Steinert 2007).

Above-mentioned correlations justified using students' pre-instructional representational consistency and scientific reasoning as covariates in ANCOVA (Sub-study III). Statistically significant gender difference in learning of the force concept favouring male students disappeared when ANCOVA was conducted. This indicated that the gender difference was affected by pre-instructional abilities. This was in line with a previous study of Kost and others (2009), which reported that gender difference in learning of forces was largely accounted for when students' prior knowledge in mathematics and physics and their incoming attitudes and beliefs were taken into account. However, different pre-instructional variables were examined in their study.

Although correlation between two variables does not mean that there is a causal relationship between the variables, causality can still be present. It is reasonable to believe that the relation between pre-instructional representational consistency and learning gain shown in Sub-study II constitutes supportive evidence for the assumption that use of multiple representations is beneficial for learning. Further, the findings of Sub-study IV cautiously support this assumption: Light intervention that stressed the use of multiple representations seemed to increase student learning, albeit results were not statistically significant. These findings agree with the sociocultural view, which considers that science learning demands that students learn to talk science (Mercer et al. 2004, Lemke 1990) and learn to fluently handle various modes of disciplinary discourse in which external representations play an essential role (Airey & Linder 2009).

There are many learning theories that may explain the benefits of multiple representations for learning. One is the cognitive theory of multimedia learning (Mayer 2003), which postulates that learning is better if information is presented using pictures (different static and dynamic graphics) and words (spoken or written) than just words alone. This postulate is founded on the dual coding theory (Paivio 1986), which assumes that verbal and visual information are processed via different channels in the human mind. Based on this, Mayer's theory proposes that learning can be more effective when both verbal and visual channels are used simultaneously. Further, it is stated that understanding and using various representations is important, because one format can be more suitable than another, depending on the situation (Larkin & Simon 1987) and different representations can complement each other as they express different information (Ainsworth 2006). In addition, multiple representations may help student learning, as one representation (e.g., familiar) can constrain interpretation of another representation (unfamiliar). Likewise, students' understanding may deepen because they need to integrate information from different representations (Ainsworth 2006).

It is often considered important to understand the practices of experts in particular domain, as they have powerful tools for solving problems in that domain (Bransford 2000). This also represents one viewpoint for explaining the benefits of multiple representations, as it has been reported that experts (i.e.,

scientists) are more skilled and flexible in their use of multiple representations than are novices (i.e., students) (Kozma 2003, Kohl & Finkelstein 2008). According to Kozma (2003), experts utilize multiple representations in both individual thinking and when working together. Actually, this fits well with the above-described sociocultural view concerning the importance of learning scientific language which includes multiple representations. Even though the Sub-studies in this dissertation did not consider expert–novice differences, the results can be interpreted, to some extent, from this viewpoint. Perhaps the positive correlation between representational consistency and learning indicates that more expert-like use of multiple representations relates to better conceptual understanding.

There are several studies in the field of physics education research analysing different background factors behind student learning gains (Coletta, Phillips & Steinert 2007, Meltzer 2002). Hake's (1998) normalized learning gain (see Section 4.4) allows a fair comparison of different student groups in the pre-post-test design, as it takes the pre-test score into account. Hence, it further enables comparison of different instructional methods. However, some have argued that the issue is not so straightforward, as there may be factors (hidden variables) (Meltzer 2002) which can limit learning gains in a student group. For example, if students' scientific reasoning ability is much lower in one group than in another group, this can explain a lower learning gain, even if similar instruction methods are used (Coletta, Phillips & Steinert 2007). In Sub-study II, it was stated that representational consistency may also be this kind of hidden variable. However, many very different factors correlate with learning and many of these may also correlate with each other. Thus, it is difficult to say which factor is most fundamental. For example, representational consistency might correlate with spatial abilities. However, physical representations have very specific meanings; therefore spatial ability does not explain very much about representational consistency. Nevertheless, more research is needed about relations between representational consistency and other factors (discussed in Section 6.4).

6.2 Implications for research and education

This dissertation produced the R-FCI test, which can be used for purposes of research and education. Dozens of researchers and teachers around the world have requested the R-FCI test, which is available now in Finnish, English, Japanese and Indonesian. In addition, the intervention material used in Sub-study IV is available on the Internet (Nieminen, Savinainen & Viiri 2012a). The exercises can be used, a minimum, in the first and fourth physics courses in upper secondary physics.

As discussed in Section 6.1, the findings in this dissertation indicate that the successful use of multiple representations is advantageous to student learning. Hence, multiple representations should be included in physics lessons so that students learn to interpret and construct different representations and

move between them. Most probably, representationally rich learning environments and a teachers' explicit use of multiple representations would also improve student metarepresentational competences (e.g., their ability to select appropriate representation for a particular problem). Use of multiple representations should be also included in summative course exams, as summative assessment has a substantial influence on what is considered important to learn (Harlen 2007).

Large sets of exercises from two Finnish physics textbooks were analysed from the multiple representations perspective in Sub-study IV, and according to this analysis textbooks could include more exercises for multiple representations than they do now. One encouraging trend for the use of multiple representations in education is the increase of information and communication technologies (ICT). These provide a versatile platform (cf. chalkboard) to present and produce multiple representations in various learning environments and situations. For example, students could construct representations using tablet computers and a teacher could then collect these and utilize them, for instance, in a formative assessment process. There are many resources available on the Internet, such as PhET simulations (Physics Education Technology Project 2013). The importance of multiple representations should be emphasised in teacher education and pre-service teachers should be familiarized with the use of new technology.

6.3 Validity

6.3.1 Validity of the R-FCI

The R-FCI was the main research instrument in this dissertation. Validity and reliability of the R-FCI was discussed in Sub-study I. Acceptable values for different reliability indices were found. It was stated that good validity of the FCI supported validity of the R-FCI as well. In addition, high correlation between FCI and R-FCI scores (correct answers) was found although R-FCI includes only part of the FCI items. This correlation indicates construct validity of the R-FCI with relation to the FCI, that is, R-FCI can be used as a measure for students' conceptual understanding.

Isomorphic items of the R-FCI (representational variants of the FCI items) were carefully designed, piloted and improved in the period of 2005-2007. We also consulted David Meltzer, who previously designed that kind of items for physics concepts (Meltzer 2005). The representational consistency, which was one of the key concepts in this dissertation, was measured using these isomorphic items. Here, one can ask how valid is the representational consistency on R-FCI for evaluating students' ability to interpret multiple representations?

One possibility for answering this question would be to use some other test for the same construct, which is called alternative-form reliability (Gall, Gall & Borg 2003). However, I am not aware of R-FCI-like tests for upper secondary

level, although Meltzer (2005) and Kohl and Finkelstein (2005) have used similar kinds of isomorphic items for university level. Hence, alternative-form reliability for the representational consistency on R-FCI is not provided in this dissertation.

Sub-study IV gives some support to the construct of representational consistency on R-FCI. In this study, students' answers to open-ended exercises were collected. Representational consistency can be straightforwardly evaluated in three sub-item pairs of these exercises in a similar manner as in the R-FCI. There is significant positive correlation between representational consistency in these three sub-item pairs and representational consistency on R-FCI pre-test (ρ = .54, p = .031, n = 16) and post-test (ρ = .55, p = .028, n = 16). This result is not published in the Article IV or elsewhere.

Other sources of supportive evidence can be seen in Sub-studies II and III: there was reported positive correlation between representational consistency on R-FCI pre-test and learning gain on FCI and R-FCI. Hence, the construct – representational consistency – seems to work as it was assumed: it is related to students' learning. However, this issue is still worth future research as discussed in Section 6.4.

6.3.2 Validity of results

The data from Sub-studies I–III were collected from students of one teacher in one upper secondary school. Methodologically, this can be seen also as a strength because it gives a certain consistency to the data analysis. On the other hand, there are several factors limiting the validity of generalization of these results. First, the homogenous sample means that these students did not adequately represent the student population in Finland. Second, the teacher in Sub-studies I–III was very experienced and has done educational research for a number of years. Third, the school has had very good academic success historically in terms of national matriculation examination results. This means that student test results could be lower in standard Finnish schools. However, it is difficult to know what correlations (Sub-study II) would be like in standard schools but it is possible that, for example, the correlation between representational consistency and learning could be even higher. This speculation is somewhat supported by the result in Sub-study II: The correlation was higher among students in the bottom half of the distribution according to their scientific reasoning ability. Likewise, the results of gender issues in Sub-study III cannot be generalized to Finnish upper secondary schools: For example, if the same ANCOVA procedure were conducted in other Finnish schools, there might be significant gender differences in some schools. Most probably, the instructional approach utilized in the Sub-studies has a strong influence on the aforementioned differences, and, there may be other effective factors such as students' general learning skills and study habits.

Generalization is also a problem with Sub-study IV, as only two schools were included. In addition, some results were not statistically significant. Alt-

hough the Sub-studies yielded interesting and useful results, their generalisation would demand replication studies in other school settings.

6.4 Future research

It would be interesting to see how representational consistency on the R-FCI relates to other measures of representational consistency or ability to construct multiple representations if these kinds of measures will be available in future. R-FCI data should also be collected among university students because it would be interesting to see if there is some kind of ceiling effect for representational consistency. This means that advanced students may be able to interpret multiple representations so well that they mostly exhibit very high representational consistency in the R-FCI which could make the test too easy.

Problems with external validity (generalizability) were discussed in Section 6.3.2, and the need for replication studies was noted. For example, Substudy IV showed promising results that very light intervention seemed to affect student learning positively. In this case, data collection in replication studies could be done quite easily (pre- and post-R-FCI and intervention exercises), excluding video recording.

As discussed in Section 6.2, the use of ICT in education will increase, which brings many possibilities for the use of multiple representations. Consequently, multiple representations with ICT will be an important and extensive research area in future. Also, R-FCI can be converted to an electronic version and administered in classrooms with tablet computers, for example. Data collection would then be very easy and as reliable as in the paper-and-pencil format.

YHTEENVETO

Väitöstyössä tutkittiin oppilaiden kykyä käyttää erilaisia fysiikan representaatioita voiman käsitteen oppimisen yhteydessä. Sosiokulttuurisen oppimiskäsityksen mukaan oppiminen tapahtuu kulttuurisessa kontekstissa sosiaalisena prosessina, jossa kielellä on keskeinen rooli (Tynjälä 1999). Puhutun ja kirjoitetun kielen lisäksi fysiikan kieli sisältää paljon erilaisia representaatioita, kuten graafeja, vektoreita ja matemaattisia lausekkeita. Oppiakseen fysiikan kielen, ja siis tullakseen kyvykkääksi ratkomaan fysiikan ongelmia, oppilaiden tulisi oppia käyttämään sujuvasti moninaisia representaatioita (multiple representations) (Van Heuvelen & Zou 2001, Mercer et al. 2004, Lemke 1990). Ongelmanratkaisutilanteessa tämä tarkoittaa sitä, että oppilas osaa tulkita ja muodostaa moninaisia representaatioita, tunnistaa niiden eroavaisuudet ja samankaltaisuudet sekä liikkua sujuvasti niiden välillä.

Fysiikan opetuksen tutkijat ovat selvittäneet moninaisten representaatioiden käyttämisen hyötyjä opetuksessa, mutta aiemmin ei ole ollut sopivaa instrumenttia, jolla oppilaiden representaatioiden tulkintakykyä (representational consistency) olisi voitu tutkia. Tätä kykyä mittaamaan suunniteltiin monivalintatesti väitöstutkimuksen alkuvaiheessa. Testillä kerättiin aineistoa lukiolaisilta (n = 322) väitöstyön kaikissa osatutkimuksissa. Aineistoa kerättiin oppilailta myös muilla kvantitatiivisilla testeillä ja avoimilla tehtävillä. Lisäksi tutkimukseen osallistuneita opettajia haastateltiin ja heidän opetustaan kuvattiin.

Tulokset osoittavat, että ennen opetusta oppilailta (n = 133) mitattu representaatioiden tulkintakyky korreloi voiman käsitteen oppimisen kanssa (ρ = .51, p < .001). Havaittiin myös, että pojat oppivat kurssin aikana voiman käsitteen tyttöjä paremmin tilastollisesti merkitsevällä tasolla (d = .71, p < .001). Sukupuolten välinen ero ei ollut enää tilastollisesti merkitsevä kovarianssianalyysissa, kun kovariaatteina käytettiin oppilaiden opetusta edeltävää representaatioiden tulkintakykyä ja tieteellistä päättelykykyä. Näin ollen havaittu sukupuolten välinen ero ei ilmeisesti ollut niinkään opetuksen (ko. kurssilla) tuottamaa vaan liittyi enemmän opiskelijoiden opetusta edeltäviin kykyihin.

Väitöstyön viimeisessä osassa toteutettiin opetusinterventio kahdessa lukiossa. Interventiossa oppilaat tekivät kinematiikkaan ja voiman käsitteeseen liittyviä kotitehtäviä, jotka vaativat moninaisten representaatioiden tulkitsemista ja muodostamista sekä representaatioiden välillä liikkumista. Vaikutti siltä, että interventioon osallistuneet oppilaat (n = 28) oppivat voiman käsitteen paremmin (d = .56) kuin ne, jotka eivät osallistuneet interventioon (n = 22): Vertailussa efektikoko oli keskikokoinen (d = .56), mutta ero ei ollut tilastollisesti merkitsevä (p = .077). Tulos ei ole näin yleistettävissä, vaan tutkimus tulisi toistaa muissa kouluissa samaa asetelmaa käyttäen.

Tutkimuksen tulokset ovat yhteensopivia sosiokulttuurisen oppimisnäkemyksen kanssa, jonka mukaan kielellä, mukaan lukien fysiikan representaatiot, on keskeinen rooli oppimisessa. Tuloksia voidaan tarkastella myös yksilökonstruktivismin näkökulmasta, kuten Mayerin (2003) multimediaoppimisen

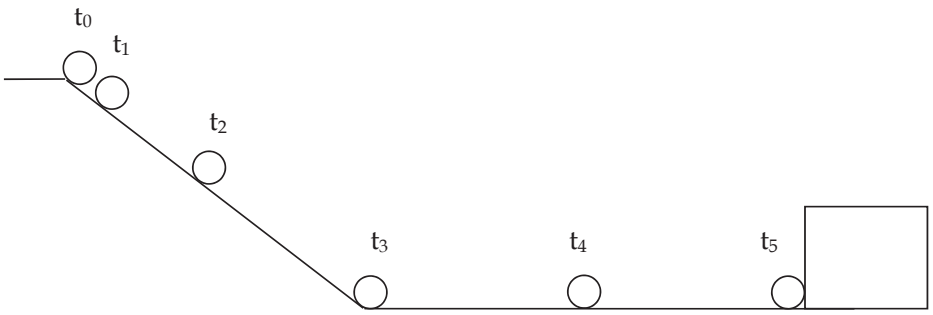
kognitiivisen teorian avulla. Tulokset osoittavat yhteyttä representaatioiden osaamisen ja voiman käsitteen oppimisen välillä. Voidaankin sanoa, että opetuksessa tulisi kiinnittää huomiota representaatioiden tulkintaan ja muodostamiseen. Tällaisia tehtäviä olisi hyvä sisällyttää myös kurssikokeisiin, koska usein arviointi ilmentää sitä, mitä pidetään tärkeänä oppia (Harlen 2007).

Uudet teknologiset innovaatiot mahdollistavat representaatioiden käyttämisen opetuksessa entistä paremmin. Esimerkiksi laboratoriotöiden rinnalla ja niiden sijasta voidaan käyttää tietokonesimulaatioita, (ks. esim. Physics Education Technology Project 2013) tai oppilaat voivat taulutietokoneiden avulla tuottaa omia representaatioita, jotka opettaja voi vaivattomasti kerätä vaikkapa formatiivisen arvioinnin tueksi. Opettajankoulutuksessa tulisi huomioida representaatioiden merkitys oppimisessa ja ohjata tulevat opettajat uuden teknologian käyttöön myös representaatioiden näkökulmasta. Väitöstyön keskeinen metodologinen tulos on representaatioiden tulkintakyvyn testi, joka on hyödyllinen sekä tutkijoille että opettajille.

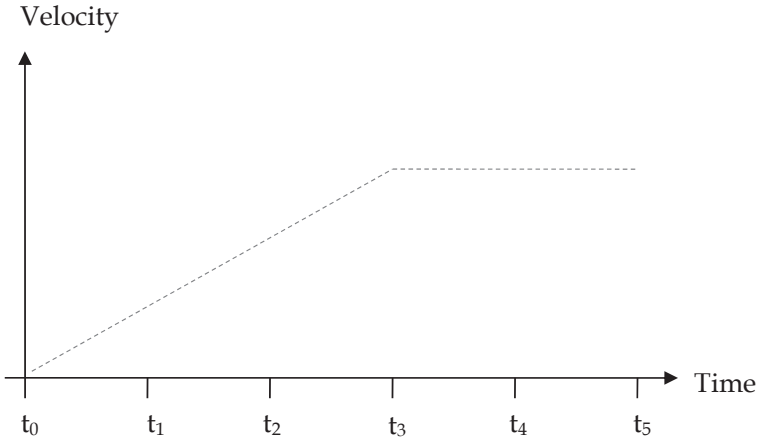
APPENDIX: AN EXAMPLE OF THE INTERVENTION EXERCISES INCLUDING MODEL ANSWERS (LIGHTENED, DASHED LINE AND ITALICS).

A ball colliding with a box

A ball starts to roll down an inclined plane at the instant of t_0. At the instant of t_5, it collides with a box. The mass of the box is ten times larger than the mass of the ball.

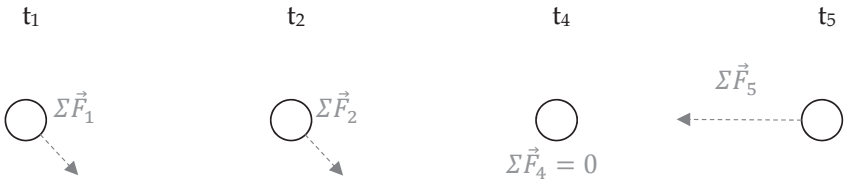


Graph velocity of the ball against time.

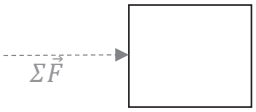


APPENDIX (CONTINUES)

Add the net force vector on the ball at t_1, t_2, t_4 and t_5 (where t_5 is the instant of collision).



Add the net force vector on the box at the instant of collision (t_5).



Compare (explain verbally) magnitudes of net force vectors on ball and box in the instant of collision (t_5).

According to Newton's third law, the interaction between the ball and the box (in the instant of collision) produces two equal, opposite forces. The one force acts on the ball and another on the box.

I think that the exercise was

1. easy.
2. quite easy.
3. not easy, but not difficult either.
4. quite difficult.
5. difficult.

REFERENCES

Acevedo Nistal, A., Van Dooren, W., Clarebout, G., Elen, J. & Verschaffel, L. 2010. Representational flexibility in linear-function problems: A choice/no-choice study. In L. Verschaffel, E. De Corte, T. de Jong & J. Elen (Eds.) Use or Representations in Reasoning and Problem Solving: Analysis and improvement.Milton Park, UK: Routledge, 74-93.

Ainsworth, S. 2006. DeFT: A conceptual framework for considering learning with multiple representations. Learning and Instruction 16 (3), 183-198.

Airey, J. & Linder, C. 2009. A disciplinary discourse perspective on university science learning: Achieving fluency in a critical constellation of modes. Journal of Research in Science Teaching 46 (1), 27-49.

American Psychological Association 2010. Publication manual of the American Psychological Association. (6th edition) Washington (D.C.): American Psychological Association.

Bao, L., Fang, K., Cai, T., Wang, J., Yang, L., Cui, L., Han, J., Ding, L. & Luo, Y. 2009. Learning of content knowledge and development of scientific reasoning ability: A cross culture comparison. American Journal of Physics 77 (12), 1118-1123.

Bransford, J. 2000. How people learn: brain, mind, experience, and school. (Expanded edition) Washington, D.C: National Academy Press.

Chi, M. T. H. 2008. Three types of conceptual change: Belief revision,mental model transformation, and categorical shift. In S. Vosniadou (Ed.) International Handbook of Research on Conceptual Change. New York: Routledge, 61-82.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. (2nd edition) Hillsdale, New Jersey: Lawrence Erlbaum.

Coletta, V. P. & Phillips, J. A. 2005. Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. American Journal of Physics 73 (12), 1172-1182.

Coletta, V. P., Phillips, J. A. & Steinert, J. J. 2007. Why You Should Measure Your Students' Reasoning Ability. The Physics Teacher 45, 235-238.

Coletta, V. P., Phillips, J. A. & Steinert, J. 2012. FCI Normalized Gain, Scientific Reasoning Ability, Thinking in Physics, and Gender Effects. In N. S. Rebello, P. Engelhardt & C. Singh (Eds.) 2011 Physics Education Research Conference. New York: American Institute of Physics, 23.

Coletta, V. P., Phillips, J. A. & Steinert, J. J. 2007. Interpreting force concept inventory scores: Normalized gain and SAT scores. Physical Review Special Topics-Physics Education Research 3 (1), 010106.

diSessa, A. A. 2004. Metarepresentation: Native Competence and Targets for Instruction. Cognition and Instruction 22 (3), 293.

diSessa, A. A. 1993. Toward an Epistemology of Physics. Cognition and Instruction 10 (2-3), 105-225.

Driver, R., Asoko, H., Leach, J., Scott, P. & Mortimer, E. 1994. Constructing Scientific Knowledge in the Classroom. Educational Researcher 23 (7), 5-12.

Gall, M. D., Gall, J. P. & Borg, W. R. 2003. Educational research: an introduction. (7th edition) Boston, MA: Allyn and Bacon.

Giancoli, D. 2005. Physics: Principles with Applications. (6th edition) Englewood, Cliffs, NJ: Prentice-Hall International.

Hake, R. R. 1998. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics 66 (1), 64-74.

Halloun, I., Hake, R. R., Mosca, E. P. & Hestenes, D. 1995. Force Concept Inventory (Revised 1995). Available in: http://modeling.asu.edu/R&E/Research.html.

Halloun, I. & Hestenes, D. 1985a. Common sense concepts about motion. American Journal of Physics 53 (11), 1056-1065.

Halloun, I. & Hestenes, D. 1985b. The initial knowledge state of college physics students. American Journal of Physics 53 (11), 1043-1055.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S. & Gernsbacher, M. A. 2007. The Science of Sex Differences in Science and Mathematics. Psychological Science in the Public Interest 8 (1), 1-51.

Harlen, W. 2007. The Quality of Learning: assessment alternatives for primary education. Primary Review Research Survey 3/4. Cambridge.

Hatakka, J., Saari, H., Sirviö, J., Viiri, J. & Yrjänäinen, S. 2005. Physica 4. Porvoo: WSOY.

Hatakka, J., Saari, H., Sirviö, J., Viiri, J. & Yrjänäinen, S. 2004. Physica 1. Porvoo: WSOY.

Hestenes, D. & Halloun, I. 1995. Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller. The Physics Teacher 33 (8), 502-506.

Hestenes, D. & Jackson, J. Revised Table II from the article (Hestenes, Wells, & Swackhamer, 1992) on FCI. Available in: http://modeling.asu.edu/R&E/research.html.

Hestenes, D., Wells, M. & Swackhamer, G. 1992. Force Concept Inventory. The Physics Teacher 30 (3), 141-158.

Kauppila, R. A. 2007. Ihmisen tapa oppia: Johdatus sosiokonstruktiiviseen oppimiskäsitykseen. Jyväskylä: PS-kustannus.

Kohl, P. B. & Finkelstein, N. D. 2008. Patterns of multiple representation use by experts and novices during physics problem solving. Physical Review Special Topics - Physics Education Research 4 (1), 010111.

Kohl, P. B. & Finkelstein, N. D. 2005. Student representational competence and self-assessment when solving physics problems. Physical Review Special Topics - Physics Education Research 1 (1), 010104.

Kohl, P. B., Rosengrant, D. & Finkelstein, N. D. 2007. Strongly and weakly directed approaches to teaching multiple representation use in physics. Physical Review Special Topics - Physics Education Research 3 (1), 010108.

Kost, L. E., Pollock, S. J. & Finkelstein, N. D. 2009. Characterizing the gender gap in introductory physics. Physical Review Special Topics - Physics Education Research 5 (1), 010101.

Kozma, R. B. 2003. The material features of multiple representations and their cognitive and social affordances for science understanding. Learning and Instruction 13 (2), 205-226.

Larkin, J. H. & Simon, H. A. 1987. Why a Diagram is (Sometimes) Worth 10000 Words. Cognitive Science 11 (1), 65-99.

Lawson, A. E. 2000. Classroom Test of Scientific Reasoning (revised version). Available in: http://www.ncsu.edu/per/TestInfo.html.

Lawson, A. E. 1978. The development and validation of a classroom test of formal reasoning. Journal of Research in Science Teaching 55 (1), 11-24.

Lawson, A. E., Banks, D. L. & Logvin, M. 2007. Self-efficacy, reasoning ability, and achievement in college biology. Journal of Research in Science Teaching 44 (5), 706-724.

Leach, J. & Scott, P. 2003. Individual and Sociocultural Views of Learning in Science Education. Science & Education 12 (1), 91-113.

Leach, J., Scott, P., Ametller, J., Hind, A. & Lewis, J. 2006. Implementing and evaluating teaching interventions: Towards research evidence-based practice? In R. Millar, J. Leach, J. Osborne & M. Ratcliffe (Eds.) Improving Subject Teaching: Lessons from research in science education. London: Routledge, 79-99.

Lehto, H., Havukainen, R., Leskinen, J. & Luoma, T. 2006. Fysiikka 4. Helsinki: Tammi.

Lemke, J. L. 1990. Talking Science: Language, Learning, and Values. New Jersey: Ablex Publishing.

Lorenzo, M., Crouch, C. H. & Mazur, E. 2006. Reducing the gender gap in the physics classroom. American Journal of Physics 74 (2), 118-122.

Mayer, R. E. 2003. The promise of multimedia learning: using the same instructional design methods across different media. Learning and Instruction 13 (2), 125-139.

McCullough, L. 2004. Gender, Context, and Physics Assessment. Journal of International Women's Studies 5 (4), 20-30.

Mehéut, M. & Psillos, D. 2004. Teaching–learning sequences: aims and tools for science education research. International Journal of Science Education 26 (5), 515-535.

Meltzer, D. E. 2005. Relation between students' problem-solving performance and representational format. American Journal of Physics 73 (5), 463-478.

Meltzer, D. E. 2002. The relationship between mathematics preparation and conceptual learning gains in physics: A possible "hidden variable" in diagnostic pretest scores. American Journal of Physics 70 (12), 1259-1268.

Mercer, N. 2000. Words and Minds : How We Use Language to Think Together. London: Routledge.

Mercer, N., Dawes, L., Wegerif, R. & Sams, C. 2004. Reasoning as a scientist: Ways of helping children to use language to learn science. British Educational Research Journal 30 (3), 359-377.

Metsämuuronen, J. 2009. Tutkimuksen tekemisen perusteet ihmistieteissä. (4th edition) Helsinki: International Methelp.

Millar, R., Leach, J., Osborne, J. & Ratcliffe, M. 2006. Research and practice in education. In R. Millar, J. Leach, J. Osborne & M. Ratcliffe (Eds.) Improving Subject Teaching: Lessons from research in science education.London: Routledge, 3-24.

Morris, S. B. S. & DeShon, R. P. R. 2002. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychological methods 7 (1), 105-125.

Nieminen, P., Savinainen, A. & Viiri, J. in press. Gender differences in learning of the force concept, representational consistency, and scientific reasoning. International Journal of Science and Mathematics Education.

Nieminen, P., Savinainen, A. & Viiri, J. 2012a. Multiple representation exercises on kinematics and the force concept. Available in: https://www.jyu.fi/edu/en/research/projects/inclass/Multiple%20representation%20exercises.pdf/view.

Nieminen, P., Savinainen, A. & Viiri, J. 2012b. Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning. Physical Review Special Topics - Physics Education Research 8 (1), 010123.

Nieminen, P., Savinainen, A. & Viiri, J. 2010. Force Concept Inventory based multiple-choice test for investigating students' representational consistency. Physical Review Special Topics - Physics Education Research 6 (2), 020109.

Osborn Popp, S. E., Meltzer, D. E. & Megowan-Romanowicz, C. 2011. Is the Force Concept Inventory Biased? Investigating Differential Item Functioning on a Test of Conceptual Learning in Physics. Available in: http://modeling.asu.edu/R&E/Research.html.

Paivio, A. 1986. Mental representations: a dual coding approach. New York: Oxford University Press.

Physics Education Technology Project 2013. PhET - Interactive Simulations. Available in: http://phet.colorado.edu/.

Pollock, S. J., Finkelstein, N. D. & Kost, L. E. 2007. Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? Physical Review Special Topics - Physics Education Research 3 (1), 010107.

Robson, C. 2002. Real world research: a resource for social scientists and practitioner-researchers. (2nd edition) Oxford: Blackwell.

Savinainen, A. 2004. High school students' conceptual coherence of qualitative knowledge in the case of the force concept. (Doctoral dissertation)

Savinainen, A., Nieminen, P., Viiri, J., Korkea-aho, J. & Talikka, A. 2007. FCI-based Multiple Choice Test for Investigating Students' Representational Coherence. In L. Hsu, C. Henderson & L. McCullough (Eds.) AIP Conference Proceedings. New York: AIP, 176.

Savinainen, A. & Scott, P. 2002. The Force Concept Inventory: a tool for monitoring student learning. Physics Education 37 (1), 45.

Savinainen, A. & Viiri, J. 2008. The Force Concept Inventory as a Measure of Students Conceptual Coherence. International Journal of Science and Mathematics Education 6 (4), 719-740.

Schnotz, W., Baadte, C., Müller, A. & Rasch, R. 2010. Creative thinking and problem solving with depictive and descriptive representations. In L. Verschaffel, E. De Corte, T. de Jong & J. Elen (Eds.) Use or Representations in Reasoning and Problem Solving: Analysis and improvement. Milton Park, UK: Routledge, 11-35.

Seufert, T. 2003. Supporting coherence formation in learning from multiple representations. Learning and Instruction 13 (2), 227-237.

Thornton, R. K. & Sokoloff, D. R. 1998. Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. American Journal of Physics 66 (4), 338-352.

Tynjälä, P. 1999. Oppiminen tiedon rakentamisena: Konstruktivistisen oppimiskäsityksen perusteita. Helsinki: Kirjayhtymä.

Van Heuvelen, A. & Zou, X. L. 2001. Multiple representations of work-energy processes. American Journal of Physics 69 (2), 184-194.

Woolfolk, A. 2007. Educational Psychology. (10th edition) Boston: Allyn & Bacon.

Zhang, J. 1997. The nature of external representations in problem solving. Cognitive Science 21 (2), 179-217.

ORIGINAL PAPERS

I

FORCE CONCEPT INVENTORY-BASED MULTIPLE-CHOICE TEST FOR INVESTIGATING STUDENTS' REPRESENTATIONAL CONSISTENCY

by

Pasi Nieminen, Antti Savinainen & Jouni Viiri, 2010

Physical Review Special Topics – Physics Education Research vol 6 (2), 020109

PHYSICAL REVIEW SPECIAL TOPICS - PHYSICS EDUCATION RESEARCH **6**, 020109 (2010)

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

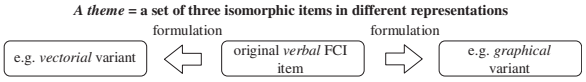


FIG. 1. A theme is a set of three isomorphic items (differing only in their representations).

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

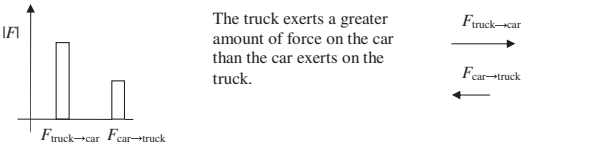


FIG. 2. Corresponding multiple-choice alternatives of theme 4 (T4) in the R-FCI. The representational formats of the alternatives are a bar chart (item 2), verbal (item 11) and vectorial (item 20). All three items include an identical original FCI question in the verbal form. The questions of bar chart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

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-

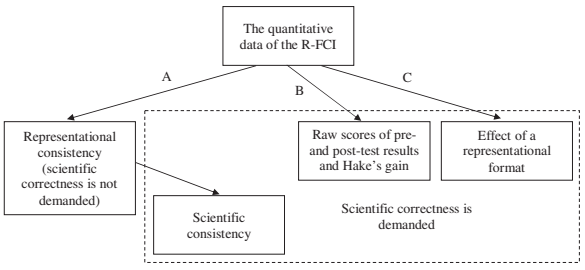


FIG. 3. Different ways for analyzing the R-FCI data.

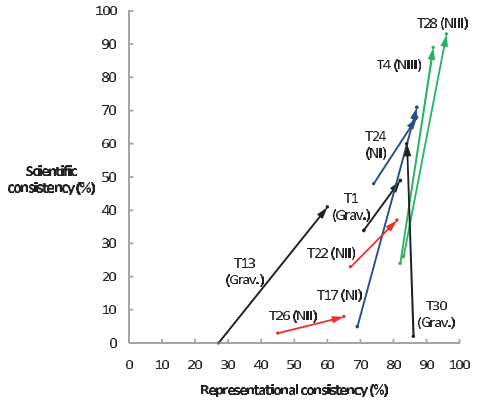


FIG. 4. (Color) Students' progress in representational and scientific consistency between pre- and post-tests. The starting point of an arrow shows pretest results and the head shows post-test results.

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'

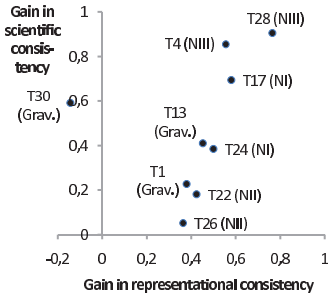


FIG. 5. (Color) Average normalized gains in representational and scientific consistency of themes.

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

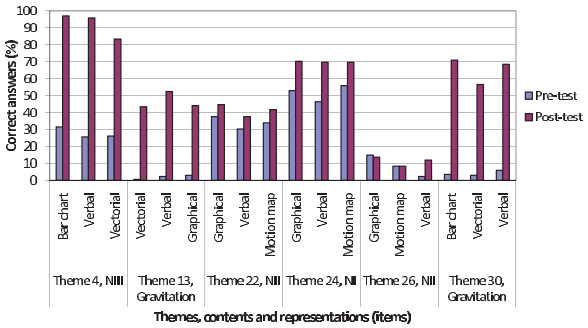


FIG. 6. (Color) The percentages of correct answers in the themes with statistically significant differences between representations.

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 $D\geq0.3$ 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 $r_{pbi}\geq0.2$ [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 $r_{test}\geq0.7$. 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

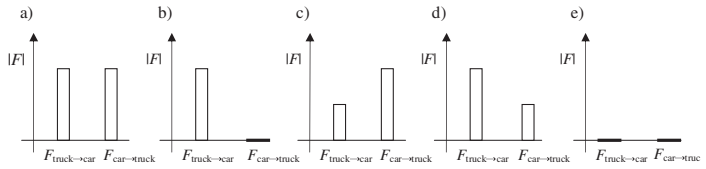


FIG. 7. Alternatives for item 2.

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.

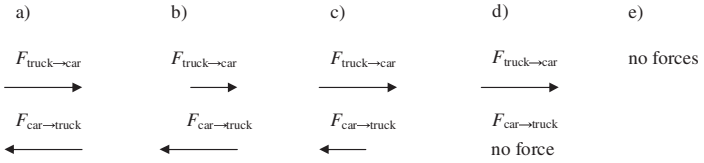


FIG. 9. Alternatives for item 20.

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).

II

RELATIONS BETWEEN REPRESENTATIONAL CONSISTENCY, CONCEPTUAL UNDERSTANDING OF THE FORCE CONCEPT, AND SCIENTIFIC REASONING

by

Pasi Nieminen, Antti Savinainen & Jouni Viiri, 2012

Physical Review Special Topics – Physics Education Research vol 8 (1), 010123

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:

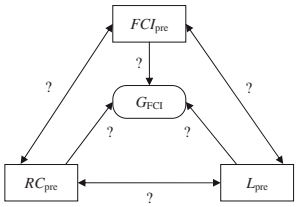


FIG. 1. The correlations between single student normalized FCI gain (G_{FCI}) and the three pretest variables: representational consistency on the R-FCI pretest (RC_{pre}), the FCI prescore (FCI_{pre}), and the Lawson prescore (L_{pre}).

(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]).

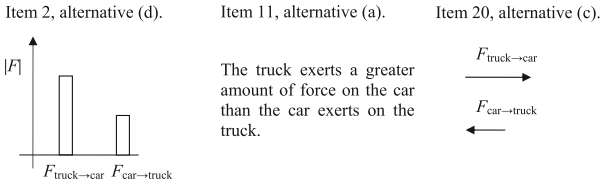


FIG. 2. Corresponding multiple-choice alternatives of a theme. The representational formats of the alternatives are a bar chart (item 2), verbal (item 11), and vectorial (item 20). All three items include an identical, original FCI question in verbal form. The questions with the bar chart and vectorial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

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	⋯

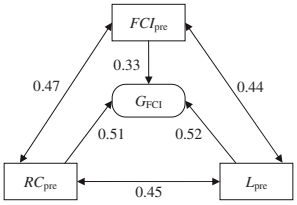


FIG. 3. Spearman's rank correlation between single student normalized FCI gain (G_{FCI}) and the three pretest variables for all the students ($n = 131$): representational consistency on the R-FCI (RC_{pre}), the FCI score (FCI_{pre}), and the Lawson test score (L_{pre}). All correlations are statistically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

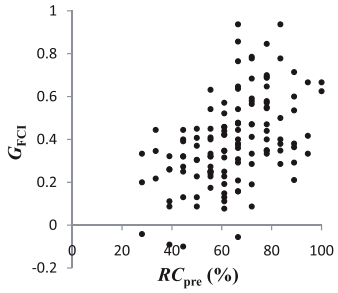


FIG. 4. Scatter plot for the students' ($n = 131$) representational consistency on the R-FCI pretest (RC_{pre}) and the single student normalized FCI gain (G_{FCI}). Spearman's rank correlation is 0.51 ($p < 0.001$).

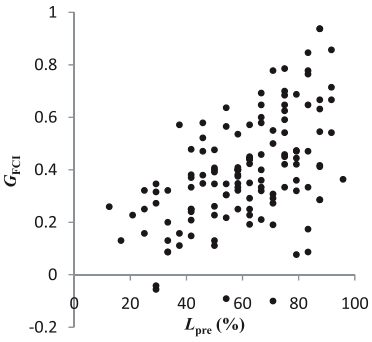


FIG. 5. Scatter plot for the students' ($n = 131$) Lawson prescore (L_{pre}) and the single student normalized FCI gain (G_{FCI}). Spearman's rank correlation is 0.52 ($p < 0.001$).

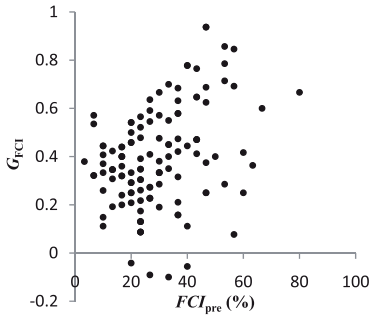


FIG. 6. Scatter plot for the students' ($n = 131$) FCI prescores (FCI_{pre}) and the single student normalized FCI gain (G_{FCI}). Spearman's rank correlation is 0.33 ($p < 0.001$).

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

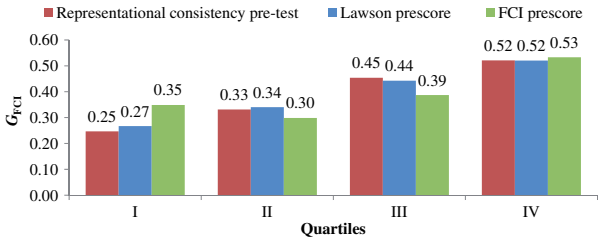


FIG. 7 (color online). Average of single student normalized FCI gain (G_{FCI}) in quartiles of representational consistency on the R-FCI pretest, Lawson prescores, and FCI prescores.

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

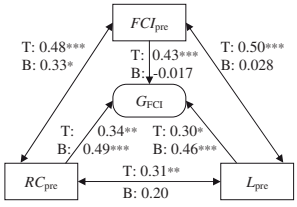


FIG. 8. Spearman's rank correlation between single student normalized FCI gain (G_{FCI}) and the three pretest variables: representational consistency on the R-FCI (RC_{pre}), the FCI score (FCI_{pre}), and the Lawson score (L_{pre}). Students were placed into the top (T, $n = 71$) or bottom (B, $n =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

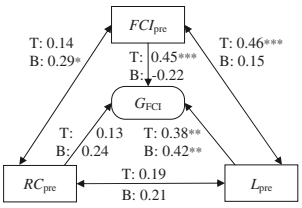


FIG. 9. Spearman's rank correlation between single student normalized FCI gain (G_{FCI}) and the three pretest variables: representational consistency on the R-FCI (RC_{pre}), the FCI score (FCI_{pre}), and the Lawson score (L_{pre}). Students were placed into the top (T, $n = 69$) or bottom (B, $n =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

III

GENDER DIFFERENCES IN LEARNING OF THE FORCE CONCEPT, REPRESENTATIONAL CONSISTENCY, SCIENTIFIC REASONING

by

Pasi Nieminen, Antti Savinainen & Jouni Viiri, in press

International Journal of Science and Mathematics Education

IV

AN INTERVENTION FOR USING MULTIPLE REPRESENTATIONS OF MECHANICS IN UPPER SECONDARY SCHOOL COURSES

by

Pasi Nieminen, Antti Savinainen, Niina Nurkka & Jouni Viiri, 2012

E-Book Proceedings of the ESERA 2011 Conference: Science learning and Citizenship

AN INTERVENTION FOR USING MULTIPLE REPRESENTATIONS OF MECHANICS IN UPPER SECONDARY SCHOOL COURSES

Pasi Nieminen[1], Antti Savinainen[1], Niina Nurkka[2], and Jouni Viiri[1]

[1]Department of Teacher Education, University of Jyväskylä, Finland and

[2]Faculty of Health Care and Social Services, Saimaa University of Applied Sciences, Finland

Previous research has emphasized the importance of multiple representations for learning and understanding physics concepts. Students should learn to construct multiple representations of physical processes and learn to move in any direction between these representations. For this purpose, we designed teaching material emphasizing multiple representations within the context of forces. The material and short instructions (= intervention) were given to two experienced upper secondary school teachers who were not involved in designing the intervention. They used the intervention material in their mechanics courses in fall 2009 (students n=28 altogether). The effect of the intervention on the learning of forces and representations was evaluated using a multiple-choice test Representational Variant of the Force Concept Inventory (R-FCI) as a pre- and post-test. In addition, students' answers to exercises of the intervention material were copied and collected. All lessons were videotaped and the teachers were interviewed after the course. The results were compared with the R-FCI baseline data (n=22 altogether) collected prior to the intervention from the same mechanics course taught by the same teachers in fall 2008. Our results suggest that the intervention material was useful for students' learning of the force concept and multiple representations.

Keywords: Multiple representations, mechanics, force, intervention

INTRODUCTION

This study concentrated on multiple representations of upper secondary school physics in the context of Newtonian mechanics. The focus was on external representations (e.g. graphs and vectors) as distinct from internal, mental representations. In the school context, external representations can be seen as communicative tools between a teacher, students and media (e.g. books), and also as cognitive tools when students are working alone, for example, in problem solving. Instead of processing everything internally in their minds, they can create an external representation and thus reduce the cognitive load (Schnotz, Baadte, Müller, & Rasch, 2010).

Multiple representations refer to the circumstances where various representations are used for learning a concept or solving a problem instead of, for example, only verbal and mathematical representations. Multiple representations have many functions in learning (Ainsworth, 1999). They can complement and constrain other representations, and construct a more complete understanding. According to Van Heuvelen and Zou (2001), multiple representations are useful in physics education as they foster students' understanding of physics problems, building a bridge between verbal and mathematical representations, and helping students develop images that give meaning to mathematical symbols. These researchers argue that students should be taught to construct multiple representations and to move in any direction between these representations. Furthermore, studies in physics education research have shown that the representational format in which a problem is posed significantly affects student performance (Meltzer, 2005; Nieminen, Savinainen, & Viiri, 2010).

A natural resource of multiple representations in studying physics would be a textbook, which indeed plays a central role in teaching and learning in Finland (Korkeakoski, 2008). Textbooks have been constantly revised in accordance with the present curriculum. Generally teachers follow the structure of textbooks and they use textbook-based exercises for class- and homework. Finnish students have become accustomed to studying with textbooks. Whilst we are aware that Finnish textbooks draw on standard representational formats used in physics, they do not necessarily offer many exercises that stress the use of multiple representations.

There is some research in science education on how evidence-informed teaching interventions can be transferred to teachers who have not been involved in designing the interventions (Leach, Scott, Ametller, Hind, & Lewis, 2006). As pointed out above, the emphasis on multiple representations in teaching and learning is important for understanding physics. However, we were unaware of any studies in which multiple representations-based teaching interventions are implemented by transfer teachers in upper secondary school. In order to study this we designed an intervention that stressed the use of multiple representations in the context of force and conducted a study concerning the use of this intervention in two upper secondary school mechanics courses taught by two transfer teachers. We hypothesized the intervention could enhance students' learning when implemented by transfer teachers.

The research questions were:

1) How did the transfer teachers integrate the intervention into their teaching without any extra training?
2) How did the intervention affect students' conceptual understanding of force and their ability to interpret multiple representations?
3) What kind of textbook exercises were used from the perspective of multiple representations?

METHOD

Design and experiment of the intervention

This paper mainly describes the intervention and baseline phases of the study, however each phase (Fig. 1) is briefly presented. The intervention material itself was designed during the design phase. The material consists of seven exercises (3–6 sub-items per exercise). These open-ended, paper-and-pencil exercises emphasize the use of multiple representations and guide movement between the representations in different contexts that address the force concept. During the pilot phase, the material was used in an upper secondary school course taught by one of the authors (AS). After the pilot phase we made some improvements to the material. Only the research group (the authors) participated in these three phases.

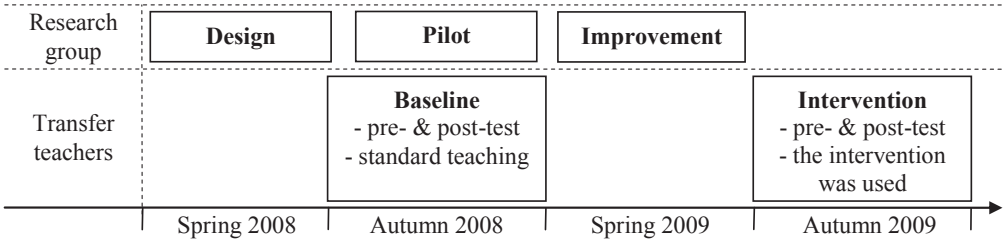


Fig. 1. Different phases of the study

The baseline phase was required to evaluate the influence of the intervention on students' learning. Between the baseline and intervention phases, the teachers remained the same, as did the courses within the same high school mechanics courses. The only difference was that

the intervention was not used in the baseline phase and, of course, the different year groups contained different students.

During the intervention phase, we gave the intervention material to two transfer teachers. We did not want to interfere too much with the teachers' plans for the course, hence we only made suggestions about when to use a certain exercise during the course in relation to the content of the course. We did not provide instructions on how to use the material.

Pre- and post-test instrument

For evaluating students' understanding of the force concept and the ability to interpret multiple representations consistently (i.e. *representational consistency*) the Representational Variant of Force Concept Inventory (R-FCI; Nieminen et al., 2010) was used. The R-FCI contains nine *themes* concerning gravitation and Newton's laws in different contexts. Each theme has three isomorphic items (the context and content remain as similar as possible) in different representations. Each item contains five multiple-choice alternatives: one scientifically correct and four incorrect distracters. Figure 2 provides an example for corresponding alternatives of the items of a theme. A more detailed description of the R-FCI is provided in our previous article (Nieminen et al., 2010).

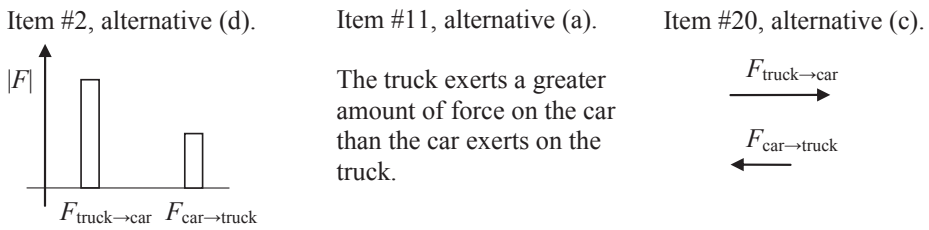


Fig. 2. Corresponding multiple-choice alternatives of the items of a theme of R-FCI

The R-FCI was used to evaluate representational consistency. When a student exhibits representational consistency, scientific correctness is not required. Figure 2 shows an example of one representationally consistent answer pattern in which each item has been answered scientifically incorrectly. Of course, if a student answers all three items of a theme correctly, he/she also exhibits representational consistency as the selected alternatives correspond between the items.

The R-FCI score (sum of scientifically correct answers) can be used for evaluating students' conceptual understanding of force. The pattern in Figure 2 exhibits no conceptual understanding: each item has been answered scientifically incorrectly and answering this way does not increase the score. Instead, answering each item of the theme correctly would add three points to the total score. It should be noted that understanding a representation is required to solve a conceptual problem with that representation. Hence the R-FCI score, as a measure for conceptual understanding, is not independent of the understanding of representations. However, there are strong correlations between R-FCI and FCI scores indicating that the R-FCI score is appropriate for evaluating conceptual understanding of force (Nieminen et al., 2010).

Participants and data collection

Both transfer teachers had over 10 years, and therefore extensive, experience in teaching upper secondary school physics. The students (aged 17) were taking their fourth physics course (mechanics, 30 lessons, 45 mins per lesson) in autumn 2008 (baseline courses, n=22 altogether) and 2009 (intervention courses, n=28 altogether). Each student took the R-FCI as

a pre- and post-test. In the results section, we have referred to the schools as School1 and School2, and the teachers as Teacher1 and Teacher2 respectively.

During the intervention phase each lesson was videotaped and the teachers were interviewed after the course. The intervention students' answers for the intervention exercises were collected before the teacher gave the correct answers. We did not, however, get all of the intervention students' answers for all of the seven exercises. The exercises were given at different points during the course and some students were not in school on certain days or did not complete the exercises for one reason or another. In the analysis of the intervention exercises we have included students (n=21) who completed 4-7 of the exercises. The baseline students did not complete the intervention exercises at all.

Video-data analysis used Atlas.ti software and PASW Statistics 18 was used for the quantitative analysis of R-FCI data and intervention exercises. To evaluate the conceptual understanding change we used normalized gain (Hake, 1998) defined as the ratio of the actual gain to the maximum possible gain:

$$\langle g \rangle = \frac{\text{Post-test \%} - \text{Pre-test \%}}{100\ \% - \text{Pre-test \%}}$$

Normalized gain is independent of students' pre-instruction score of measured variable, useful for the comparison of student groups. Normalized gain can be calculated from class averages (average normalized gain) or single-student scores (single-student normalized gain). This study used the latter for statistical comparisons of student groups and correlation analysis.

RESULTS

Implementation of the intervention

Both teachers used the intervention exercises as homework: They gave an intervention exercise to students as homework and they presented the correct answers at the beginning of the next lesson. Hence, the teaching time used for the intervention was very limited: 7% in School1 and 4% in School2.

There were differences, however, in how the teachers talked through the solutions of the exercises: Teacher2 only showed the correct answers, whereas Teacher1 elicited the correct answers from the students. On the other hand, Teacher2 criticized some exercises and our example solutions. Both teachers closely followed our suggestions as to when to use a certain exercise, albeit that the final exercise #7 was not completed in School2 at all. Most of the time, interaction between students and the teacher was minor or non-existent in both courses and teacher presentation was the dominant activity in class. Teacher2 gave much more time (39% of teaching time) to student work without his involvement than Teacher1 (only 6%).

Text book exercises in the intervention courses

School1 used a different textbook than School2 (School1: Lehto, Havukainen, Leskinen, & Luoma, 2006; School2: Hatakka, Saari, Sirviö, Viiri, & Yrjänäinen, 2005). Each textbook exercise used for class- or homework was analysed. We found that 77% of the textbook exercises in School1 were presented via verbal format (text) only. The other exercises (23%) also included one other representation: a graph, a picture or a table. In School2, 50% of the exercises were in verbal format only, and 48 % also included a graph, a picture, a table, vectors, or a motion map. One exercise (2%) included verbal, pictorial and graphical representation. Many exercises included mathematical markings, such as 7 m/s, but no exercise included equations, for example $F=ma$. This analysis did not concentrate on which

representations were demanded for solving the exercises, but only which were presented on the textbook page.

We also categorized the textbook exercises as quantitative (calculations were needed in the solution) or qualitative (no calculations). In this categorization, a certain quantitative exercise may contain some qualitative characteristics, but a qualitative exercise does not contain any calculations. Table I shows that the most of the exercises were quantitative in School1 (81%) and School2 (63%), but these were more frequently employed in School1. The majority of exercises (67%) were quantitative in verbal representation in School1. This was the most general type also in the School2, however less frequent there (39%). Qualitative exercises were performed more frequently in School2 (37%) than in School1 (19%). In particular qualitative exercises which included verbal and one other representation were more common in School2.

Table I. Classification of textbook exercises in School1 and 2. Some of the exercises were expressed via *only verbal* representation and the other exercises via verbal and one other representation (*verbal +*). *One exercise (2%) included verbal and two other representations.

School	Number of exercises	Quantitative		Qualitative	
		Only verbal (%)	Verbal + (%)	Only verbal (%)	Verbal + (%)
1	58	67	14	10	9
2	62	39	24*	11	26

Conceptual understanding of force

On average the normalized gain was higher among the intervention than baseline students in both schools (Table I), but the differences were not statistically significant. However, for all students (School1&2) the nearly significant p-value (.077) and the effect size of .56 suggested that the learning outcomes were better among intervention students than baseline students (medium $d \approx 0.5$; (Cohen, 1988).

We studied the students' answers to the intervention exercises and found various misconceptions (e.g. impetus or dominance conception) and errors in the use of representations. We also graded the correctness of students' answers and found that the students' (n=21) intervention exercise scores correlated with students' normalized learning gain (ρ=.50, p=.022). Hence, the R-FCI and intervention material seem to evaluate the same content and skills at least to some extent.

Table I. Normalized gain and standard errors (parentheses) in different student groups. For differences between gain of intervention and baseline groups Mann-Whitney U-test was conducted and Cohen's effect sizes calculated.

School	Intervention course		Baseline course		For gain differences	
	Gain	n	Gain	n	p-value	Effect size (d)
1	.39 (.08)	16	.19 (.10)	13	.18	.58
2	.47 (.08)	12	.27 (.17)	9	.39	.50
1&2	.42 (.06)	28	.22 (.09)	22	.077	.56

Representational consistency

The pre-test representational consistency did not differ between the intervention and baseline course students (Fig. 3): for the intervention students representational consistency was 76% and 75% for the baseline students (School1&2). The post-test consistency, in contrast, was higher among the intervention students (83%) than the baseline students (78%), although the difference was statistically almost significant (Mann-Whitney U-test, z=1.71, p=.088).

With regard to the baseline students, there was no significant change in representational consistency (i.e. the difference between pre- and post-test; Fig. 3). With regard to the intervention students, this change was significant (Wilcoxon Signed Rank Test, z=2.80, p=.005). The change was greater in School2 than in School1. In School2, the change of representation consistency was significant for both the intervention (z=2.68, p=.007) and baseline students (z=2.07, p=.038), but it was clearly greater for the intervention students as the effect sizes also indicate (1.25 and .33 respectively). In School1 the change was not significant among intervention or baseline students. The intervention students of School2 exhibited more representational consistency in post-test than the intervention students of School1 (z=2.37, p=.018).

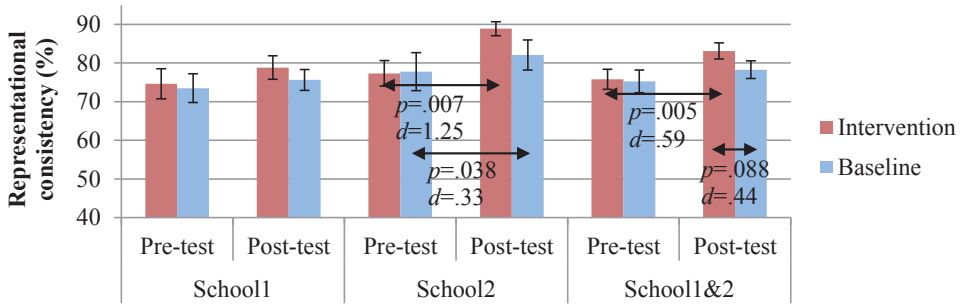


Fig. 3. Students' representational consistency in intervention and baseline groups

DISCUSSION AND CONCLUSIONS

The first research question asked how transfer teachers integrated the intervention in their teaching without any extra training. The teachers received no instructions on how to use the intervention material. Both teachers used the material as homework exercises, limiting the teaching time used for the intervention in both schools. The teachers had taught the mechanics course many times before, and they said in the interviews that they had not changed their teaching with the exception of including the new homework exercises (intervention material).

The second research question concerned the effect of intervention on students' conceptual understanding of force and their ability to interpret multiple representations. Despite the lightness of the implementation, the results suggest that the intervention increased students' understanding of force (medium effect sizes for normalized gain of the R-FCI score). In addition, the change of the representational consistency was greater for the intervention students than baseline students, especially in School2. The intervention was used only in the two courses with the small amount of students hence one needs to be careful with the generalization of results to a bigger population. A replication study in other courses and schools would be of value.

The third research question asked what kind of textbook exercises were used from the perspective of multiple representations. We assumed that if the intervention material contained something that the textbook exercises did not include, such as the use of multiple representations and qualitative thinking with regard to forces, this could explain some of the increase of intervention students' learning. Indeed, especially in School1, textbook exercises included few multiple representations and mainly focused on calculation. In School2 exercises contained more representations. Qualitative exercises, which did not demand calculation, were also used more in School2.

Why did the intervention students in School2 seem to benefit more from the intervention than their peers in School1, especially when the textbook exercises of School1 included only a few

representations? Maybe School2 students were more accustomed to studying with representations: the students had used the same textbook series in their former physics courses. Or perhaps School2 students benefited from both the intervention material and the multiple representation exercises of the textbook. It should, however, be emphasized that differences between the schools can be caused by many other reasons than just the intervention material. Nevertheless, the material seemed to improve learning in both schools.

The results from this intervention support the assumption that the use of multiple external representations benefits student learning. The ease with which the transfer teachers integrated the material also suggests that this "transfer approach" can be more widely introduced without too much difficultly. Furthermore, the positive results with this light intervention may well suggest that this type of approach could be used more extensively to support student learning.

ACKNOWLEDGEMENT

This work has been supported by the Academy of Finland (Project No. 132316).

REFERENCES

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education, 33*(2-3), 131-152.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64-74.

Hatakka, J., Saari, H., Sirviö, J., Viiri, J., & Yrjänäinen, S. (2005). *Physica 4*. Porvoo: WSOY.

Korkeakoski, E. (2008). Tavoitteista vuorovaikutukseen. Perusopetuksen pedagogiikan arvioinnin tulosten tiivistelmä ja kehittämisehdotukset. Jyväskylä: The Finnish Education Evaluation Council.

Leach, J., Scott, P., Ametller, J., Hind, A., & Lewis, J. (2006). Implementing and evaluating teaching interventions: Towards research evidence-based practice? In R. Millar, J. Leach, J. Osborne & M. Ratcliffe (Eds.), *Improving subject teaching: Lessons from research in science education* (pp. 79-99). London: Routledge.

Lehto, H., Havukainen, R., Leskinen, J., & Luoma, T. (2006). *Fysiikka 4*. Helsinki: Tammi.

Meltzer, D. E. (2005). Relation between students' problem-solving performance and representational format. *American Journal of Physics, 73*(5), 463-478.

Nieminen, P., Savinainen, A., & Viiri, J. (2010). Force concept inventory based multiple-choice test for investigating students' representational consistency. *Physical Review Special Topics - Physics Education Research, 6*(2), 020109-1-12.

Schnotz, W., Baadte, C., Müller, A., & Rasch, R. (2010). Creative thinking and problem solving with depictive and descriptive representations. In L. Verschaffel, E. De Corte, T. de Jong & J. Elen (Eds.), *Use of representations in reasoning and problem solving: Analysis and improvement* (pp. 11-35). Milton Park, UK: Routledge.

Van Heuvelen, A., & Zou, X. L. (2001). Multiple representations of work-energy processes. *American Journal of Physics, 69*(2), 184-194.