International Journal of Science Education

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tsed20

What Makes the Finnish Different in Science? Assessing and Comparing Students' Science Learning in Three Countries

Cornelia Geller[a], Knut Neumann[b], William J. Boone[c] & Hans E. Fischer[a]

[a] Faculty of Physics, University Duisburg-Essen, Essen, Germany

[b] Department of Physics Education, Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany

[c] Department of Educational Psychology, Miami University, Oxford, OH, USA

Published online: 30 Aug 2014.

To cite this article: Cornelia Geller, Knut Neumann, William J. Boone & Hans E. Fischer (2014) What Makes the Finnish Different in Science? Assessing and Comparing Students' Science Learning in Three Countries, International Journal of Science Education, 36:18, 3042-3066, DOI: 10.1080/09500693.2014.950185

To link to this article: http://dx.doi.org/10.1080/09500693.2014.950185

International Journal of Science Education, 2014
Vol. 36, No. 18, 3042–3066, http://dx.doi.org/10.1080/09500693.2014.950185

What Makes the Finnish Different in Science? Assessing and Comparing Students' Science Learning in Three Countries

Cornelia Geller[a], Knut Neumann[b*], William J. Boone[c] and Hans E. Fischer[a]

[a]*Faculty of Physics, University Duisburg-Essen, Essen, Germany;* [b]*Department of Physics Education, Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany;* [c]*Department of Educational Psychology, Miami University, Oxford, OH, USA*

This manuscript details our efforts to assess and compare students' learning about electricity in three countries. As our world is increasingly driven by technological advancements, the education of future citizens in science becomes one important resource for economic productivity. Not surprisingly international large-scale assessments are viewed as significant sources of information about the effectiveness of science education. However, these assessments do not provide information about the reasons for particular effectiveness—or more importantly a lack thereof—as these assessments are based on one-time measurements of student achievement. In order to identify reasons for the effectiveness of science education, it is necessary to investigate students' learning as a result of science instruction. In this manuscript we report about the development of an instrument to assess students' learning in the field of electricity and the use of this instrument to collect data from $N =$ 2,193 middle school students in Finland, Germany and Switzerland prior to and after instruction on the topic of electricity. Our findings indicate that the differences in students' science achievement as observed in large-scale assessments are a result of differences in students' science learning. And our findings suggest that these differences are more likely to stem from differences in science instruction than from systemic differences: a result that needs to be further explored by analyzing instruction in the three countries and its effect on students' learning.

Keywords: *Student learning; Learning progression; Electricity; Electrical energy; Large-scale assessment*

*Corresponding author. Knut Neumann, Department of Physics Education, Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany. Email: neumann@ipn.uni-kiel.de
This article was originally published with errors. This version has been corrected. Please see Corrigendum http://dx.doi.org/10.1080/09500693.2014.969495

Introduction

In a globalized world, countries must continually strive for educational excellence in order to not fall behind other countries with respect to economic and cultural well-being (deBoer, 2011). Certainly there is little debate regarding the need for a sound education in mathematics and science in order to achieve both economic and cultural goals. One prominent technique which was initiated worldwide to inform policy-makers' decisions regarding student achievement in mathematics and science is large-scale international assessments such as the Programme for International Student Assessment (PISA) (Organisation for Economic Co-operation and Development [OECD], 2001).

The PISA has revealed considerable differences in the mathematics and science achievement of students from different countries. In some Group of Eight (G8) countries, for example, students' mathematics and science achievement was only average compared to the achievement of students from other countries. Countries such as Germany viewed such average performance as a bellwether regarding the potential loss of their economic advantage (OECD, 2001). Another significant finding of PISA has been the consistently high science achievement of Finnish students. A number of hypotheses have been suggested to explain Finnish students' performance; amongst them, high quality Finnish pre-service teacher education (Lavonen & Laaksonen, 2009; Simola, 2005) leads to better instruction in mathematics and science and thus to higher student achievement.

There are limitations, however, to the evidence large-scale assessments such as PISA can provide about the reasons for differences in students' mathematics and science achievement. For example, large-scale assessments usually are one-time measurements. That is, the data obtained present a measure of students' achievement at a certain point during schooling, but the data do not document students' learning over a certain period of instruction on a particular content (i.e. differences in students' achievement between two points in time). As a consequence of this (and of course of many other possible reasons for the differences), data from large-scale assessments such as PISA can be used for cross-country comparisons (e.g. that Finnish students' science achievement is indeed higher than German students' science achievement), but provide no information about the provenance of differences (e.g. whether Finnish teachers are better at teaching than their German counterparts). That is, the rankings provided by PISA can inform policy-makers about differences in students' achievement across countries, but not about the origin of these differences. Thus, the key unanswered question still is: What are the origins of country-specific differences in students' achievement?

In order to shed more light on one possible origin of such differences, students' learning as a result of instruction needs to be investigated. In order to address this issue, we initiated a multinational project to assess and compare students' learning as a function of science instruction in Finland, Germany and Switzerland. This required an instrument suitable to assess students' learning; that is, an instrument that is sensitive to differences in students' achievement prior to and after a particular

period of science instruction. In this manuscript we report about the development of such an instrument to assess students' learning about electricity and the use of this instrument to compare students' learning as a result of instruction in the field of electricity in Finland, Germany and Switzerland. In future publications, we will, based on videos we took of the instruction students received, further investigate the origins of the differences in students' science learning beyond systemic differences in science education across countries.

Theoretical Background

The PISA was started in 2000 and is repeated every 3 years. In each testing cycle students' reading and mathematical and scientific literacy are assessed. Additionally, in each testing cycle students' literacy in one of these three domains is assessed in greater depth. In 2006 scientific literacy was assessed in greater depth. The focus upon the literacy aspect in student achievement instead of mastery of a particular curriculum is one major feature of PISA.

Students' Science Achievement

In large-scale assessments such as PISA, students' achievement is commonly reported on a specifically constructed scale. On this scale, a value of 500 scale points corresponds to the average achievement in the overall population. A difference of 100 scale points equals 1 standard deviation in students' achievement in the overall population. That is, a score of less than 500 scale points for a country suggests that student achievement in this country is lower, and a score of more than 500 scale points suggests that student achievement is better than average. A standard deviation in students' achievement for one country of 200 scale points indicates that student achievement is more heterogeneous than student achievement in the other countries, and a standard deviation of 50 scale points would suggest that student achievement is more homogeneous than in other countries. Utilizing respective scales, PISA provides a ranking of countries to inform policy-makers and the public regarding how a country ranks with respect to students' mathematics and science achievement. Usually these rankings are divided into three tiers: countries with an average student achievement significantly below average, countries with an average student achievement statistically similar to that of the international average and countries with an average student achievement significantly above the international average. One result of PISA was that some of the economically leading countries were found to exhibit less-than-expected achievements, amongst those countries were the USA, Germany and the UK (for details, see Bybee, McCrae, & Laurie, 2009; OECD, 2001).

In the 2006 iteration of PISA, in which students' scientific literacy was investigated in greater depth, Finland was the top ranked country with an average student science achievement of $M = 563$ points (SD = 86 points) and Mexico was found to be the country ranked at the very end of the list with an average student achievement of $M = 410$ points (SD = 95 points). The USA was found to rank slightly, yet

significantly below the OECD average ($M = 489$ points, SD = 106 points). Germany and Switzerland were found to rank significantly, yet slightly, above average ($M = 516$ points, SD = 100 points and $M = 512$ points, SD = 99 points, respectively) (for details, see OECD, 2007). However, neither the rankings, nor the points provide information about the nature of the differences in student achievement. The question is, if students from one country (i.e. Finland) exhibit in average a higher achievement than students from another country (i.e. Germany), in which way are they more scientifically literate? To answer the question, in addition to the distribution of student achievement being reported on a numerical scale, sometimes results are also reported on the basis of proficiency levels. For this purpose the scale is divided into segments (i.e. below 336 points, 336–409 points, 410–484 points, 485–559 points, 560–633 points, 634–708 points and above 708 points). Each segment (above 336 points) is assumed to represent a level of proficiency. Qualitative descriptions of what level of literacy can be expected from students at each level were obtained from an analysis of the content of the items with a difficulty within the range related to a respective segment. At the lowest level, for example, students were assumed to have very limited knowledge only and to be able to apply this knowledge to only very few, familiar situations. At the next levels, students were assumed to have an increasingly complex knowledge that they can apply to a broader range of situations—up to the point where students at the highest level (above 708 points) are expected to have a knowledge that reflects a deep understanding of science and the ability to use this understanding to identify scientific issues, to explain scientific phenomena, and to use scientific evidence (for details, see Bybee et al., 2009). Students are assigned a level of proficiency based on the points scored (i.e. a student with an achievement of 500 points would be assumed to have just reached the third level of science proficiency).

In large-scale assessments such as PISA, opportunities to identify the origins of the observed differences in achievement are limited. The one-time measurement, the computation of scale scores and even the utilization of proficiency levels in PISA provide an appraisal of student achievement. Regression analyses of students' achievement utilizing data on students' socioeconomic status (SES), motivation or interest can provide valuable information on how achievement depends on, or can be predicted by, other student variables (see, for example, Neumann, Fischer, & Kauertz, 2010). What is lacking, however, is evidence about country-specific differences in students' learning (meaning not only where students are at a specific time point, but also what changes occur amongst students over time). In order to obtain information about students' learning and origins of differences in students' learning, instead of student achievement, students' learning needs to be assessed and compared across countries.

Students' Science Learning

Investigating students' learning requires a model of students' proficiency in a particular field that incorporates a developmental perspective (Pellegrino, Chudowsky, & Glaser, 2001). In previous research, several attempts have been undertaken to

describe students' learning in a particular field (for an overview, see, for example, Bransford, Brown, & Cocking, 2000)—two of the most prominent organizational frameworks in the field of science education being the theory of *conceptual change* and the theory of *knowledge integration*. Both of these theories developed from Piaget's (1985) theory of cognitive development and have been successfully adapted to the teaching and learning of science over the past decades.

Conceptual change. This framework is based on the idea that learning (as described by Piaget, 1985) corresponds to the process of how scientific theories are developed (Posner, Strike, Hewson, & Gertzog, 1982). Another central idea within this framework is that students, when entering schooling, may have developed their own individual theories about scientific phenomena—commonly referred to as everyday conceptions. In this framework learning is understood as a change in how students conceptualize a scientific concept. This may include conceptual growth, i.e. students obtaining a broader understanding of a scientific concept (assimilation), as well as conceptual change, i.e. students obtaining a different understanding of a scientific concept (accommodation) (Posner et al., 1982; see also Vosniadou, 1994). That is, learning results in the student obtaining a different conception. This conception may not turn out to be a scientifically accurate conception. In fact, in the conceptual change framework students, when taught about scientific theories, are assumed to interpret the content based on their everyday conceptions potentially leading to new, yet inaccurate, interpretations of scientific phenomena (referred to as misconceptions).

Conceptual change research has identified numerous everyday conceptions and misconceptions about science (for an overview, see, for example, Duit, 2007; Vosniadou, 2008). In the domain of electricity for instance, popular misconceptions of students include the confusion of voltage and current (Engelhardt & Beichner, 2004; Maloney, O'Kuma, Hieggelke, & van Heuvelen, 2001; Shipstone et al., 1988) and various misconceptions about current, such as the idea of 'clashing currents' (Osborne, 1981). With respect to student learning, conceptual change research has focused on describing the conceptions of students as they progress in their learning about a particular concept—for example, energy (Duit, 1986), matter (Andersson, 1990) or electricity (Lee & Law, 2001). In recent years some conceptual change research has worked toward identifying a general sequence of conceptions along which students progress in mastering a particular concept or domain. Examples of science topics for which there has been interest to evaluate a sequence include energy (Lee & Liu, 2010; Liu & McKeough, 2005; Neumann, Viering, Boone, & Fischer, 2013) and force/motion (Alonzo & Steedle, 2009). Findings from this research suggest that a general progression of most students can only be expected over extended periods of time (e.g. several grades). In summary, the conceptual change framework has proven to be a valuable framework for better understanding particular difficulties students have in mastering a domain despite receiving instruction.

Knowledge integration. This framework is based upon the idea that learning corresponds to the development of increasingly complex cognition (i.e. cognitive abilities) through reflective abstraction (Songer & Linn, 1991; for details on reflective abstraction, see Piaget, 1952). Abilities of lower complexity are viewed as a prerequisite for abilities of higher complexity (Gagne & White, 1978). Over time this idea has been further refined (e.g. in Fischer's 1980 skill theory or Commons & Pekker's 2008 model of hierarchical complexity) and even utilized to develop a program to foster students' science learning through accelerating cognitive development (Adey & Shayer, 1990; Shayer & Adey, 1992a, 1992b).

The perspective of learning corresponding to the development of increasingly complex abilities has also been utilized to describe the process of knowledge acquisition (Aebli, 1980). In this view learning corresponds to (1) acquiring new knowledge elements (i.e. assimilation in the notion of Piaget, 1985) and (2) establishing new links between elements of an existing knowledge base (i.e. accommodation in the notion of Piaget, 1985). Learners' development of an increasingly complex knowledge base characterizes this perspective (Fischer & Von Aufschnaiter, 1993; Osborne & Wittrock, 1983; see also Bransford et al., 2000).

Recently, science education researchers have utilized the idea of the complexity of a knowledge base to describe different levels of proficiency as a product of learning (see Table 1). Although these three examples may differ in naming nomenclature, a commonality is the use of five or six 'levels'. Typically, levels range from a lowest complexity level ('single isolated elements') to a highest level ('multiple interrelated elements') (Bernholt & Parchmann, 2011; Kauertz & Fischer, 2006; Liu, Lee, Hofstetter, & Linn, 2008). Kauertz and Fischer (2006) have suggested six levels of complexity ('one fact', 'several facts', 'one relation', 'several unconnected relations', 'several connected relations' and 'conceptual understanding') to describe the knowledge required to solve physics tasks. Tasks that require the knowledge of facts (e.g. that watt is the unit of power) were considered to be the easiest, whereas those tasks requiring the knowledge of a system of intertwined relations between a series of facts (e.g. that the electrical energy of an 'empty' battery was converted, not consumed) were

Table 1. Overview of three models describing learning through the process of knowledge integration

Content complexity (Kauertz & Fischer, 2006)	Knowledge integration (Liu et al., 2008)	Hierarchical complexity (Bernholt & Parchmann, 2011)
One fact	Off task	Everyday experiences
Several facts	No-link: non-normative ideas	Facts
One relation	Partial link: normative ideas	Processes
Several relations	Full-link: single link between two normative ideas	Linear causality
Several interconnected relations	Complex-link: two or more links between normative ideas	Multivariate interdependencies
Overarching concept		

viewed as requiring a conceptual understanding and therefore were considered to be amongst the most difficult tasks. Kauertz and Fischer (2006) have provided evidence that the difficulty of items can indeed be described as a function of the complexity of the knowledge required to solve test items. Results showed that (1) the higher the complexity level of a task (determined a priori), the more difficult it was for students to solve the task (Kauertz & Fischer, 2006) and (2) item complexity levels describe a latent trait. Researchers have reported similar findings using models that differ slightly (Bernholt & Parchmann, 2011; Liu et al., 2008).

Research Questions

Large-scale assessments have provided valuable information about differences in students' achievement across countries. Moreover, such assessments have provided insights into predictors of students' achievement on the level of the individual student (e.g. SES). However, until now, little to nothing is known about differences in students' learning and instructional characteristics that may explain differences in students' learning.

In order to investigate students' learning across countries and to explore instructional characteristics that may explain differences in students' learning, a measurement instrument is needed that can assess students' learning as a result of instruction in a particular domain (e.g. science) or about a particular topic (e.g. electricity). That is, the instrument must be built on a theory of learning (Pellegrino et al., 2001)—more specifically, a theory of learning about the particular domain or topic in question (Duncan & Hmelo-Silver, 2009). For the design of such instruments, the 'knowledge integration approach', which describes students' learning as the growth of an increasingly complex knowledge base in a particular domain or about a particular topic, has proven itself to be a sound theory that can be used to guide instrument development (Kauertz, Fischer, Mayer, Sumfleth, & Walpuski, 2010). Such an instrument would facilitate the comparison of the complexity of students' knowledge prior to and after instruction on the particular topic as a measure for students' learning.

The primary aim of the study presented in this manuscript is to investigate students' learning in terms of the increase in the complexity of their knowledge base across countries to provide insights into the origins of differences in student achievement as observed in large-scale studies. The following research questions were formulated to guide our research:

(1) To what extent can students' learning be described in terms of an increasingly complex knowledge base?
(2) To what extent does students' learning differ across countries?

Method

In order to address these research questions, one important first step is to consider the countries to be included in the study. Obviously countries, or more specifically

researchers from countries, that would be interested in participating in the project would be those countries that performed unexpectedly low in international large-scale assessments. Such countries would include Germany, the UK or the USA. Researchers from these countries would obviously be interested in how students' learning in their countries differs from students' learning in countries ranging at the top of the rankings in international large-scale assessments. The latter countries would include Finland, Hong Kong or Canada. As one important part of the overall project was to also videorecord and analyze instruction in the participating countries to be able to explore factors affecting students' learning beyond systemic factors, what was also important was logistics. Researchers in each of the countries needed to be able to collect videos from a particular number of classrooms. With researchers in Germany and Switzerland having considerable experience with video studies in science classrooms, and Finnish researchers having demonstrated a particular interest in the origins for Finnish students' success in PISA, the project was set up as a multinational effort of researchers in Finland, Germany and Switzerland. This selection of countries also provided us with a reasonable variety with respect to differences in students' performance and education systems. Finnish students performed at the highest level of all tested countries since 2003, whereas German and Swiss students mostly exhibited an average performance (OECD, 2004, 2007, 2010). And although Germany and Switzerland share some similarities with respect to the education system (i.e. a multiple-school-tracks system), at the same time the structure and pacing of the teacher education system in Switzerland differ from that of Germany. The study was designed to collect data that would address selected limitations of current international large-scale science assessments, in particular the lack of information regarding students' learning as a function of classroom instruction.

Design

In order to investigate students' learning as a result of instruction, the study was designed as a pre-post-comparison with regular instruction on a specific topic, taking place between pre- and post-time points. This allowed not only for determining students' achievement (students' knowledge at pre- and post-time points), but also for determining students' learning (i.e. students' knowledge gain) from pre to post as a result of instruction.

As the goal of the study was to assess students' learning in terms of the complexity of their knowledge about a particular topic (which students received instruction on between pre- and post-time points), a topic had to be chosen which is part of the curricula of all three nations. One such topic was electrical energy. Electricity is a key component of most national science standards for secondary school age students. With regard to energy, Liu and Ruiz (2008) have commented that 'the teaching of the energy concept may appear again and again in the science curriculum at different grades, but the teaching of the energy concept may not take place at the same cognitive demand' (p. 557). That is, over time students are expected to

develop a more complex knowledge base around electrical energy. As a consequence the topic of electrical energy was the perfect match to our theoretical framework. Another reason for the choice of electrical energy was the broad societal relevance of the topic.

Instrument

The aim of the study reported in this paper was to investigate students' learning about electrical energy and compare students' learning across three countries. As previously discussed, one important step in the collection of meaningful data is the use of a theoretical model to conceptualize learning (see also Pellegrino et al., 2001). That is, what it means to progress from one level of knowledge to a higher level of knowledge of a particular topic. A second decision was to focus the study on one topic in order to facilitate a measurement of students' knowledge as precisely as possible. For development of the instrument, the model of complexity developed by Kauertz and Fischer (2006) was used. This model details a six-level complexity model, which can be utilized to describe students' learning in terms of quality (see the left-hand column in Table 1 for details).

Instrument development. A number of steps were taken to develop this project instrument—including (1) selection, adaption and authoring of items, (2) expert review of items, (3) revision of items and (4) compilation of test booklets utilizing the final set of test items.

(1) Item selection. A rigorous process of selection, adaption and authoring of items was carried out to develop the project instrumentation. The instrument involved items on the chosen topic of electrical energy, as well as items from the related topics of electricity and energy. This meant that items ranged from items concerning the broad topics of electricity and energy and items focused specifically on electrical energy. Electricity items and energy items were included in the test in order to be able to measure students' prior knowledge and growth over time. The development of the instrument started with selecting items from existing tests (Engelhardt & Beichner, 2004; Kauertz & Fischer, 2006). Items from these existing instruments were classified with respect to Kauertz and Fischer's (2006; cf. Table 1) complexity model. If items could not be classified utilizing the complexity model, when possible, items were adapted in order to fit the model. In order to ensure a similar number of items for the topic of electricity, for the topic of energy and for the topic of electrical energy, additional items were authored as needed. When authoring these items, Finnish and German school textbooks were utilized (the latter being similar to Swiss school textbooks) to guide item content selection. In addition to textbooks, research concerning students' energy misconceptions and electricity misconceptions were used to aid in the development of distractors for items assessing conceptual understanding (cf. Table 1). Although care was taken to present items representing all six complexity levels to students, a particular emphasis concerned the lower

complexity level items. This was done because previous research suggested that most of the students of this study would typically be at the lower level of complexity (Kauertz & Fischer, 2006; Neumann et al., 2010). So, more items of lower complexity levels were utilized for the instrument to facilitate the classification of students who had not progressed beyond the lower level of complexity. In general, about two-thirds of the test items were multiple choice in construction and one-third of the test items were open ended. This mix of item types selected for open-ended items can provide additional information regarding respondent knowledge, but can take longer for respondents to answer (in comparison to multiple-choice items) and are costly to score (cf. Briggs, Alonzo, Schwab, & Wilson, 2006). Table 2 shows a selection of sample instrument items for different complexity levels.

The test items were arranged in 'context units' (Adams, 2009), e.g. three to six items were grouped together and were introduced through the use of a common stimulus—e.g. a short text and a photo. Such 'context units' are used in PISA. PISA asserts, and we concur, that such an approach to item presentation provides a familiar item context to students. The context units utilized in our instrument included domestic appliances (e.g. washing machine), technical devices (e.g. builder's hoist) and common physics contexts (e.g. simple electric circuits). The whole set of items arranged in their context units are provided as online supplemental material. Note, how the sample items from Table 2 are integrated into their respective units.

The developed item pool was reviewed by experts from Germany, Finland and Switzerland to assess (1) curricular validity and (2) conformity to the selected model of complexity. After translation and piloting, 54 items (40 multiple-choice and 14 open-ended items), grouped in 12 context units, were selected for the main study (cf. Table 3).

(2) Expert ratings. Following the selection of the items for the study, an expert rating was conducted with regard to the complexity level of each of the 54 test items. Each item was reviewed by at least one external reviewer. The reviewers were asked to (1) assign each item to one of the six complexity levels and (2) identify which of the three item topics pertained to an item. The agreement between the assignment of items to topics and complexity levels by the authoring team and the experts was evaluated with Cohen's κ. The rating of the items with respect to the six complexity levels ($\kappa = .53$) suggested good agreement—taking into account that judging the complexity of a task is a high inferent process. For high inferent processes, Wirtz and Caspar (2002) suggested that kappa values higher than .4 may be considered sufficient. The ratings of the items with respect to the topics electricity, energy and electrical energy ($\kappa = .78$) accordingly suggested very good agreement (Wirtz & Caspar, 2002).

(3) Test booklet compilation. The 54 items were equally distributed into three test booklets (booklets A, B and C), so that each test booklet contained four context

Table 2. Sample items

Sample Item A: Electric Mixer [One Fact]
A '450 Watt label is affixed to a handheld mixer. The handheld mixer uses electric energy which is converted into kinetic energy to a large extent.
Which physical quantity is indicated by Watt? Please tick only one answer!
☐ Work
☐ Energy
☐ Power
☐ Force

Sample Item B: Wind Turbine [One relation]
A wind turbine operator assumes that a wind turbine at a certain location provides 800,000 kWh of energy annually. He further assumes that the average output of the wind turbine when it is operating is 500 kW. How many hours does the turbine have to operate? Please show your work.

Sample Item C: High Voltage [Several relations]
Why are high voltages used to transmit ectrical energy from the power station to the consumer?. Please tick only one answer!
☐ Because constant power is provided, there is a lower electric current and less heating of the wire.
☐ Because constant power is provided, there is a higher electrical current so that charges are flowing faster.
☐ The resistance in the wire causes the voltage to decrease so that an appropriate voltage is provided at the outlet.
☐ The resistance in the wire causes charges to get lost so more charges have to be provided initially.

Sample Item D: Electrical circuit [Overarching concept]
What is meant by the term 'Voltage'? Please explain in one sentence if possible!

Table 3. Distribution of items as a function of topic and complexity level

Item type	One fact	Several facts	One relation	Unconnected relations	Interconnected relations	Overarching concept	Total
Electricity: circuits, voltage, current, resistance	6	2	4	2	1	1	16
Energy in general: forms, conversion, dissipation, conservation	7	2	2	4	1	3	19
Electrical energy and electrical power	1	0	11	3	3	1	19
Total	14	4	17	9	5	5	54

units (i.e. a total of 18 items). A mix of similar item difficulty was presented in each booklet. In order to control for order effects (Carstensen, Prenzel, & Baumert, 2008), three additional test booklets (named 'a', 'b' and 'c') were constructed using the same

test items (e.g. booklet 'A' contained the identical items as booklet 'a'), but items were presented in a different order. In total six test item booklets (A, B, C, a, b and c) were developed utilizing the 54 items.

Sample

In this study entire classes of students were tested in order to be able to investigate students' learning as a result of instruction (which takes place at the classroom level).[1] The classes were sampled from grade levels 9 in Finland and Switzerland and grades 9 and 10 in Germany. This is for several reasons: for one, we aimed to facilitate the comparison with PISA results. In PISA, students who are about 15 years old are tested and in grade 9 students in the participating countries are typically 15 years old. In Germany, however, electricity is typically taught in grade 10. Thus in order to ensure comparability with respect to the typicality of the instruction, sampled classes from grade 10 were appropriate, resulting in German students being slightly older than their counterparts from Finland and Switzerland. Table 4 presents an overview of the sample and sample characteristics. For each of the three studied countries, the sample can be considered typical with respect to these characteristics. In Finland the participating schools, or classes respectively, were sampled from a 150 km radius around Jyväskylä, a medium-sized city in the heart of Finland. In Finland, compulsory schooling starts with six years of primary school (age 7–12), followed by three years of lower secondary school (age 13–15). During this phase of basic education, there is only one school track. That is, all students attend the same type of school, the comprehensive school. However, after basic education, students can chose to either attend general upper secondary schools leading to a potential university attendance, or vocational institutions leading to the attendance of a polytechnic school (for details, see Neumann, Fischer, Labudde, & Viiri, 2014). In Germany, classes were sampled from a 150 km radius around the city of Essen, NorthRhine-Westphalia. North Rhine-Westphalia is the German state with the highest population. In North Rhine-Westphalia (like in all of Germany) schooling starts with a four-year primary school (age 6–9). After attending primary school students attend one of four different types of lower secondary schools (age 10–16)—typically depending on their achievement. Low-achieving students typically attend Hauptschule. The Hauptschule represents the lowest of three career tracks, leading to a career in a practical vocation

Table 4. Number of classes (and students) as a function of country and school track

Country	Grade	Age	Comprehensive school	Lowest track	Medium track	Highest track	Overall
Finland	9	$M = 15.59$ (SD = 0.37)	25 (382)	X	X	X	25 (382)
Switzerland	9	$M = 15.58$ (SD = 0.66)	4 (73)	9 (149)	11 (195)	7 (143)	31 (560)
Germany	9/10	$M = 16.12$ (SD = 0.58)	21 (536)	0 (0)	10 (228)	16 (429)	47 (1,193)

Note: X: school track does not exist in this country.

(e.g. car mechanic). The Realschule represents a medium track, preparing students mainly for vocations requiring a more in-depth knowledge of mathematics and science (e.g. accountant). The Gymnasium represents the highest track, offering an upper secondary level and thus leading to a qualification that allows students to attend university. A fourth school type, the Gesamtschule (comprehensive school), combines all three tracks into one school type. Since the Hauptschule did not meet the curricular requirements (i.e. the contents taught in grades 9 and 10 do not correspond to the contents taught in the other school types and countries), instead of classes from Hauptschule additional classes from the lowest track in Gesamtschule were sampled (for details, see Neumann et al., 2014). In Switzerland classes were sampled from the German-speaking part of Switzerland, which is essentially the geographic area of about 150 km radius around Bern. The Swiss school system is very similar to the German one. However, schooling in Switzerland starts with a two-year preschool at the age of 4 or 5, followed by a six-year primary school (age 6–12) and three years of lower secondary school (age 12–15). As in Germany, lower secondary school consists of different school tracks. The lowest track is called Realschule, comparable to the German Hauptschule. The second track, the Sekundarschule, corresponds to the German Realschule and the highest track is also called Gymnasium. Like in Germany, the Gymnasium leads to a certification that allows for university attendance. Finally, a comprehensive school type combines all four tracks into one. Since in Switzerland there were no curricular constraints, classes from each school type were sampled (for details, see Neumann et al., 2014).

Data Collection and Data Entry

In this project data were collected before instruction of electrical energy was initiated (pre) and data were collected after teaching of the topic was completed (post). Because there was variability as to when the topic was presented to students, when electrical energy was taught in the first half of the school year, the pre- and post-data were collected at the beginning (in fall) and the end (winter) of this *first* half of the school year. In the case that electrical energy was taught in the second half of the school year, pre- and post-data were collected at the beginning and the end of this second half school year (winter, summer). Besides students' knowledge about electricity prior to and after instruction, the data collection included multiple other instruments such as a student background questionnaire based on the one used in PISA (OECD, 2005).

For the instrument developed to assess students' learning, a multi-matrix-design was used (for a general introduction to matrix designs, see Frey, Hartig, & Rupp, 2009). Using a matrix-design allowed us to administer different items to each student at pre- and post-time. Each student randomly received one of the six test booklets in the pretest and the two corresponding booklets in the posttest (see Table 5). Students were given 20 min to complete the one booklet in the pretest and 36 min to complete the two booklets for the posttest. Whereas these time

Table 5. Multi-matrix-design for the pre- and posttest

	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Pretest	A	B	C	a	b	c	A
Posttest	B C	C A	A B	b c	c a	a b	B C

frames may seem small, they had proven to be sufficient during pilot-testing. Also, a comparison of item difficulty (measured by the relative frequency of correct answers per item) as a function of item position did not suggest that items were more difficult when located at the end of a booklet.

In each country the test was administered by a member of the local project team. Once completed the tests were scanned and sent to the project coordination team based at the University of Duisburg-Essen (Germany). Data were processed using the TeleForm software. Multiple-choice items were scored automatically (1 for correct answers, 0 for incorrect answers). Open items were evaluated by native speakers and also scored dichotomously. A random sample of 5–10% of the open-ended item answers was coded by two raters. Of the 54 items, one open-ended item had to be excluded because of unreliable codings. As a result, the final item pool that was used to evaluate students' knowledge consisted of 53 items.

Results

In this study the complexity of students' knowledge about electricity prior to and after a unit on electricity was assessed in three different countries. The analysis (1) investigated to what extent it was possible to describe students' learning in terms of an increasingly complex knowledge base and (2) explored differences in students' learning across countries.

(1) In order to investigate to what extent students' learning can be described by an increasingly complex knowledge base, item difficulty was investigated as a function of complexity. If—as hypothesized—students' learning can be described by obtaining an increasingly complex knowledge base, items requiring a more complex knowledge base should be more difficult and only be solved by more able students, whereas items requiring factual knowledge should be relatively easy and solved by most students. In order to obtain measures of item difficulty in relation to students' ability, we utilized Rasch analysis. Rasch analysis is a psychometric procedure that provides, amongst many things, measures for item difficulty and measures for person ability on the same scale. This in turn allows identifying the likelihood of a person of a particular ability being able to solve a particular item (for a detailed description of Rasch analysis, see Bond & Fox, 2007; for applications in science education, see, for example, Liu, 2010; Liu & Boone, 2006).

Figure 1 presents the average difficulty in logits (i.e. on a logarithmic scale of success probability, see Bond & Fox, 2007 for details) of each item type presented

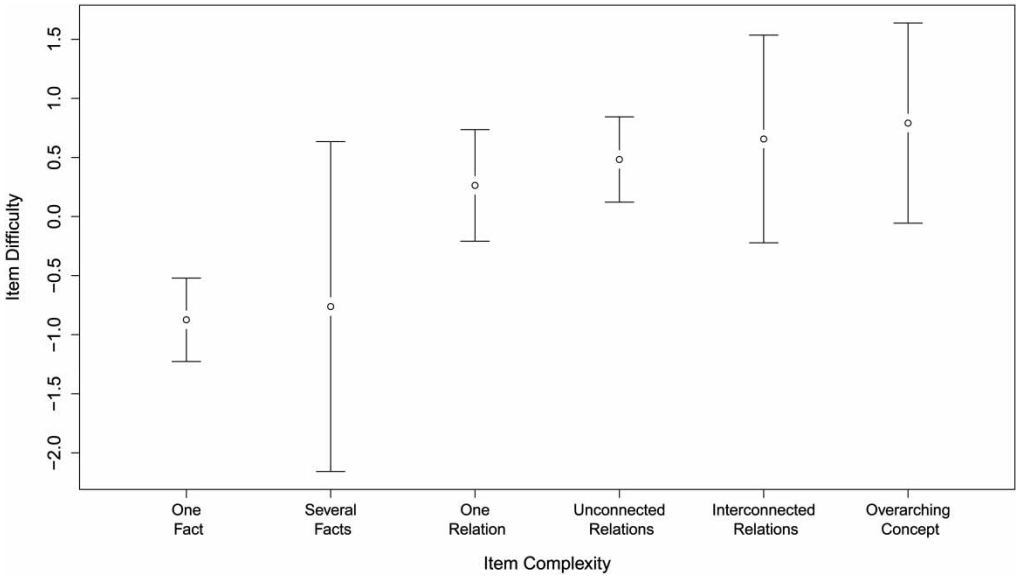

Figure 1. Mean of item difficulties with standard errors (in logits) for each complexity level

in the instrument. Generally, there is an observed increase in item difficulty as a function of the six item types. As the number of items for each complexity level is not identical (see Table 3), wider 95% confidence interval bands (in part) reflect a smaller number of items within a specific category. A trend of increasing item difficulty as a function of complexity level is present. This is the same pattern observed by Kauertz and Fischer (2006) who conducted initial investigations following their proposal of the complexity model.

An analysis of variances of the complexity of the item measures was computed to determine whether item difficulty indeed increases with higher levels of item complexity (Bernholt & Parchmann, 2011; Kauertz & Fischer, 2006). The results indicated a large significant effect of the complexity on the item difficulty, $F(5, 47) = 7.90$, $p < .001$, $\eta^2 = .46$. The differences between mean item difficulties of items from adjacent complexity levels were not significant.

(2) To compare students' learning across countries, the initial first step was to investigate item functioning by country. Rasch analysis, and specifically differential item functioning (DIF), was used to investigate the data. Based on the analysis, items were classified with respect to three DIF levels: negligible, intermediate and large (Wilson, 2005, p. 167). The results showed that 25 items exhibited negligible DIF, and 28 items exhibited intermediate or large DIF. This means that these 28 items, to some degree, differ in how each item defines the trait measured by the set of test items in each country. There are many reasons for the presence of DIF; for example, it could be that in Germany a stronger emphasis is put on students understanding the difference between voltage and current. As a consequence a respective

item may define a different portion of the (same) trait (with respect to the other set of items) in Germany than in Switzerland. In order to correct for DIF, another Rasch analysis was performed treating the 25 items with no DIF as defining in the same manner the same portion of the trait across all three countries (i.e. as item difficulty was assumed to be the same across all three countries) and the other 28 items with intermediate or large DIF were viewed as different items in each country (i.e. items were allowed to have different difficulties in each of the three countries). This compares to using different tests in each country to measure the same trait. In these tests 28 items are unique to every country and 25 items are common across the countries. Using test equating procedures (Kolen & Brennan, 2004) students' measures can be expressed on the same scale and thus be compared to each other. This way all test items, the 25 test items with no DIF and the 28 items with intermediate or large DIF, could be used for the computation of students' measures.

In order to evaluate model fit, we investigated the commonly used mean square (MNSQ) and z-standardized (ZSTD) fit statistics (for details, see Bond & Fox, 2007; Neumann, Neumann, & Nehm, 2011). Applying the typical cutoff values for the weighted MNSQ of between 0.8 and 1.2 two items had to be excluded due to a weighted MNSQ above 1.2. Another 10 items were found to have a weighted ZSTD value of above 2.0. These 10 items were also excluded from further analyses.

The results of the Rasch analysis suggested that the final pool of 41 items could be used for the computation of a person measure for each student (i.e. a measure of students' knowledge) prior to and after instruction on a single unidimensional scale. The item separation reliability was .998, suggesting that item difficulty measures were reliable enough to differentiate between different items. The person reliability was .63. Person separation reliability can be considered to be similar to Cronbach's alpha (see, for example, Wu, Adams, & Wilson, 2007). The value of .63 suggests average to good reliability, which is sufficient for group comparisons (Kline, 2000). The analyses presented below are based on the data from all students who completed both pre- and posttests ($N = 1{,}813$). Table 6 presents the mean student achievement expressed on the Rasch scale as a function of country and time point. The Rasch scale is a logit scale (see Bond & Fox, 2007 for details), which ranges from negative values to positive values. Higher values indicate a higher achievement, with zero marking the mean score across all students and points in time.

The results suggest a significant effect of the country on students' achievement in the pretest, $F(2,\ 1{,}810) = 14.23$, $p < .001$, $\eta^2 = .02$. A *post hoc* comparison with Bonferroni correction revealed that German students' achievement at pretest time was significantly higher than Finnish, $p < .001$, $d = .28$, and Swiss students' achievement, $p < .05$, $d = .18$. No significant differences were observed for Finnish and Swiss students, $p = .34$. Regarding student achievement in the posttest, again a significant effect of the country was found, $F(2,\ 1{,}810) = 16.25$, $p < .001$, $\eta^2 = 0.02$. *Post hoc* comparison of students' achievement confirmed that this time Finnish students outperformed both Swiss students, $p < .001$, $d = .31$ and German students, $p < .001$, $d = .36$. There was no statistical difference between German and Swiss students' achievement on the posttest. Figure 2 presents a graphical summary of

Table 6. Mean student achievement for the countries and time points on the Rasch logit scale

	Finland $N = 316$		Switzerland $N = 483$		Germany $N = 1{,}014$		Overall $N = 1{,}813$	
	M	SD	M	SD	M	SD	M	SD
Pretest	−0.37	0.83	−0.47	0.97	−0.21	0.89	−0.31	0.90
Posttest	0.16	0.98	−0.15	1.01	−0.19	0.95	−0.12	0.98
Difference	0.53	0.93	0.32	0.94	0.02	0.90	0.19	0.94

Note: The Rasch logit scale ranges from negative to positive values with zero marking the mean score across all students and points in time.

students' achievement at pre- and post-time. Since Rasch analysis places students and items on the same (logit) scale, students' achievement can be interpreted in terms of item features (i.e. item complexity). As a consequence, in addition to students' achievement, Figure 2 indicates the quality of students' knowledge (i.e. factual knowledge or knowledge of relations). However, for interpretation of the data what is important is the magnitude, direction and statistical significance of the change in student achievement (i.e. students' learning).

A comparison of students' achievement at pre- and posttest times across all countries suggests a gain over time (i.e. student learning), $t(1{,}812) = -8.48$, $p < .001$, $d = .20$. However, the gain is different across countries, $F(2, 1{,}810) = 43.43$,

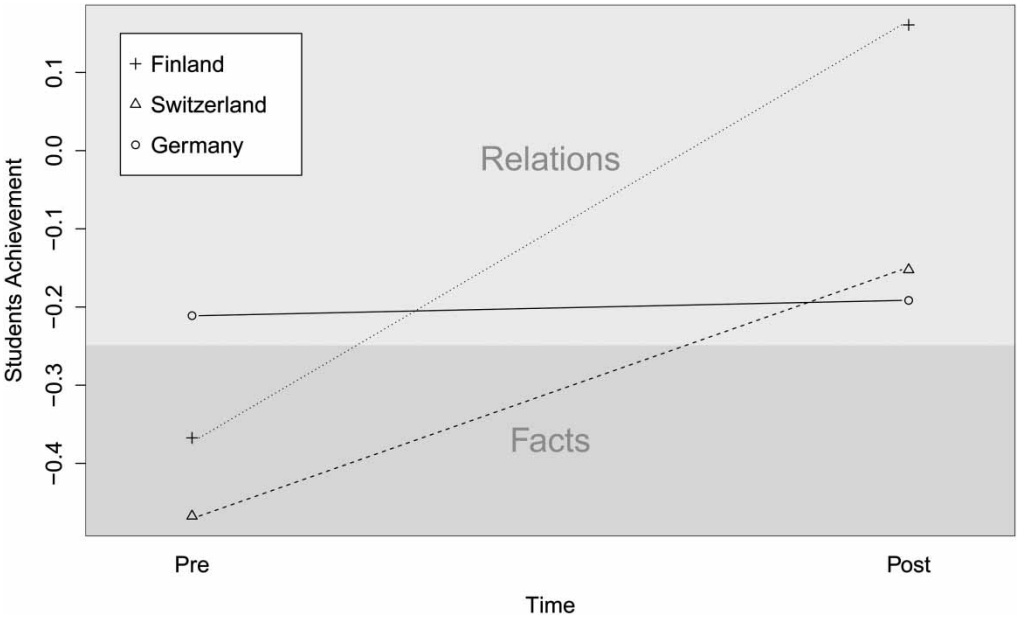

Figure 2. Students' achievement and learning (in logits) as a function of country

$p < .01$, $\eta^2 = 0.05$. The analysis of students' gain for each country revealed a positive gain in Finland $t(315) = 10.09$; $p < .001$; $d = 0.61$ and Switzerland $t(482) = 7.36$, $p < .001$, $d = 0.32$, but no significant gain in Germany $t(1{,}013) = 0.69$, $p = .49$. As Figure 2 indicates, Finnish and Swiss students progress from the knowledge of isolated facts toward the knowledge of relations. And, Finnish students clearly acquire more relations between knowledge elements than German and Swiss students.

In order to explore differences in students' learning, we investigated the effect of school tracks and classrooms (i.e. class membership) on students' learning. One reason for doing so was that the aforementioned results suggest that whereas students from Finland exhibit remarkable learning gains, and students from Switzerland show less remarkable yet still positive gains, students from Germany show no gains. And while this may suggest that German students do not learn anything, it might as well be that, for example, in Germany the gains of students from the higher tracks are compensated by losses of students from lower tracks. However, no significant effect of the school track on students' learning gains was found in Germany, $F(2, 1{,}014) = 2.04$, $p = .13$, and only a small effect was found for Switzerland, $F(3, 479) = 4.73$, $p < .01$, $\eta^2 = .03$. Thus, in a next step, we investigated the effect of the classroom on students' learning gains in order to explore if the missing learning gains are a principal feature of German science classrooms, or if at least in some classrooms students show noticeable learning gains. And indeed, our results suggested medium to large effects of the classroom on students' learning gains for Germany $F(46, 967) = 2.16$, $p < .001$; $\eta^2 = .09$ and Switzerland, $F(30, 452) = 3.47$, $p < .001$, $\eta^2 = .19$, but no significant effect for Finland, $F(24, 291) = 1.25$, $p = .20$. Obviously, in Finland students' learning does not depend on the specific class—as opposed to Germany and Switzerland where learning seems to take place only in specific classrooms. This is particularly remarkable as the samples from the three countries are (mostly) comparable with respect to parameters such as the SES, $F(2, 1{,}795) = .57$, $p = .57$.[2]

Discussion

Large-scale assessments such as PISA can provide valuable data in order to compare students' mathematics and science achievement across countries. One particular problem of these studies, however, is the lack of longitudinal data. Also, in order to be able to identify causes for differences in students' achievement, students' learning needs to be investigated. In this study we sought to move forward by developing a test instrument to assess students' knowledge about electrical energy which allowed us to investigate students' achievement at two time points (before and after instruction) and thus students' learning as a function of country. The research test instrument is based on a model of complexity that differentiates between six levels of complexity of students' electrical energy knowledge. In our analysis we have (1) investigated to what extent it was possible to describe students' learning in terms of an increasingly complex knowledge base and (2) explored differences in students' learning across countries.

Student Learning in Terms of an Increasingly Complex Knowledge Base

The instrument was administered to $N_S = 2{,}193$ students in $N_C = 103$ classes from schools in Finland, Germany and Switzerland. Findings revealed that it was in fact possible to describe students' learning in terms of complexity. About 46% of item difficulty could be explained by the complexity of the respective items. This result aligns with previous findings by Kauertz and Fischer (2006) and Bernholt and Parchmann (2011) who found that complexity explained about 29% and 57% of the variance in item difficulty, respectively. And like Kauertz and Fischer (2006) and Bernholt and Parchmann (2011), we detected a significant increase in item difficulty as a function of item complexity. These findings suggest that indeed low-achieving students typically are at the lowest complexity (factual knowledge only), while high achieving students exhibit an achievement which characterizes being at a more complex knowledge level (well above factual knowledge level) with regard to the topic of electrical energy. These findings provide sound evidence supporting that the instrument is suitable to assess students' knowledge about electrical energy. Of additional importance is that this study shows that the hierarchy of complexity levels utilized as the basis for developing different items of the instrument indeed describes different levels in mastering the topic of electricity. These findings confirm that it is possible to assess students' learning about electricity in terms of the complexity of their knowledge base. This takes the idea of proficiency levels in PISA one step further. In PISA the proficiency levels describe qualitatively different levels of mastering a domain (i.e. science). The hierarchy of complexity levels utilized in our study can describe students' progression in mastering a domain. This is important when investigating the effect of instruction on students' learning (Pellegrino et al., 2001).

Students' Learning in Germany, Finland and Switzerland Compared

Considerable differences in students' learning across countries were observed, suggesting that country differences found in large-scale assessments such as PISA may in fact be a product of differences in students' learning. No significant learning was observed for German students, and a significant but small increase in Swiss students' achievement was detected. A medium (nearly large) effect size characterized Finnish students' learning (see Figure 2 for details). These findings for German and Swiss students align well with previous findings on students' learning as obtained from cross-grade comparisons. Beaton et al. (1996), for example, compared the performance of 15-year-old US students in biology in grades 7 and 8 and found a small-to-medium effect size ($d = .33$). An analysis of several meta-analyses by Hattie (2009, p. 20) concludes that the average effect size for students' learning across one school year can be expected to be between $d = .15$ and $d = .40$ (independent from the subject)—suggesting that our findings are well in line with the suggestion that German students appear to be typically at the very low end of growth while Finnish students excel. With respect to German students' learning, findings of this study were not completely as expected. While German students were predicted to show

Table 7. Effect sizes of differences in student achievement between Finland, Germany and Switzerland in our study and PISA 2009 compared[a]

Cohen's *d*	Finland—Germany	Finland—Switzerland	Germany—Switzerland
Posttest our study	0.36	0.33	n.s.
PISA 2009	0.35	0.39	n.s.

[a]Effect sizes for differences between countries were computed using the formula for Cohen's *d* (see Bortz & Döring, 2006, p. 606 for details) based on the means and standard deviations published in Klieme et al. (2010, p. 184). In case of the standard deviation, the larger value was used for computation.

less progress than their Swiss and Finnish counterparts, the lack of any learning from pre- to posttest was not expected. One reason may be that the time span in our study was only half of a school year in comparison to a whole school year in the other studies. As a consequence of the differences in students' learning, Finnish students were found to outperform both Swiss and German students. The differences between the countries in our study is identical to that which is observed in the PISA data in which Finnish students are found to be exhibiting the highest achievement, whereas German and Swiss students typically exhibit lower and medium levels of proficiency (OECD, 2004, 2007, 2010). In fact, our analyses showed differences between the three countries at the time of the posttest that are very similar to the differences observed in PISA (see Table 7).

Also, based on the recurring findings from large-scale assessments such as PISA, at the time of the posttest (1) German students were predicted to have knowledge on the levels of 'isolated facts' and (2) Finnish students would be expected to exhibit knowledge on the level of functional relations well above the level of isolated facts. And in fact Finnish students were found to have knowledge of the level of relations at the time of the posttest—like German and Swiss students. However, while Finnish students exhibited a well-established knowledge of relations, German and Swiss students achieved knowledge barely above the level of isolated facts, a level German students already had achieved at the time of the pretest (where Finnish and Swiss students only exhibited factual knowledge). These findings are not unexpected when the design and impact of the curricula in the three countries are considered. German students typically receive teaching on electricity in lower grades of middle school (i.e. grades 6 and/or 8). Hence German students at the time of the pretest already knew some facts and some relations. Swiss and Finnish students only receive instruction in the field of electricity at the end of middle school. However, while both Swiss and Finnish students showed considerable learning and Finnish students obtained a considerably higher integrated knowledge as a result of instruction, German students remained at the somewhat intermediary level of knowing the facts and with some few relations between them. A similar pattern has been observed by other researchers (Baumert et al., 1997).

In order to further explore our findings, students' learning as a function of school track and classroom (i.e. class membership) was investigated. However, in contrast

to previous research suggesting that tracking has a considerable influence on opportunities to learn (Oakes, 1990), we found no considerable differences in students' learning across school tracks in Germany or Switzerland. Yet we found particular differences regarding students' learning amongst classrooms in Germany and Switzerland. In our study, a significant number of German and Swiss classes were found to exhibit no significant learning at all. However, there were also German and Swiss classes observed in which students' learning compares to the average learning of Finnish students. And whereas these findings may suggest that the performance of Finnish students is a result of features that uniformly apply to all science classrooms (e.g. a better teacher education system), the findings also indicate that instruction in German and Swiss classrooms can be good enough to foster learning gains similar to those of Finnish students. However, these findings require further investigations by means of video analysis to obtain evidence about which instructional characteristics foster learning to such extent. Such investigations will also be able to shed more light on why students in Finnish classrooms exhibit consistently high learning.

Conclusion

This study aimed to investigate if the country-specific differences in student achievement found in international large-scale studies can be explained as a result of differences in students' learning. The findings from our study confirm the findings from large-scale assessments such as PISA. Finnish students were found to have the largest learning gains over half a school year of teaching on the topic of electricity, whereas Swiss students demonstrated only a small gain and German students exhibited no learning gain.

By investigating students' learning based on a knowledge integration perspective, we were also able to confirm that German students do not develop an integrated knowledge base as a result of instruction, but do progress beyond a somewhat fragmented knowledge base including a broad knowledge about facts and some connections between these facts (see Figure 2). This matches the findings suggested by previous research (Baumert et al., 1997). We also found a particular homogeneity in students' learning across Finnish classes in contrast to a considerable heterogeneity in students' learning amongst classes from Germany and Switzerland. Since the samples in the three countries were pretty similar with respect to parameters such as SES, this may suggest that the origin for the differences in students' learning lie in a particular heterogeneity of the quality of science instruction in Germany and Switzerland. The reason for the homogeneity of Finnish teachers obviously lies in a feature of Finnish science education that uniformly applies to all classrooms. This may be science instruction of consistently high quality. However, in order to obtain evidence for this assumption and to identify what exactly the characteristics of the high quality teaching exhibited by Finnish teachers are, further research is needed. In order to further investigate the data patterns observed in this study, video recordings of the Finnish, German and Swiss classrooms were collected

between pre- and post-time points. These video recordings will be evaluated to further understand the mechanisms at work in classrooms of the three studied nations.

Notes

1. This is different from the purposeful sampling of students that is used, for example, in PISA. In those two efforts, a very small number of students are tested in each class.
2. In line with PISA and PISA-related studies, we used the Highest International Social and Economic Index, which reflects the highest occupational status of both the parents for each student.

Supplemental data

Supplemental data for this article can be accessed at http://dx.doi.org/10.1080/09500693.2014.950185.

References

Adams, R. J. (2009). *PISA 2006 technical report*. Paris: OECD. Retrieved from http://www.worldcat.org/oclc/302313707

Adey, P. S., & Shayer, M. (1990). Accelerating the development of formal thinking in middle and high school pupils. *Journal of Research in Science Teaching*, *27*(3), 267–285.

Aebli, H. (1980). *Denken: Das Ordnen des Tuns* [Thinking: The ordering of the doing] (Vol. 1). Stuttgart: Klett-Cotta.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389–421. doi:10.1002/sce.20303

Andersson, B. R. (1990). Pupils' conceptions of matter and its transformations (age 12–16). In P. L. Lijnse, P. Licht, & W. d. W. A. J. Vos (Eds.), *Relating macroscopic phenomena to microscopic particles: A central problem in secondary science education* (pp. 12–35). Utrecht: CD-ß Press.

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., & Köller, O. (Eds.). (1997). *TIMSS. Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde* [TIMSS. Mathematics and science instruction in international comparison. Descriptive findings]. Opladen: Leske+Budrich.

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's third international mathematics and science study (TIMSS)*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Bernholt, S., & Parchmann, I. (2011). Assessing the complexity of students' knowledge in chemistry. *Chemistry Education Research and Practice*, *12*(2), 167–173. doi:10.1039/c1rp90021h

deBoer, G. E. (2011). The globalization of science education. *Journal of Research in Science Teaching*, *48*(6), 567–591. doi:10.1002/tea.20421

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers. Retrieved from http://www.worldcat.org/oclc/141188100

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler* [Research Methods and Evaluation for Human and Social Scientists] (4th ed.). Heidelberg: Springer.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience and school*. Washington, DC: National Academy Press.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*(1), 33–63.

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching*, *46*(8), 865–883. doi:10.1002/tea.20333

Carstensen, C. H., Prenzel, M., & Baumert, J. (2008). Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt? [Trend analyses in PISA: How have [student] competencies in German developed from PISA 2000 to PISA 2006]. *Zeitschrift für Erziehungswissenschaft*, (Special Issue 10), 11–34.

Commons, M. L., & Pekker, A. (2008). Presenting the formal theory of hierarchical complexity. *World Futures*, *64*(5–7), 375–382. doi:10.1080/02604020802301204

Duit, R. (1986). *Der Energiebegriff im Physikunterricht* [The energy concept in Physics instruction]. Kiel: Institut für die Pädagogik der Naturwissenschaften.

Duit, R. (2007). *STCSE—Bibliography: Students' and teachers' conceptions and science education*. IPN. Retrieved from http://www.ipn.uni-kiel.de/aktuell/stcse/stcse.html

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching*, *46*(6), 606–609. doi:10.1002/tea.20316

Engelhardt, P., & Beichner, R. J. (2004). Students' understanding of direct current resistive electrical circuits. *American Journal of Physics*, *72*(1), 98–115.

Fischer, H. E., & Von Aufschnaiter, S. (1993). The development of meaning during physics instruction: Case studies in view of the paradigm of constructivism. *Science Education*, *77*(2), 153–168.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, *87*(6), 477–531.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53. doi:10.1111/j.1745–3992.2009.00154.x

Gagne, R. M., & White, R. T. (1978). Memory structures and learning outcomes. *Review of Educational Research*, *48*(2), 187–222.

Hattie, J. (2009). *Visible learning*. London: Routledge.

Kauertz, A., & Fischer, H. E. (2006). Assessing students' level of knowledge and analysing the reasons for learning difficulties in physics by Rasch analysis. In X. Liu & W. J. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 212–246). Maple Grove, MN: JAM Press.

Kauertz, A., Fischer, H. E., Mayer, J., Sumfleth, E., & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I [Modeling competence according to standards for science education in secondary schools]. *Zeitschrift fuer Didaktik der Naturwissenschaften*, *16*, 132–153. Retrieved from http://ipn.uni-kiel.de/zfdn/pdf/16_Kauertz.pdf

Klieme, E., Artelt, C. Hartig, J., Jude, N., Köller, O., Prenzel, M., … Stanat, P. (Eds.). (2010). *PISA 2009*: *Bilanz nach einem Jahrzehnt* [PISA 2009: Results after one decade]. Münster: Waxmann.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed). London: Routledge. Retrieved from http://www.worldcat.org/oclc/40489260

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Lavonen, J., & Laaksonen, S. (2009). Context of teaching and learning school science in Finland: Reflections on PISA 2006 results. *Journal of Research in Science Teaching*, *46*(8), 922–944.

Lee, H.-S., & Liu, O. L. (2010). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education*, *94*(4), 665–688. doi:10.1002/sce.20382

Lee, Y., & Law, N. (2001). Exploration in promoting conceptual change in electrical concepts via ontological category shift. *International Journal of Science Education*, *23*(2), 111–149.

Liu, O. L., Lee, H.-S., Hofstetter, C., & Linn, M. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, *13*(1), 33–55. doi:10.1080/10627190801968224

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch modeling approach. Science and engineering education sources*. Charlotte, NC: Information Age Pub.

Liu, X., & Boone, W. J. (Eds.). (2006). *Applications of Rasch measurement in science education*. Maple Grove, MN: JAM Press.

Liu, X., & McKeough, A. (2005). Developmental growth in students' concept of energy: Analysis of selected items from the TIMSS database. *Journal of Research in Science Teaching*, *42*(5), 493–517. doi:10.1002/tea.20060

Liu, X., & Ruiz, M. E. (2008). Using data mining to predict K–12 students' performance on large-scale assessment items related to energy. *Journal of Research in Science Teaching*, *45*(5), 554–573. doi:10.1002/tea.20232

Maloney, D. P., O'Kuma, T. L., Hieggelke, C. J., & van Heuvelen, A. (2001). Surveying students' conceptual knowledge of electricity and magnetism. *Physic Education Research, American Journal of Physics*, *69*(7), 12–23.

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education*, *10*(33), 1–33. doi:10.1080/09500693.2010.511297

Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: The impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education*, *8*(3), 545–563.

Neumann, K., Fischer, H. E., Labudde, P., & Viiri, J. (2014). Design of the study. In H. E. Fischer, P. Labudde, K. Neumann, & J. Viiri (Eds.), *Quality of instruction in Phyiscs – Results from a trinational video study* (pp. 31–48). Münster: Waxmann.

Neumann, K., Viering, T., Boone, W. J., & Fischer, H. E. (2013). Towards a learning progression of energy. *Journal of Research in Science Teaching*, *50*(2), 162–188. doi:10.1002/tea.21061

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science*. [Report]/Rand: Vol. 3928. Santa Monica, CA: Rand.

Organisation for Economic Co-operation and Development. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris: Author.

Organisation for Economic Co-operation and Development. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: Author.

Organisation for Economic Co-operation and Development. (2005). *PISA 2003 technical report*. Paris: Author.

Organisation for Economic Co-operation and Development. (2007). *Programme for international student assessment (PISA) 2006: Science competencies for tomorrow's world*. Paris: Author. Retrieved from http://public.eblib.com/EBLPublic/PublicView.do?ptiID=359823

Organisation for Economic Co-operation and Development. (2010). *Education at a glance 2010: OECD indicators*. Paris: Author. Retrieved from http://site.ebrary.com/id/10421659

Osborne, R. (1981). Children's ideas about electric current. *New Zealand Science Teacher*, *29*, 12–19.

Osborne, R. J., & Wittrock, M. C. (1983). Learning science: A generative process. *Science Education*, *67*(4), 489–508. doi:10.1002/sce.3730670406

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Piaget, J. (1952). *The origins of intelligence in children*. New York: National Universities Press.

Piaget, J. (1985). *The equilibration of cognitive structures: The central problem of intellectual development*. Chicago: University of Chicago Press.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*, 211–227.

Shayer, M., & Adey, P. S. (1992a). Accelerating the development of formal thinking in middle and high school students II: Post project effects on science achievement. *Journal of Research in Science Teaching*, *29*(1), 81–92.

Shayer, M., & Adey, P. S. (1992b). Accelerating the development of formal thinking in middle and high school students III: Testing the permanency of effects. *Journal of Research in Science Teaching*, *29*(10), 1101–1115.

Shipstone, D. M., Rhoeneck, C. v., Jung, W., Kaerrquist, C., Dupin, J. J., Johsua, S., & Licht, P. (1988). A study of students' understanding of electricity in five European countries. *International Journal of Science Education*, *10*(3), 303–316.

Simola, H. (2005). The Finnish miracle of PISA: Historical and sociological remarks on teaching and teacher education. *Comparative Education*, *41*(4), 455–470.

Songer, N., & Linn, M. C. (1991). How do students' views of science influence knowledge integration? *Journal of Research in Science Teaching*, *28*(9), 761–784.

Vosniadou, S. (1994). Capturing and modelling the process of conceptual change. *Learning and Instruction*, *4*(1), 45–69.

Vosniadou, S. (Ed.). (2008). *International handbook of research on conceptual change*. New York: Routledge.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen* [Rater agreement and rater reliability. Methods for determining and optimizing the reliability of assessments by means of category systems and rating scales]. Göttingen: Hogrefe.

Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). *Acer ConQuest: Version 2.0*. Mulgrave: ACER.