

Review Article

Exploring Epistemological Trends in Students' Understanding of Science from the Perspective of Large-Scale Studies

Anders Jakobsson,¹ Eva Davidsson,¹ Karl-Göran Karlsson,² and Magnus Oskarsson²

¹ *Learning and Society, Malmö University, 205 06 Malmö, Sweden*

² *Teacher Education, Mid Sweden University, 871 88 Härnösand, Sweden*

Correspondence should be addressed to Anders Jakobsson; anders.jakobsson@mah.se

Received 13 February 2013; Accepted 12 March 2013

Academic Editors: M. Martin, G. Velan, and J. Widera

Copyright © 2013 Anders Jakobsson et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper highlights how results from large-scale studies can be used to understand students' knowledge of science. Several scholars express critique of today's PISA framework, especially with regard to the presentation of the results as national rankings, and suggest alternative and complementary methods. The present study has used PISA data to reveal hidden patterns in the results. The results show a general descending trend in items focusing on the nature of science and how new scientific knowledge is generated. On the other hand, there is an obvious upward trend regarding tasks that measure fact-based elementary or root knowledge. These trends are slightly differentiated at a national level, as the time and magnitude of the decline or increase may vary.

1. Introduction

Since the 1990s, the OECD (Organisation for Economic Co-operation and Development) has conducted large-scale PISA studies about students' knowledge of, and attitudes toward, science, mathematics, and reading. One of the common methods of describing the results from these studies involves comparing the mean values between countries in order to evaluate different educational systems, which also constitutes an important aim of the studies [1–3]. Jakobsson et al. [4] argue that the results of the surveys and tests of educational achievement play an increasingly important role in monitoring educational performance and in political discussions around the world. The test results are used as institutional efficiency indicators, quality assurance measures, and instruments through which politicians, school administrators, and teachers are held accountable [5, 6].

There is an ongoing discussion in the field of science education about the value of these tests, and several scholars express critical opinions about the validity and reliability of the measurements. For example, Sjøberg [7] and Bautier and Rayou [8] argue that the tests do not constitute a valid representation of students performance and knowledge at

a national level, and that it is hard to draw any conclusions from the results. They also call attention to the fact that national science curriculum goals in their countries have diverged from the framework of OECD and IEA, and they point out the risk of the tests being considered as “hidden curricula” of science education. Other scholars highlight problems concerning the cultural bias of the tests (e.g., [9]), and the fact that the translating procedure favors English-speaking students [10]. Bottani and Vignaud [11] also call attention to the inherent conflict between political and scientific interests. According to them, this implies a situation in which science education researchers strive to differentiate complex patterns about people's understanding of and attitudes toward science, at the same time as political society is calling for simple answers and rapid solutions. Additionally, they argue that large-scale studies mostly focus on one-dimensional rankings between countries that mainly serve politics and not science.

However, an important question is whether the data material from these studies could be used in alternative or complementary ways, rather than simply presenting mean values and league tables across participating countries. One suggestion could be to use Fensham's [12] arguments regarding context-based curriculum. This describes the results in

terms of how many students can use their knowledge to solve problems related to contemporary society. Another example is the Norwegian research project PISA+ [13], which aims to conduct complementary qualitative data collection in classrooms in order to explore and explain trends in the results.

Therefore, the purpose of the present paper is to highlight and exemplify ways in which PISA science data is currently being used and to discuss how the data could be used to conduct alternative or complementary analyses. We also suggest ways in which quantitative large-scale studies can be complemented by qualitative analyses in order to obtain an in-depth understanding of quantitative findings. Additionally, we present an example of how to interpret the national trends of the results of three Nordic countries using Roberts' [14, 15] Curriculum Emphases as an analytic tool in order to explicate epistemological trends in students' understanding of science in these countries.

2. The Framework of PISA and OECD

The Programme for International Student Assessment (PISA) was created in 1997 by the governments of OECD countries, with the aim of monitoring education outcomes in terms of student achievement, within a common internationally agreed framework. The explicit purpose and aim of the surveys has been to evaluate the strengths and weaknesses of different school systems, which "will allow national policy makers to compare their education system with those of other countries" [1, page 7]. The tests are individual paper-and-pencil tests for 15-year-old students and consist of a mixture of multiple-choice items and questions that require students to construct their own responses. Approximately 4500–10000 students from each country participate in the surveys, and the national results are described as a mean value for each country, compared to a constructed mean value of all participating OECD countries.

The number of countries that participate in the surveys has increased from 43 in 2000 to 67 in 2009, and different years (2000–2009) have focused on different domains, such as science, mathematics, or reading. Science was the main focus of the 2006 survey, and the measured variables were three scientific competencies related to OECD's definition of scientific literacy. These competencies were described as students' ability to identify scientific issues, scientifically explain phenomena, and use scientific evidence. According to the framework [16], the aim was to assess students' ability to use their knowledge in a way that corresponds to the needs and demands of modern society, rather than just assess specific or subject-related facts and procedural knowledge. Additionally, the report entitled "Assessing Scientific, Reading and Mathematical Literacy: A framework for PISA 2006" (2006) states that the students' level of scientific literacy is assessed in relation to their understanding of scientific concepts, processes, and contexts. This method of conducting the survey describes the results on a proficiency scale from 1 to 6 that ranked individual students' abilities related to their scientific literacy performance on the test.

This method also implies that it is possible to describe the percentage of students in each country who have reached a specific proficiency level.

3. Examples of PISA's Impact on Governmental and Educational Policies

A number of articles and book chapters call attention to the increased impact of studies such as PISA and TIMSS on educational policy and national education systems in Europe and beyond. For example, Grek [17] argues that these types of surveys have become an indirect but influential tool of the new political technology of governing educational systems in the global perspective. Local policy makers use PISA results to legitimize discussions by presenting their policies as being based on robust evidence. For example, in an interview study [17] of key policy-makers in three European countries, it was evident that the PISA 2000 results had a major effect on the German educational system. The mean value rankings placed Germany in 20th place in terms of reading, mathematical, and scientific literacy among 32 countries which sent shock waves to German policy-makers, teachers, and parents. The poor results dominated German media for several weeks, project leaders gave interviews, and roundtable discussions were held on television. This resulted in a number of new national school projects and tests, and many other changes to educational practice. All of these changes were conducted despite criticism of PISA's testing frame and statistical validity (e.g., [18]).

According to Grek [17], the German example only constitutes a common pattern in relation to PISA, and TIMSS surveys were the scientific community initially criticized the statistical reliability or validity, but the data findings were gradually accepted and appropriate policy responses were then put in place. Another example of this tendency was seen in Sweden where the decreasing results from the PISA and TIMSS surveys (2000–2009) led to a crucial change in teacher education organization and content. The results were used as arguments in political documents SOU 2008:109 [19] and debates to effect change. A third example is Norway, where results from PISA surveys "provided war-like headings in most national newspapers" [7], and a situation in which media uses the results to draw its own conclusions about the quality of the school system, teachers' work, and development of "ignorant" citizens.

According to Uljens [20], these tendencies only constitute examples of a worldwide development of the promotion of a neoliberal policy, controlled by OECD and the educational-assessment movement. The idea is to support an increasingly competitive mentality combined with common standards across nations, as this is expected to be beneficial for a common market. Additionally, he argues that this mentality is supported by commodification or "marketization" of knowledge, and a stronger view of education as a vehicle for international competition. Using a transnational evaluation procedure assumes a single measurement standard, which supports the development of increased homogeneity

conducted by a self-adjusting process among participating countries.

4. The Relationship between the Education Research Community and PISA

Hopmann and Brinek [21] argue that, despite the educational impact of the PISA survey results, discussions within the educational research community about the reliability of PISA methodology are rare. One exception is Goldstein's [22] analysis of the statistical methodology of the country rankings. He concludes that it is essential to recognize that the reality of comparing countries is a complex multidimensional issue that is "well beyond PISA's ineffectual and one-dimensional attempt." Goldstein also stresses that cross-sectional data makes it impossible to draw satisfactory conclusions about the effects of different educational systems. Goldstein notes that comparative studies should move toward becoming longitudinal in order to reveal trends. In addition, Fertig [23] expresses critical viewpoints about the statistical methodology of the ranking behind the PISA surveys and calls attention to the fact that study conclusions can only be drawn if strong assumptions about the school systems are made. He argues that since the education systems of the participating countries differ in more than one aspect, it is impossible to identify the driving force behind differences in specific issues. Allerup [24] notes that the scales used are not homogenous with respect to sex, ethnicity, and the variation of item difficulties, and that these shortcomings risk misinterpretation of the Rasch model that actually constitutes the starting point for the framework.

Other scholarly criticism focuses on the validity of the studies. For example, Sjøberg [7] argues that the tests do not actually constitute a valid representation of students' performance and knowledge at a national level. According to him, one problem consists in the PISA claim that the surveys test real-life skills and competencies in authentic contexts, despite the fact that they use paper-and-pencil tests. He concludes that coping with life in modern societies requires a range of competencies and skills that cannot possibly be measured by test items of that kind. Olsen [9] highlights problems concerning cultural bias and the tendency of culture-related groups in countries with similar languages to produce similar response patterns. Puchhammer [10] calls attention to problems with the translation procedure and the fact that first-generation immigrants are tested in their second language. This implies that the number of immigrants in a country will have a great impact on that country's results. Bautier and Rayou [8] raise further questions regarding what conclusions can be drawn from the PISA studies. Their reanalyses of students' answers found that the items did not necessarily measure what they were supposed to measure. The analyses revealed that the large group of "midperformers" in PISA surveys showed great variability and instability in the descriptions of their competencies and proficiency levels. According to the authors, the results indicate that the students' responses could be correct or incorrect for reasons

that are not envisioned through a priori analysis of items. Brunner et al. [5] question the validity of the PISA tests by revealing that it is possible to coach students for the test. Their study indicates that the combined effects of pretesting and coaching have significant positive effects on students performance.

5. Suggestions of Alternative or Complementary Models for Analyses

The political impact and the controversial characteristics of the results, as well as the country rankings, have encouraged researchers to carry out alternative and complementary research studies based on available PISA data. Many of these studies use the data to analyze factors behind the different results between countries. One example is Lietz's [25] metastudy, which indicates that gender differences in reading proficiency may have a crucial impact on students' performance on natural science tests; this is because reading ability seems to have a decisive impact on how individual students understand the questions. Other researchers have used the data to reinterpret results or to conduct cross-national comparisons. For example, Kjærnsli and Lie's [26] item-by-item analysis focuses on the differences and similarities between countries, while the Northern Lights on PISA 2006 Project [27] aims to compare the science curricula in Nordic countries in relation to the PISA framework. Other examples include Kjærnsli and Lie [28], which use the data to reveal international patterns about students' preferences and attitudes [29] toward professional careers as scientists.

A somewhat different suggestion is to use the raw survey data to find new or hidden patterns. For example, Bonnet [30] argues that it would be more informative to reflect on students' errors than to simply construct macroindicators as country rankings. Reflecting on errors could help to qualify the analysis by identifying differences between countries that may point to necessary improvements in specific areas. Fensham [12] suggests secondary analyses of students' responses using a contextual set of items as the unit of analysis. The study reveals that the procedure of using a means of percentage correctness to describe national results provides an image that is quite different from the traditional national ranking. The results indicate relatively small differences in the mean of percentage correctness on different sets of items between the top-ranked and significantly lower-ranked countries. However, in some cases and in some of the sets, Fensham argues that the differences could be useful for identifying topics or content to which science teaching is already contributing or where there is a need for improvement. Bonnet [30] uses similar arguments when suggesting qualifying the analysis by identifying constructive differences between countries that may point to necessary improvements in specific areas, rather than emphasizing global differences that are not particularly helpful to teachers. Mortimore [31] and Lundgren [32] suggest changing the PISA test organization toward a more nuanced interpretation of countries' strengths and weaknesses in developing

citizens' lifelong learning. Mortimer's model presupposes an extension of the methodology to include longitudinal elements, analysis of trends in the surveys, and refocusing on how schools and school systems could promote achievement and increase educational outcome equity. This method of conducting the surveys involves the teachers and includes information from them that will enrich the context of the data.

6. The Study

In the science education community, as well as in other educational research areas, it is possible to identify relatively critical standpoints related to today's large-scale studies. Some scholars approach the discussion from the perspectives of validity (e.g., [8, 10]) or reliability (e.g., [22, 23]), which question the scientific or educational value of such large-scale studies. Others (e.g., [7, 22, 31]) express a more general critique of the methodology or framework, especially the presentation of the results in the form of mean value national rankings across participating countries. According to these scholars, such a method of conducting the surveys and presenting the results seldom offers any conceivable information regarding how to improve education at the national level. However, Fensham [12] and Bonnet [30] assert the possibility of using the data to conduct alternative or complementary analyses. Their aim is primarily related to an ambition to increase knowledge about students' understanding of, or attitudes toward, specific areas or domains in science and, in doing so, to create opportunities to enhance science instruction in a national or international perspective.

The present study proposes a complementary way of analyzing the data and shedding light on some of the epistemological trends in PISA science results from three Nordic countries, exemplified by Sweden, Denmark, and Finland. These countries represent different result developments communicated in international reports from the OECD [3, 6, 33]. We intend to use the word "trend" as a way of describing significant changes or developments in the national result on specific items. This involves measuring 15-year-old students' performances on recurring science items at various occasions from 2000 to 2009. The aim is to identify and discuss what may constitute these trends and, in doing so, advance the possibility of drawing conclusions related to science instruction and education. The chosen countries only constitute examples and could easily be replaced by others, which imply that the main purpose is to explore possibilities and problems when conducting such analyses. Our expectation is that complementary analyses of the PISA data over time may increase our knowledge of epistemological trends in students' understanding of science and, in doing so, create incitement to improve science education in a Nordic as well as an international perspective. Thus, the main aims of the study are to explore the possibilities to interpret epistemological trends in national PISA data by analyzing the results of recurring science items at different measurement occasions (2000–2009).

7. Methodology and Methodological Considerations

One option for exploring the content and character of possible national trends is through analyzing the percentage of correctness [12] related to individual recurring items (link items) at different measurement times. This means interpreting whether the changes in national mean values (P values) on individual items over time may be connected to specific competences or epistemological understanding of science. This study will look at students' performances on items that have been included in three, or in some cases four, subsequent measurements in order to interpret trends. Items that recur three times constitute the basis for the analysis and clarify trends in the material. The results section consists of two parts. The first provides a detailed description of some of the released items' content and design, how these items were categorized, results from the three countries, and the OECD mean value. The second part illustrates the categorization and results of recurring and unreleased items. The content of these items is not presented because OECD does not allow researchers to publish them as the items must remain secret for future measurements.

One option for categorizing items would have been to use the existing PISA framework of scientific literacy [3], defined as students' competencies in identifying scientific issues, explaining phenomena scientifically, and using scientific evidence. However, the main aim of the study was not to propose alternative categories for interpreting students' competences, skills, or performance in PISA science, but rather to use the data to analyze whether there are any epistemological trends related to individual items and, if so, to interpret whether these trends can be related to students' understanding of science. Therefore, our analysis has used Roberts' [14, 15] descriptions of *curriculum emphasis* as an analytical tool with which the PISA items categorized. An advantage of this alternative is built on the assumption that this procedure will offer a richer description of students' epistemological understanding related to specific areas within science.

Roberts' curriculum emphases were originally used as a tool for analyzing the content and intentions of science curricula and school textbooks. One important conclusion in relation to this research is that science education and instruction consist of at least one of the emphases that accentuate different aims in the curriculum and, in doing so, bring consequences to the alignment of it. In this way, the concept of curriculum emphases also strives to capture the explicit or implicit orientations and different sets of messages given in science teaching that imply what science actually is about, its intent, and meaning. Furthermore, it is possible to understand and interpret written assignments in tests or items in surveys as representations or expressions of similar emphases, which implies that each of the items may agree with one or more of the emphases. In relation to this, we argue that it is possible to use the emphases as a tool for approaching students' epistemological understanding of science. We have found it possible and fruitful to analyze each item from the following emphases: solid foundation; correct explanations; self as explainer; everyday coping; scientific

skill development; science, technology, and decisions; and structure of science.

The categories of solid foundation and correct explanations concern issues related to the solid knowledge products and facts of science. The first category focuses on the subject matter that is necessary for studying science at the next level in the educational system, whereas the second category accentuates the ability to give correct explanations without demanding any underlying understanding. Knowledge connected to these two categories often constitutes de-contextualized facts and explanations related to an intradisciplinary-specific subject area. These two categories may be related to Hofers' [34] two epistemic dimensions, which include students' views on the *certainty of knowledge* or the *simplicity of knowledge*. The first dimension describes a continuum of students' understanding of knowledge as fixed and certain as opposed to tentative and evolving. The second dimension concerns students' view of knowledge as either isolated discrete truths and facts or as interrelated ideas and concepts. The self as explainer emphasis focuses on students' ability to explain and understand natural phenomena and theories. In this category, students' own thoughts and hypotheses are important as a starting point for developing a more advanced understanding. Thus, this category may be related to a third dimension that involves the *source of knowledge* [35] and concerns students' view of knowledge as either being transmitted from an external authority or as being actively constructed by individuals in interaction. In this perspective, this category is also related to a view that appropriation of knowledge consists of a personal or interpersonal interpretation.

The everyday coping emphasis stresses the knowledge that facilitates everyday life when students use or apply scientific knowledge. This could include practical applications of electricity at home, how to protect oneself from sexually transmitted diseases, understanding the nutritional content on a milk carton, or how to repair a bike, all of which are related to the practical application or use of scientific knowledge. Scientific skill development accentuates the practical parts of the disciplines of science, such as how to conduct an inquiry, plan an experiment, handle equipment, and what can be learned from these situations. This emphasis also includes sorting, observing, or describing a course of events or phenomena, together with understanding graphs or other representations. Considering the practical intradisciplinary issues, this category may also be related to an understanding of the relation between conducting experiments and the process of knowledge production. It implies that this category is, to some extent, related to the dimension of the *justification of knowledge* [34], that is, a view about science as objective reality versus science, as evaluated by scientific methods based on evidence.

The science, technology, and decisions category is related to how students use, apply, and consider scientific knowledge in order to reach valid and attentive decisions as citizens. This emphasis is often based on multi-disciplinary socioscientific issues and aims to contribute to students' becoming active members in society. Finally, structure of science describes issues about natural science as a discipline, its history, how

TABLE 1: Overview of how the emphases are used when categorizing the PISA items.

Solid foundation	Items that focus on science content that may facilitate future science instruction and education
Correct explanations	Items that focus on scientific facts and products without any requirement to understand underlying scientific models or theories
Self as explainer	Items demanding explanations and understanding of natural phenomena and theories
Everyday coping	Items emphasizing science as an important means for understanding and controlling everyday situations
Scientific skill development	Items that accentuate the practical parts of science, how to conduct an inquiry, plan an experiment, and handle equipment along with what can be learned from these situations
Science, technology, and decisions	Items focusing on applying and considering scientific knowledge to make valid and attentive decisions about the future and society
Structure of science	Items highlighting how science intellectually functions in its own growth and development, the interplay between evidence and theory, and the adequacy of a model to explain phenomena

new knowledge is created in the scientific community, and what constitutes valid and trustworthy knowledge. It also focuses on developing students' critical thinking in relation to science. In this way, this category is strongly related to the above-mentioned dimension about the sources of knowledge and how scientific knowledge is generated and to a broad definition of the *nature of science*. According to Lederman [36], the nature of science can be viewed as an understanding of scientific knowledge as tentative, empirically based, subjective, partly the product of human imagination and creativity, and socially and culturally embedded. An overview of the emphases is presented in Table 1.

As mentioned, the analytic procedure in this study includes a description and an analysis of the differences and similarities of national results on individual recurring items at different measurement occasions between 2000 and 2009. The results (P values) from each of the individual items and from different measurements have already been published by the PISA organization and are available for use by researchers and the public (see <http://www.oecd.org/pisa/pisaproducts/>). Consequently, we have not conducted any calculations of the mean values but have instead used the published results in order to interpret trends. However, the changes in the results between different measurement occasions have been calculated if there is a clearly ascending or descending trend through all measurement occasions. If we found the trend to be statistically significant ($P < 0.05$), we marked the figure in the table with an asterisk (*). In order to increase the reliability, four researchers conducted the analysis and the categorization of the items into emphases independently; if there were different interpretations, the material was reanalyzed in order to reach a consensus.

TABLE 2: The national mean results in three countries for 2000–2009, as well as the mean of all OECD countries.

Year	2000	2003	2006	2009
Sweden	512p	506p	503p	495p
Denmark	481p	475p	495p	499p
Finland	538p	548p	563p	554p
OECD	500p	500p	500p	501p

8. Results

When using the result descriptions from PISA [33], which emphasize participating countries' mean values, Sweden shows a decline between 2000 and 2009 compared to the mean value of participating OECD countries. The Danish results indicate the opposite, with an obvious increase during the same period. For Finland, the upward trend between 2000 and 2006 was broken in 2009. However, Finland displays significantly higher mean values compared to the OECD mean during all measurement occasions and, together with Shanghai (575p, year 2009) and Hong Kong (549p, year 2009), belongs to an exclusive group of top-scoring countries. Table 2 shows the results from the three Nordic countries, expressed as national mean values between 2000 and 2009.

9. Examples of Categorizations

In order to approach the aims of the study, we categorized all recurring items from PISA surveys during 2000–2009 into one or more curriculum emphases [14, 15]. The results of this analysis revealed that the most frequent categories related to these items were correct explanations, scientific skill development, and structure of science. This implies that the most frequently explored knowledge domains were related to situations in which students were requested to give fundamental explanations, handle experimental problems, interpret diagrams and tables, and solve problems related to the nature or structure of science. However, it is important to ensure that these results are related to an analysis of link items and not to all appurtenant items. The self as explainer and solid foundation categories were less frequent, while the everyday coping and science, technology, and decision categories were not represented at all. This means that respondents were not asked to consider decision-making processes in relation to science, technology, and society, or science as a means to handle everyday situations.

10. The Greenhouse Effect: Fact or Fiction?

An example of recurring items (2000–2006) was “The greenhouse effect: fact or fiction?” (item 114 Q03 and Q04). The students were given a background text in which the radiation balance of Earth is described briefly. The text also explained that “the Earth's atmosphere has the same effect as a greenhouse” and “the greenhouse effect is said to have become more pronounced during the twentieth century.” The students were asked to compare two diagrams concerning the discharge of carbon dioxide and the mean temperature

TABLE 3: Mean of percentage of correctness on two PISA items (114 Q3 and 114 Q4) in three countries and OECD during 2000–2006.

Item	Year	OECD	Den	Swe	Fin
114 Q03	2000	57.30	55.20	61.60	61.80
	2003	54.02	57.34	61.59	64.44
	2006	53.94	55.78	54.83	66.63
Diff		−3.36*	—	−6.77*	+4.83*
114 Q04	2000	39.50	37.70	40.50	49.10
	2003	35.99	33.98	40.13	44.53
	2006	34.47	34.62	29.91	47.64
Diff		−5.03*	—	−10.59*	—

of Earth between 1860 and 1990. The item is divided into separate tasks. In the first (Q03), students were expected to use the diagrams to find evidence that supports the hypothesis that the rise of the Earth's mean temperature is related to the increased discharge of carbon dioxide during the period. In the second (Q04), students were expected to interpret the diagrams to find counterevidence for the same hypothesis. The two tasks are formulated through a discussion between two imaginary students who present different opinions and arguments of how to interpret the graphs. One of the students (André) concludes: “It is certain that the increase in the average temperature of the Earth's atmosphere is due to the increase in the carbon dioxide emission.” Therefore, the first assignment (Q03) is to consider the graphs and find arguments that support this conclusion. In the second assignment (Q04), another imaginary student (Jeanne) disagrees with André's conclusion and argues that some parts of the graphs do not support his interpretation. The respondents are asked to “give an example of a part of the graphs that does not support André's conclusion” and explain their answer.

These two tasks have been categorized as belonging to the emphasis structure of science and scientific skill development. Regarding structure of science, this interpretation is derived from the assumption that one important prerequisite for solving these tasks is understanding that new scientific knowledge is generated from a discussion of the trustworthiness of different hypotheses and scientific evidence. The interpretation is enhanced through an argumentation between the imaginary students using the same graphs to support different hypotheses about the greenhouse effect. In order to solve the tasks, respondents must understand that scientific results are sometimes contradictory. For the scientific skill development emphasis, the interpretation is related to students' ability to use and interpret diagrams. Solving the assignments requires the use of separate parts of the diagrams and a comparison of the two. Table 3 describes the results from these two tasks in the three countries and the OECD mean.

One obvious result in relation to Table 3 is that the solution frequency are significantly higher on the first (Q03) task than the second (Q04) in all three countries, as is the case in all OECD countries. In other words, it seems more difficult for the students to interpret the diagrams to find

arguments that contradict the first hypothesis than to simply find proof from the curves that support it. One possible interpretation is that the first task (Q03) is easier, as it only requires students to find a common covariation of the two curves, while the second (Q04) requires them to find a lack of covariation during some periods. This interpretation is related to students' abilities to analyze diagrams and to the category of scientific skill development. Another possibility is that students generally find it more difficult to understand a situation or task in which several different hypotheses and interpretations are involved and negotiated, and a single correct answer could not be found. The last interpretation is more related to the category structure of science, as it requires an understanding of the process of how new knowledge in science may be generated from a discussion about the trustworthiness of different hypotheses.

With regard to the first task (Q03), there are some clear differences in the results between the presented countries. For example, the mean of percentage of correctness in the OECD countries shows a descending trend during all measurement occasions. This is also the case for the Swedish results, and the analysis shows that the decline is statistically significant. The Danish results do not indicate any trend at all, while the Finnish results on this item show a third pattern and display a clear and significant upward trend through all measurements.

Additionally, the results from the second task (Q04) show a clear descending trend in the results in the OECD countries, as well as in Sweden. The drop in the Swedish results is related to a decline between the measurements of 2003 and 2006, while the results in Finland and Denmark indicate unchanged results during all measurement periods. Although it is not possible to draw any conclusions from these results about the underlying causes, one can look to the existence of possible trends in the PISA data in relation to different types of assignments and content.

11. Clothes

Another example of a released link item is "clothes," which consists of two multiple-choice questions (213 Q01 and Q02). In the background text, the students were given a description about a British research team developing "intelligent clothes." The idea was to produce waistcoats made of a "unique electrotexile," linked to a "speech synthesizer" that enables disabled children to communicate with their surroundings. The first task (Q01) is to consider four different statements and decide whether these statements could be tested through scientific investigation. The four statements are as follows:

- (1) the material can be washed without being damaged.
- (2) The material can be wrapped around objects without being damaged.
- (3) The material can be scrunched up without being damaged.
- (4) The material can be mass-produced cheaply.

In the second task (Q02), the students are asked to choose what laboratory equipment they need in order to measure

TABLE 4: Mean of percentage of correctness on two PISA items (213 Q1 and 213 Q2) in three countries and in OECD during 2000–2006.

Item	Year	OECD	Den	Swe	Fin
213 Q1	2000	40.30	43.20	35.90	53.80
	2003	41.98	47.18	42.59	59.25
	2006	47.86	50.54	48.51	67.88
Diff		+7.56*	+7.34*	+12.61*	+14.08*
213 Q2	2000	75.90	68.60	81.80	90.40
	2003	76.22	72.36	80.37	91.66
	2006	79.37	75.91	83.99	94.93
Diff		+3.47*	+7.31*	—	+4.53*

whether the fabric is conducting electricity. The four alternatives are "voltmeter, light box, micrometer, and sound meter."

The first task (Q01) has been categorized into the scientific skill development and correct explanations emphases. This interpretation is related to the fact that the assignment assesses students' ability to determine what can be studied and measured in a simple scientific investigation. In this way, the described situation is similar to a typical experimental setting in the natural sciences. On the other hand, the task does not demand any deeper understanding of scientific inquiry, as it does not require knowledge of dependent and independent variables, or which variables are needed to keep constant. This implies that students may provide the correct explanation on the task, regardless of whether they understand the meaning of a scientific investigation. Similarly, the second item (Q02) is categorized as belonging to the correct explanations category. This interpretation is associated either with knowing that the voltmeter is the only equipment for measuring electricity, or how to exclude the other choices. Table 4 describes the results from these two tasks (2000–2006) in the three countries and the OECD mean.

As Table 4 shows, students in all countries achieved a higher mean of percentage of correctness on the second item (Q02) than the first (Q01). There was general improvement on both items during the period. Regarding the first question, the lower scores could be interpreted as meaning that more than one answer could be considered correct and that students probably lack experience from similar investigations. For the second task, it is reasonable to assume that most students had encountered a voltmeter in the practical parts of their science education and/or that the alternatives of "light box, micrometer, and sound meter" are connected to other areas. Nearly 80 percent of students in all OECD countries are able to connect measurements of electricity to a voltmeter and/or can exclude the other equipment. When it comes to the results of the involved countries, the progress on the first item (Q01) is distributed relatively evenly between the three measurement times. The Finnish and Swedish results show the most obvious improvement between the countries, although the OECD mean also indicated a general improvement on this item. On the second task (Q02), the Danish and Finnish results show significant improvement

during the period, and nearly all of the Finnish students (94.93 percent in 2006) answered this item correctly.

12. Result Description in Relation to Different Emphases

The result description in the following section focuses on student performances on items related to different emphases. This means that we intend to present the results from both released and unreleased PISA items from 2000 to 2009 and analyze the trend of percentage of correctness [12] on items categorized into the same emphasis. The aim of this approach was to explore and interpret the structure of the national trends in the PISA science results exemplified through the three countries. As mentioned, some of the items are categorized into more than one emphasis, so they are presented in more than one table.

Table 5 is an overview of the items categorized into the emphasis of correct explanations and the mean percentage of correctness on three or four measurement occasions between 2000 and 2009. Eight of the 10 presented items belonged only to this emphasis. One result is the ascending trend of several items in the three countries, as well as in the OECD. However, the upward trends are more pronounced in some countries. In seven of the listed items, the upward trend is apparent or indicates unchanged results in all presented countries, as well as in the OECD. Two of these items (213 Q01, Q02) belong to the released item “clothes,” which was discussed in the previous section. The third and fourth items (326 Q03 and 326 Q04) are only categorized into this emphasis and constitute typical examples as the students are only required to provide or reproduce a single correct answer without any further explanation of the phenomenon.

Further on, Table 5 displays four other items (256 Q01, 269 Q01, 326 Q01, and 326 Q02) for which student performances are relatively similar throughout the various measurement times. This means that there is no evident upward or downward trend in student achievements during the period. However, the Danish results diverge on item 326 Q01, as there is a downward trend and the percentage of correctness is apparently lower than in the other countries. One item (131 Q02) differs in that there is an evident downward trend in all countries. One possible explanation for this could be that this item consists of a relatively comprehensive background text, combined with the requirement of an open-ended response. Another item (269 Q4) diverges from the pattern, as the results display different developments or trends. The Danish results indicate an obvious upward trend, while the results from Sweden and the OECD countries show a weak descending trend. Finally, the Finnish results show a large drop between 2003 and 2006, although these results are strengthened to some extent in the 2009 measurement.

The analysis of students performances on most of the items categorized into correct explanations indicates either an obvious upward trend or unchanged performance. From these results, it can be concluded that the students in the three countries and OECD countries perform at about the same level or better on items that only require a single

TABLE 5: Students’ performances in terms of mean of percentage of correctness related to items categorized into the emphasis of *correct explanation*.

Item	Year	OECD	Den	Swe	Fin
131Q2	2000	50.60	50.10	58.00	67.40
	2003	46.41	45.22	53.16	67.20
	2006	46.24	44.36	52.07	65.50
Diff		−4.36*	−5.74*	−5.93*	−1.90
213Q1	2000	40.30	43.20	35.90	53.80
	2003	41.98	47.18	42.59	59.25
	2006	47.86	50.54	48.51	67.88
Diff		+7.56*	+7.34*	+12.61*	+14.08*
213Q2	2000	75.90	68.60	81.80	90.40
	2003	76.22	72.36	80.37	91.66
	2006	79.37	75.91	83.99	94.93
Diff		+3.47	+7.31*	—	+4.53*
269Q4	2003	35.76	30.19	40.15	41.29
	2006	34.11	35.91	38.70	33.58
	2009	33.03	41.93	37.99	34.95
Diff		−1.73	+11.74*	−2.16*	—
256Q1	2000	88.30	86.20	87.20	89.50
	2003	87.12	84.89	84.63	91.43
	2006	87.75	87.25	85.84	92.48
	2009	88.57	89.11	87.24	94.20
Diff		—	—	—	+4.70*
326Q1	2003	58.33	57.25	64.48	71.80
	2006	59.01	53.34	60.66	73.51
	2009	58.59	52.64	61.12	70.17
Diff		—	−4.61*	—	—
326Q2	2003	62.65	66.60	64.52	79.98
	2006	63.74	66.63	65.98	82.62
	2009	63.89	67.97	62.10	78.28
Diff		+1.24	+1.37	—	—
326Q3	2003	56.72	57.27	61.07	73.89
	2006	58.27	57.79	60.38	78.18
	2009	60.62	57.83	63.31	77.72
Diff		+3.90*	+0.56	—	—
326Q4	2003	22.33	15.57	23.50	34.52
	2006	23.31	22.01	26.72	37.68
	2009	25.32	23.89	28.58	37.63
Diff		+2.99*	+8.32*	+5.08*	—
269Q1	2000	59.10	41.10	64.10	64.00
	2003	59.65	43.14	64.58	65.95
	2006	57.76	42.99	56.27	66.15
	2009	58.00	44.18	61.25	65.71
Diff		—	—	—	—

correct answer. This implies a trend whereby students tend to achieve the same or higher scores on items that only required the correct word or an ability to find the correct alternative in multiple-choice questions. The question is how this trend may be understood and interpreted in relation to

TABLE 6: Students' performances in terms of mean of percentage of correctness related to items categorized into the emphasis *structure of science*.

Item	Year	OECD	Den	Swe	Fin
114Q03	2000	57.30	55.20	61.60	61.80
	2003	54.02	57.34	61.59	64.44
	2006	53.94	55.78	54.83	66.63
Diff		-3.36*	—	-6.77*	+4.83*
114Q04	2000	39.50	37.70	40.50	49.10
	2003	35.99	33.98	40.13	44.53
	2006	34.47	34.62	29.91	47.64
Diff		-5.03*	—	-10.59*	—
114Q05	2000	24.80	21.40	16.50	31.40
	2003	22.26	16.57	24.04	30.77
	2006	18.89	14.75	17.31	32.43
Diff		-5.91*	-6.65*	—	—
268Q02	2000	40.40	37.70	34.80	50.60
	2003	37.08	30.52	26.71	51.75
	2006	36.17	37.16	25.52	47.53
Diff		-4.23*	—	-9.28*	—

science education and instruction in the three countries. This question is examined further in the discussion section.

Table 6 presents an overview of student performance on the four items that are categorized into the structure of science emphasis. The results on the first two items (114 Q03 and 114 Q04), which showed an evident descending trend in Sweden as well as in OECD countries, were discussed above in relation to the "greenhouse effect." The third item (114 Q05) has been categorized into both structure of science and self as explainer. This means that the item contains elements of understanding how scientific knowledge is generated through the interplay between scientific evidence and theory and the adequacy of a model to explain a phenomenon. Furthermore, in items categorized as self as explainer, students are required not only to find a correct answer, but to provide an explanation for the scientific phenomenon. As the results show, solution frequencies are significantly lower on this item (114 Q05) at all measurements than they are in the previous items in this emphasis. Another result is the descending trend in both OECD countries and Denmark. The Finnish results on this item display a comparatively very high mean of percentage of correctness through all measurement occasions. The Swedish results do not show any specific trend on this item.

The final item in Table 6 (268 Q02) has been categorized into both structure of science and self as explainer. Furthermore, the results are similar, as the descending trend is evident in Sweden and the OECD countries. The trend is considerable in Sweden when comparing the 2000 results to those of 2003 and 2006. However, the Danish results do not reveal any obvious trend, as the 2003 measurement is significantly different from those of 2000 and 2006. Together, the analyses of students' performances on items related to the structure of science emphasis display a descending trend,

TABLE 7: Students' performances in terms of mean of percentage of correctness related to items categorized into the emphasis *self as explainer*.

Item	Year	OECD	Den	Swe	Fin
114 Q05	2000	24.80	21.40	16.50	31.40
	2003	22.26	16.57	24.04	30.77
	2006	18.89	14.75	17.31	32.43
Diff		-5.91*	-6.65*	—	—
268 Q02	2000	40.40	37.70	34.80	50.60
	2003	37.08	30.52	26.71	51.75
	2006	36.17	37.16	25.52	47.53
Diff		-4.23*	—	-9.28*	—
268 Q06	2000	57.80	60.40	57.00	52.40
	2003	55.86	53.27	57.11	54.44
	2006	55.20	61.88	53.03	56.41
Diff		-2.60*	—	—	+4.01*

which is most evident in Sweden, but consistent with the mean values of OECD countries. In fact, no country showed an upward trend on any items in this category. Therefore, these results imply a declining trend in which students from all of the included countries tend to perform lower on items that require an understanding of the structure or nature of science and the interplay between hypotheses, evidence, and theory.

The next results concern items categorized into the self as explainer emphasis. The first (114 Q05) and second (268 Q02) items were categorized into both self as explainer and structure of science and have been discussed previously. The third item (268 Q06) was categorized into self as explainer and solid foundation, which means that this item contains essential features of both understanding and explaining natural phenomena, as well as a crucial focus on science content that is important for future instruction. When analyzing the results on this item, only the mean value of OECD shows a descending trend. However, the Finnish students scored lower in the 2006 measurement than they did in 2000 and 2003. In comparison, Danish students performed higher in 2000 and 2006 than in 2003. Although this emphasis contained only three items (Table 7), the overall results suggest a descending trend in students performances in all three countries, as well as in the OECD mean value.

The solid foundation category had three items, an overview of which is presented in Table 8. The first item (131 Q04) agreed with both the emphasis of solid foundation and scientific skill development, since it focuses on science content for future instruction and also requires knowledge of how to conduct an inquiry, plan an experiment, and draw conclusions. Students performance on this item indicates an evident upward trend, both in Sweden and the OECD countries. The results do not indicate an obvious trend through all measurement occasions, either in Denmark or in Finland. However, the results in Denmark are significantly higher in the 2006 measurement than in 2000 and 2003, while the negative change in the Finnish results is apparent between 2000 and 2003. The next item (268 Q01) was

TABLE 8: Students' performances in terms of mean of percentage of correctness related to items categorized into the emphasis solid foundation.

Item	Year	OECD	Den	Swe	Fin
131 Q04	2000	24.70	18.20	19.10	48.30
	2003	26.12	15.93	33.10	40.26
	2006	31.11	24.56	33.15	45.44
Diff		+6.41*	—	+14.05*	—
268 Q01	2000	73.60	71.70	77.90	80.20
	2003	71.30	71.03	75.94	83.53
	2006	72.48	71.37	76.60	85.10
Diff		—	—	—	+4.90*
268 Q06	2000	57.80	60.40	57.00	52.40
	2003	55.86	53.27	57.11	54.44
	2006	55.20	61.88	53.03	56.41
Diff		-2.60	—	—	+4.01*

also categorized into solid foundation and scientific skill development. However, students performance on this item does not display any evident upward or downward trend, but shows similar results on all measurements. The exception is the Finnish result, which displays significant improvement during the period. The third item in this emphasis (268 Q06), presented earlier, was categorized into solid foundation and self as explainer.

Overall, there was no obvious trend regarding the items that belong to the solid foundation emphasis. The results on these items indicate both upward and downward trends, as well as unchanged performances in all three countries. This could mean that it is difficult to draw far-reaching conclusions from this way of conducting the analysis, or it may simply mean that there is no trend.

The final emphasis is scientific skill development; Table 9 presents an overview of the items included in this emphasis. All of these items are also categorized into a second emphasis, as has been presented previously. At first sight, it seems as though no evident trend existed, since the included items display both upward and downward trends, it is possible to discern a pattern. The descending trend on items 114 Q03 and 114 Q04 can be explained by the fact that they are also categorized into structure of science. The other three items (131 Q04, 213 Q01, and 268 Q01), all of which display an upward or maintained trend, are categorized into solid foundation or correct explanations. It seems that, in this context, scientific skill development is subordinate to the other emphases, which means that the trend of the results may be related to whether the content of the item agreed with another emphasis. This, in turn, implies that the other three categories appear to influence the trend development more than the scientific skill development category.

13. Discussion

This paper highlights the discussion within the international research community concerning how the results from large-scale studies are used in order to understand or interpret

TABLE 9: Students' performances in terms of mean of percentage of correctness related to items categorized into the emphasis scientific skill development.

Item	Year	OECD	Den	Swe	Fin
114Q03	2000	57.30	55.20	61.60	61.80
	2003	54.02	57.34	61.59	64.44
	2006	53.94	55.78	54.83	66.63
Diff		-3.36*	—	-6.77*	+4.83*
114Q04	2000	39.50	37.70	40.50	49.10
	2003	35.99	33.98	40.13	44.53
	2006	34.47	34.62	29.91	47.64
Diff		-5.03*	—	-10.59*	—
131Q04	2000	24.70	18.20	19.10	48.30
	2003	26.12	15.93	33.10	40.26
	2006	31.11	24.56	33.15	45.44
Diff		+6.41*	—	+14.05*	—
213Q01	2000	40.30	43.20	35.90	53.80
	2003	41.98	47.18	42.59	59.25
	2006	47.86	50.54	48.51	67.88
Diff		+7.56*	+7.34*	+12.61*	+14.08*
268Q01	2000	73.60	71.70	77.90	80.20
	2003	71.30	71.03	75.94	83.53
	2006	72.48	71.37	76.60	85.10
Diff		—	—	—	+4.90*

students' knowledge in and about science. In this context, the reliability of the PISA framework has been discussed and questioned (e.g., [21, 23, 24]), as well as to what extent PISA results constitute valid representations of students knowledge (e.g., [5, 7, 8]). To conclude this discussion, several scholars (e.g., [20, 31]) have asserted critical standpoints about the PISA framework and OECDs description and presentation of the results as one-dimensional rankings and league tables. They also argue that this way of displaying the results seldom increases understanding of science teaching and learning from a classroom perspective, but instead serves educational policy at a superficial level.

However, some scholars have suggested alternative or complementary approaches for analyzing and exploring students' knowledge and understanding of science in national and international perspectives (e.g., [12, 27]). A common feature of these proposals is an interest in going beyond the constructed national mean values to reveal hidden patterns in the material. In this way, the data can be used for different purposes and research perspectives. The present study suggests a complementary analysis that uses the raw data from four PISA measurements to explore possible epistemological trends in students understanding of science based on the framework of *curriculum emphases* [14, 15]. The analyses were facilitated and clarified by relating the emphases to Hofer's [34] and Hofer and Pintrich's [35] *epistemic dimensions*. We argue that an important conclusion is that the use of the emphases provides possibilities for interpreting epistemological trends in students' understanding from the perspective of large-scale studies. This implies that we have categorized all

recurring items (link items) from 2000–2009 into different emphases, compared the results as mean of percentage of correctness, and analyzed the trends exemplified here by three Nordic countries.

Consequently, one of the study's main results is that it is possible and feasible to categorize the items into curriculum emphases and discern general ascending or descending trends in the material. For example, items categorized into correct explanations display a general upward or maintained trend in all three countries and the OECD mean. These results indicate a tendency for students to perform higher or at the same level on items that only required a single correct answer or one right answer on multiple-choice questions. These results may also be considered from the perspective of Hofer's [34] *simplicity of knowledge*, whereby students risk viewing scientific knowledge exclusively as isolated discrete truths and facts, which means that they risk losing important knowledge dimensions in science.

Another obvious result is the general downward trend in Sweden and in the OECD countries overall for items categorized into structure of science. The drop is rather dramatic for certain items; in some cases, students performance decreased by 10 percentage points or more. With some exceptions, however, the Finnish and Danish results for these kinds of items are unchanged during the measurement period. These results indicate an epistemological trend in Sweden as well in OECD whereby students' understanding of science as tentative, empirically based, socially and culturally embedded, and as a process of knowledge production seems to be decreasing. An interesting related question is what this trend actually indicates and if it can be understood as a change in educational focus and ways of presenting science in some of the countries and in the OECD.

The analysis also indicates a general downward trend throughout the OECD countries for items categorized as self as explainer. The Swedish and Danish results partly confirm this image, although they also display an unchanged performance related to specific items, while the Finnish results indicate an upward or maintained trend on all items. To solve items in this category, students need to understand and independently explain natural phenomena and theories. In other words, it requires that students are able to use scientific knowledge as explanatory models and in addition have the ability to use these models in different contexts. From this perspective, it is related to Hofer and Pintrich's [35] third dimension; the *source of knowledge* and it concerns a view of knowledge as being either transmitted from an external authority or as actively constructed by individuals in interaction. It also requires students to be somewhat aware that learning and appropriation of knowledge consist of personal or interpersonal interpretations. However, only a few items were categorized into this emphasis, which implies that this finding demands further attention and investigation. In addition, there does not seem to be any evident trend for items categorized into solid foundation or scientific skill development.

Accordingly, the results of the study indicate that the trends cannot exclusively be understood from the perspective

that science as a school subject constitutes a one-dimensional entity, but rather that it consists of many different knowledge forms that require different competencies and understanding. In this context, the study clearly indicates that there are different trends depending on which epistemological focus the items intend to measure. From an international OECD perspective, the results indicate a general downward trend for items that focus on the nature of science, how new scientific knowledge is generated, and how different theories and hypotheses are negotiated in order to reach consensus in the scientific community. It is also possible to discern a similar trend when it comes to the ability to use explanatory models to solve scientific problems in different contexts and understand that science knowledge is built on models of reality. On the other hand, there is a general upward trend regarding tasks that are intended to measure fact-based elementary knowledge. These trends are slightly differentiated at the national level, as the changes may occur at different times and have different magnitudes. In connection to these results, it is important to ask whether the focus and intention of science teaching and instruction in OECD countries has changed toward an increasingly one-dimensional and reproducible view on scientific knowledge. In this context, it is important to observe that the Finnish and, to some extent the Danish results, indicate an opposite or more complex image.

In order to evaluate or validate the proposed method, we should note that the data material was rather limited for some specific curriculum emphases and more extensive regarding others; therefore, caution should be taken against drawing too extensive conclusions. Another weakness related to the method was that PISA items often seem to be connected to more than one of the emphases, which implies problems when it comes to refining how the students' performances were related to some of the emphases. This was the case with the scientific skill development category, which tended to depend on other emphases into which the item was categorized.

Nevertheless, by way of conclusion, we suggest using complementary or alternative frameworks from the education research community to interpret and understand results from large-scale studies. This will allow researchers to approach students' epistemological understanding of science from national and international perspectives. We argue that the data from these studies offers opportunities to move beyond the superficial level of national mean values and league tables and explore possible trends and tendencies in the material. The data material, in the form of the percentage of correctness on each PISA item, is also easily available for the public; this may provide prerequisites for professional science teachers to understand and make use of the results in their everyday classroom activities.

Funding

This research was funded by the Swedish Research Council (Dnr 2008-4717).

References

- [1] OECD, *Measuring Student Knowledge and Skills: A New Framework For Assessment*, OECD Publications, Paris, France, 1999.
- [2] OECD, *PISA, 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, OECD Publications, Paris, France, 2003.
- [3] OECD, *Executive Summary PISA, 2006: Science Competencies For Tomorrow's World*, OECD Publications, Paris, France, 2007.
- [4] A. Jakobsson, R. Säljö, and Å. Mäkitalo, "Conceptions of knowledge in research on students. Understanding of the greenhouse effect: methodological positions and their consequences for representations of knowing," *Science Education*, vol. 93, no. 6, pp. 978–995, 2009.
- [5] M. Brunner, C. Artelt, S. Krauss, and J. Baumert, "Coaching for the PISA test," *Learning and Instruction*, vol. 17, no. 2, pp. 111–122, 2007.
- [6] OECD, *Learning for Tomorrow's World. First Results from PISA 2003*, OECD Publications, Paris, France, 2004.
- [7] S. Sjöberg, "PISA and 'Real Life Challenges' : mission impossible?" in *PISA According To PISA: Does PISA Keep What It Promises?* T. Hopmann, G. Brinek, and M. Retzl, Eds., pp. 203–224, LIT, Wien, Vienna, 2007.
- [8] E. Bautier and P. Rayou, "What PISA really evaluates: literacy or students' universes of reference?" *Journal of Educational Change*, vol. 8, no. 4, pp. 359–364, 2007.
- [9] R. Olsen, "An exploration of cluster structure in scientific literacy in PISA: evidence for a Nordic dimension?" *Nordic Studies of Science Education*, vol. 1, no. 1, pp. 81–94, 2005.
- [10] M. Puchhammer, "Language-based item analysis—problems in intercultural comparisons," in *PISA According to PISA: Does PISA Keep What It Promises?* T. Hopmann, G. Brinek, and M. Retzl, Eds., pp. 127–138, LIT, Wien, Vienna, 2007.
- [11] N. Bottani and P. Vignaud, *La France Et Les Évaluations internationales*. National Report To the HCEE, HCEE, Paris, France, 2005.
- [12] P. J. Fensham, "Real world contexts in PISA science: implications for context-based science education," *Journal of Research in Science Teaching*, vol. 46, no. 8, pp. 884–896, 2009.
- [13] K. Klette, S. Lie, M. Ødegaard et al., *Prosjekt om Lærings- og Undervisnings-Strategier i Skolen [Project about Learning and Educational Strategies in Schools]*, Forskningsrådet, Oslo, Norway, 2008.
- [14] D. A. Roberts, "Developing the concepts of "curriculum emphases" in science education," *Science Education*, vol. 66, no. 2, pp. 243–260, 1982.
- [15] D. A. Roberts, "Promoting scientific literacy: science education research in transition," in *Proceedings of the Linnaeus Tercentenary Symposium*, C. Linder, L. Östman, and P.-O. Wickman, Eds., pp. 9–17, Uppsala University, Uppsala, Sweden, 2007, Linné Scientific Literacy Symposium Opening Re-marks.
- [16] OECD, *Assessing Scientific, Reading and Mathematical Literacy: A Framework For PISA 2006*, OECD Publications, Paris, France, 2006.
- [17] S. Grek, "Governing by numbers: the PISA "effect" in Europe," *Journal of Education Policy*, vol. 24, no. 1, pp. 23–37, 2009.
- [18] J. Wuttke, "Fehler, Verzerrungen, Unsicherheiten in der PISA Auswertung [Mistakes, biases, uncertainties in the PISA analysis]," in *PISA & Co: Kritik Eines Programms [PISA & Co: Critique of A programme]*, T. Jahnke and W. Meyerhöfers, Eds., pp. 101–154, Berlin, Germany, 2006.
- [19] SOU, "En hållbar lärarutbildning [A Sustainable Teacher Education]," SOU: 109, Statens offentliga utredningar, Ministry of Education in Sweden, 2008.
- [20] M. Uljens, "The hidden curriculum of PISA—the promotion of neo-liberal policy by educational assessment," in *PISA According To PISA: Does PISA Keep What It Promises?* T. Hopmann, G. Brinek, and M. Retzl, Eds., pp. 295–304, LIT, Wien, Vienna, 2007.
- [21] T. Hopmann and G. Brinek, "PISA according to PISA: does PISA keep what it promises?" in *PISA According To PISA: Does PISA Keep What It Promises?* T. Hopmann, G. Brinek, and M. Retzl, Eds., pp. 9–20, LIT, Wien, Vienna, 2007.
- [22] H. Goldstein, "International comparisons of student attainment: some issues arising from the PISA study," *Assessment in Education*, vol. 11, no. 3, pp. 319–330, 2004.
- [23] M. Fertig, "What can we learn from international student performance studies—some methodological remarks," Discussion Paper 23, RWI, Essen, Germany, 2004.
- [24] P. Allerup, "Identification of group differences using PISA scales—considering effects of inhomogeneous items," in *TPISA According To PISA: Does PISA Keep What It Promises?* T. Hopmann, G. Brinek, and M. Retzl, Eds., pp. 175–202, LIT, Wien, Vienna, 2007.
- [25] P. Lietz, "A meta-analysis of gender differences in reading achievement at secondary school level," *Studies in Educational Evaluation*, vol. 32, no. 4, pp. 317–344, 2006.
- [26] M. Kjærnsli and S. Lie, "PISA and scientific literacy: similarities and differences between the Nordic countries," *Scandinavian Journal of Educational Research*, vol. 48, no. 3, pp. 271–286, 2004.
- [27] S. Lavonen, S. Lie, A. Macdonald, M. Oscarsson, C. Reistrup, and H. Sørensen, "Science education, the science curriculum and PISA," in *Northern Lights on PISA, 2006: Differences and Similarities in the Nordic Countries*, T. Matti, Ed., vol. 547, pp. 31–58, Nordic Council of Ministers, Copenhagen, Denmark, 2009.
- [28] M. Kjærnsli and S. Lie, "Students' preference for science careers: International comparisons based on PISA 2006," *International Journal of Science Education*, vol. 33, no. 1, pp. 121–144, 2011.
- [29] R. V. Olsen and S. Lie, "Profiles of students' interest in science issues around the world: analysis of data from PISA 2006," *International Journal of Science Education*, vol. 33, no. 1, pp. 97–120, 2011.
- [30] G. Bonnet, "Reflections in a critical eye: on the pitfalls of international assessment," *Assessment in Education*, vol. 9, no. 3, pp. 387–399, 2002.
- [31] P. Mortimore, "Alternative models for analysing and representing countries' performance in PISA," Paper commissioned by Education International Research Institute, Brussels, Belgium, 2009.
- [32] U. P. Lundgren, "PISA as a political instrument—one history behind the formulating of the PISA programme," in *PISA Under Examination—Changing Knowledge, Changing Tests, and Changing Schools*, M. Pereyra, H.-G. Kotthoff, and R. Cowen, Eds., pp. 17–30, Sense, The Netherlands, 2011.
- [33] OECD, *PISA Results 2009. What Students Know and Can Do. Student Performance in Reading, Mathematics and Science*, vol. 1, OECD Publications, Paris, France, 2010.
- [34] B. K. Hofer, "Personal epistemology research: implications for learning and instruction," *Educational Psychology Review*, vol. 13, no. 4, pp. 353–383, 2001.
- [35] B. K. Hofer and P. R. Pintrich, Eds., *Personal Epistemology: the Psychology of Beliefs About Knowledge and Knowing*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1st edition, 2002.

- [36] N. G. Lederman, "Nature of science: past, present, and future," in *Handbook of Research in Science Education*, S. K. Abell and N. G. Lederman, Eds., pp. 831–880, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2007.

