Eva Davidsson är forskare inom naturvetenskapernas didaktik vid Malmö högskola med vetenskapskommunikation som specialområde. Sedan flera år deltar hon inom det VR- finansierade forskningsprojektet Perspectives of Large Scale Studies (PELS) som har som syfte att analysera svenska elevers kunskaper i naturvetenskap utifrån de storskaliga studierna.

Helene Sørensen är Associate professor i naturvetenskaps didaktik vid DPU, Aarhus universitet och har under lång tid varit involverad i de danska PISA-studierna. Hon har en lång forskarerfarenhet av genusaspekter i relation till naturvetenskaplig undervisning. Helene är även engagerad i PELS-projektet.

Peter Allerup är professor i pedagogisk statistik vid DPU, Aarhus universitet och är sedan flera år engagerad och ansvarig för flera storskaliga studier såsom danska TIMSS. Peter är även engagerad i PELS-projektet.

EVA DAVIDSSON

Malmö University, Sweden
eva.davidsson@mah.se

HELENE SØRENSEN

Department of curriculum research, Danish School of Education, Aarhus University, Denmark
helene@dpu.dk

PETER ALLERUP

Institute for learning, Danish School of Education, Aarhus University, Denmark
nimmo@dpu.dk

Assessing scientific literacy through computer-based tests – consequences related to content and gender

Abstract

Large-scale studies are now common tools for assessing students' knowledge in science, and there is an increased interest in conducting evaluations using computers. However, several studies indicate that students' achievements on computer-based science tests often involve significant gender differences: boys outperform girls. This paper explores how students' performances are affected by the content and the context of the test, Computer-Based Assessment in Science (CBAS), from a gender perspective. The framework of curriculum emphases is used to describe and analyze the CBAS items and to accomplish a rich description of what science may be. The results indicate different gender patterns depending on the items' content and context. Furthermore, the analysis reveals that the portrayal of scientific literacy tends to be dependent on the test mode.

INTRODUCTION

In recent years, large-scale studies such as TIMSS (IEA, 1997, 2008) and PISA (OECD, 2001, 2004, 2007) have increasingly been used to assess and compare students' achievements in science. Among other things, the tests have been designed to measure students' scientific literacy, knowledge of scientific facts, conceptual understanding, and scientific reasoning. In educational policy and decision-making, such tests play a reinforced importance in monitoring, for example, reorganizations of school systems, and curriculum development (Jakobsson, Säljö, & Mäketalo, 2009). However, there has

been increasing interest in administering such tests by means of computers. This interest in computer-based tests may be sparked by potential advantages, such as eliminating printing and distribution of booklets as well as reducing the costs of marking the students' answers. Computer-based tests also allow for a deeper assessment of other competencies such as conducting experiments and simulations. The increased use of computer-based tests is evident in the Organisation for Economic Co-operation and Development (OECD) and, according to Martin (2009), these tests have the potential to one day replace the PISA paper-and-pencil-based study. However, certain important questions remain: (1) How does the computer-based approach affect how science is represented and characterized, and (2) What is the impact of this approach on girls' and boys' achievements?

The development toward computer-based tests in relation to the PISA test has already started. The study Computer-Based Assessment in Science (CBAS) was carried out in the same year as the main study of PISA 2006 and within the same theoretical framework for assessment of scientific literacy. One finding of the CBAS study was the large gap between girls' and boys' achievements, where the boys perform significantly better than the girls, in all of the participating countries (OECD, 2010). The PISA main study did not demonstrate the same unambiguous image. However, what do the gender differences consist of and is it possible to distinguish gender patterns in the results of CBAS regarding the items' content and context?

This paper aims to explore in what ways girls' and boys' performances are affected by the content and the context of the CBAS test. The analytic framework of curriculum emphases (Roberts, 1982, 1998, 2011) is used to describe and characterize the CBAS items and to accomplish a richer description of what science may be as compared to the PISA framework of scientific competencies. Curriculum emphases refer to different sets of messages, or foci, that science teaching and learning could have in the classroom and that could be communicated both explicitly and implicitly. It was possible to use the framework for analyzing the CBAS items in similar ways. Furthermore, the paper discusses how the test mode may influence the portrayal of scientific literacy and how the image of science can depend on the aspects of science chosen to be included in the test.

BACKGROUND

Computer-Based Testing and Large-Scale Studies

Within the PISA administration, there is an increased interest and an explicit will to conduct future large-scale tests on computers (OECD, 2010). However, managing computer-based science tests raises a number of new issues, both in relation to technical and participant matters, as well as concerning content and epistemological issues. For example, paper-and pencil-tests and computer-based tests are often compared to and revolve around benefits of computer-based tests, such as practical issues and administrative gains, but also around technical possibilities of the media. This means that factors such as reduced costs of, for example, coding and the opportunities to include sound and visual effects in the items are often emphasized in this discussion (e.g. Bridgeman, 2009; Lee, 2008; Bennett et al., 2003). Other advantages include a range of options in the test design: the opportunity to interact with the test items, use simulations, and include video and audio applications (e.g. Lee, 2009). The question is whether the same skills and knowledge could be assessed equally by computer-based tests as by tests on paper. An ongoing discussion exists within the field of assessment concerning equivalence between the two test modes, and McDonald (2002) argues that reaching equivalence is crucial in order to compare results from separate studies. However, it has proven difficult to reach equivalence as the different test modes seem to provide the respondents with different experiences. Several studies have explored the extent to which different factors could influence the participants' achievements on computer-based tests (e.g., Leeson, 2006).

Leeson (2006) distinguishes between technical factors and participant issues, where the first refers to a number of factors related to, for example, screen size and resolution, line lengths, fonts used, and need for scrolling; those factors could influence the performance of the respondents. Research has shown that low screen resolution has a negative impact on reading tests but not on mathematics tests (Bridgeman, Lennon, & Jackenthal, 2002), that medium line length increases reading comprehension compared to short and long lengths (Dyson & Haselgrove, 2001), and that the possibility to review individual items usually improves the respondent's results (Revuelta, Ximénez, & Olea, 2003). However, Leeson (2006) argues that, although the availability for item review is taken for granted in paper-and-pencil-based tests, it is usually not an option when taking the test on computers and definitely not an option on computer adaptive tests, where the next item on the test is selected on the basis of previous answers. In their study, Goldberg and Pedulla (2002) conclude that the undergraduate students who took a paper-and-pencil-test outperformed the group that took the same test on the computer without the option to review the items.

McDonald (2002) identifies three participant-related factors that could influence the possibility to compare results from tests on paper and tests on computers: computer anxiety, computer experience and familiarity, and computer attitudes. Computer anxiety refers to the fear experienced when taking a test on a computer. It can be argued that computer anxiety derives from lack of computer experience; however, Chua, Chen, and Wong (1999) describe a more complex relationship in their meta-analysis of computer anxiety. The researchers confirm a correlation between these two factors, but conclude that the strength of this correlation varies considerably between different studies. Computer experience and familiarity have a significant impact on the test-takers' results. Horkay, Bennett, Allen, Kaplan, and Yan (2006) compared scores on essay writing on computers to essay writing on paper and found that computer familiarity significantly predicted the performance of the online writing test after controlling for paper writing skill. The authors suggest that there is a risk of underestimating the performance of a whole population if students are not allowed to choose their preferred test mode, because a substantial number of students perform better on the computer-based tests than on paper-and-pencil-based tests and vice versa.

Volman, Eck, Heemskerk, and Kuiper (2005) explored students' computer familiarity and controlled for gender and foreign or domestic background. The researchers conclude that greater differences exist between the children of different backgrounds than between girls and boys. However, boys stated that they had experiences with computers more often than girls. Furthermore, there was a gender difference in computer attitudes. The girls were less positive toward computers than the boys. Colley and Comber (2003) arrived at similar results, but argue that the gender gap has decreased, both regarding computer experience and also concerning computer attitudes since the 1990s, particularly for older children.

Computer-Based Assessment of Scientific Literacy

Apart from participant and technical issues, computer-based tests may raise concerns around how the science content is represented and outlined. For example, in what way does the test mode bring consequences to what is tested and how students' scientific literacy could be assessed?

In order to approach the aim of conducting the future PISA tests on computers, the OECD launched the computer-based assessment in science (CBAS) in 2006. The CBAS test was carried out in connection with PISA when students in Denmark, Iceland, and South Korea, in addition to taking the main test, also took the CBAS test. The central aim of CBAS was the same as for the PISA test—to evaluate students' scientific literacy. The purpose was to explore the students' (a) Scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena,

and to draw evidence-based conclusions about science-related issues, (b) Understanding the characteristic features of science as a form of human knowledge and enquiry, (c) Awareness of how science and technology shape our material, intellectual, and cultural environments (d) Willingness to engage with science-related issues, and with the ideas of science, as a reflective citizen. In order to evaluate these goals, three science competencies are in focus in both tests:

- Identifying scientific issues
- Explaining phenomena scientifically
- Using scientific evidence (OECD, 2007)

The CBAS test was thus developed within the same frame as the PISA study, but with the aim to include aspects of scientific literacy, which could not adequately be tested through the PISA main test (Martin, 2009). Accordingly, items that could be included in the PISA test were avoided in the CBAS test. In addition, visual stimuli were to be used when possible to reduce the reading load, and there were no open-ended response items. One consequence of this approach was that the CBAS test provided faster translation processes and automated coding procedures (OECD, 2007). Another consequence was that the total amount of text was reduced by about 30% in the CBAS test.

The results of the CBAS test showed interesting similarities and differences compared to the PISA main study. For example, Halldórsson, McKelvie, and Björnsson (2009) argue that the items of the CBAS test were easier for the students than those in the PISA study, because the percentages of correct answers were significantly higher in the CBAS test in all three participating countries. Another prominent feature, independent of gender patterns on the PISA test, is that the boys outperform the girls on the CBAS test in all three countries. This implies that the CBAS test showed new or enhanced gender differences that were not explicit on the main test. One possible explanation to these patterns could, according to several studies (e.g., Turmo & Lie, 2006; Sørensen & Andersen, 2009), be the overrepresentation of physical and earth system items, where boys usually perform better and fewer items of living systems where girls traditionally score better. In addition, Sørensen and Andersen (2007) highlight the gendered context in which, for example, all actors in the video clips are males. The fact that the reading load was reduced could, according to Halldórsson, McKelvie, and Björnsson (2009) also favor the boys. Furthermore, the results of CBAS show a gender difference regarding the level of interactivity of the items. According to OECD (2010), the boys outperform the girls on items, demanding high interactivity, and the more advanced the task became, the wider the gender gap.

Girls' and Boys' Attitudes Toward Science

When discussing students' performances on large-scale tests in relation to a gender perspective, several studies within the area of science education point to the significance of students' general attitudes toward science. In the previous section, we have highlighted several factors such as reading load and level of interactivity, which could have influenced the results, but what importance may the representation and characterization of science play, from a content perspective? This means, in what ways do girls and boys experience different subject matters and thematic content?

For the PISA test 2006, students were asked to answer a number of questions about their interests and attitudes towards science. About half the students stated that they were interested in topics about physics and chemistry (OECD, 2007). From the OECD average, nearly 70% of the students expressed an interest in human biology. When asked about the biology of plants and the way scientists design experiments, 47% and 49%, respectively, expressed positive attitudes. There were no significant gender differences in these results. In PISA 2006, students in the Nordic countries were among the most sceptical when it comes to considering the value of science in society as well as appreciating the personal value of science. Moreover, analysis indicates a correlation between test scores and to what extent the students find a personal value in science. In 43 of the participating countries, students who consider science to have a high personal value perform better than those who do not hold that view (ibid).

Osborne, Simon, and Collins (2003) argue that the gender aspect is crucial for understanding students' attitudes toward science. Based on analyses of a number of studies (e.g., Hendley et al., 1996; Jones, Howe, & Rua, 2000; Sjøberg, 2000), researchers conclude that girls' attitudes toward science, in general, are significantly less positive than boys' attitudes. Warrington and Younger (2000) explored English students' (ages 15–16) views about science and found that 63% of the boys stated that they liked science compared to only 42% of the girls. Moreover, 37% of the boys considered science their favorite subject, compared to just 6% of the girls. Several other studies also reveal gender differences in the interest of biology, chemistry, and physics. Osborne and Collins (2001) conclude that students (age 16) responded less favorably on the relevance and appeal of chemistry than on physics. They also found that both girls and boys expressed positive attitudes toward biology, but a gender difference was still evident: the girls tended to express many more positive aspects of biology than of physics and chemistry, and girls, more than boys, expressed an interest in aspects related to themselves, such as human biology.

Jones, Howe, and Rua (2000) showed significant gender differences regarding science experiences, attitudes, and perceptions of science courses. Boys, to a larger extent than girls, reported more extracurricular experiences with tools and expressed greater interest in atoms, cars, computers, and technology. Girls expressed greater interest in healthy eating, animal communication, and diseases. Murphy and Whitelegg (2006) discuss a number of studies related to girls' and boys' interests in, motivation for, perceived relevance of, and perception of difficulty of physics education. The authors argue that teachers must integrate and strengthen the relation between physics and social application by using a humanistic approach in teaching.

Curriculum Emphases and Science Competencies

Within the PISA framework, three different science competencies have been in focus, and those correspond to the needs and demands from the modern society of being a scientific literate citizen (OECD, 2007). An advantage of this approach is that the competencies express an ambition to cover students' knowledge in and about science. However, there are several comprehensive and overarching frameworks within the research area of science education for describing how science may be constituted and outlined in science teaching and learning situations as well as from a curriculum perspective (e.g. AAAS, 1990).

For example, Roberts (1982, 1998, 2011) uses the framework of curriculum emphases, which refer to different sets of messages in science education about science, which could be communicated both explicitly and implicitly. Such messages go beyond facts, theories, or principles and could instead provide answers to the students' question: Why am I learning this? The curriculum emphases show that different orientations are given in teaching and learning situations, and imply what science is about as well as its intent. This framework not only provides a more extensive and richer description of the students' knowledge and competencies, but it also describes the lack of knowledge related to specific aims. An overview of the emphases is presented in Table 1.

Roberts' emphases originally were set out to analyze the content and intentions of science curricula and schoolbooks. He concludes that science education and instruction always consist of at least one of the curriculum emphases that highlight different aims in the curriculum and, because of that, bring consequences to the alignment of the teaching and learning situation. This means that the framework of curriculum emphases could provide a resource for deepening and broadening the discussion of science competencies related to demands of being a scientific literate citizen and also could serve as an analytic tool for investigating what, in fact, is tested.

Table 1. Overview of different curriculum emphases (Roberts, 1982, 1998).

Everyday coping	Science is considered as an important means for understanding and controlling one's environment.
Structure of science	Concerns how science functions intellectually in its own growth and development. It focuses on the interplay between evidence and theory, the adequacy of a model for explaining phenomena, the changing and self-corrected nature of scientific knowledge, etc.
Science, technology, and decisions	Concentrating on the limits of science in coping with practical affairs
Scientific skill development	Competence in using processes basic in science and communicating the implicit message that skilful use of means (processes) automatically yield a correct end (product)
Correct explanations	Stresses the products
Self as explainer	Emphasises science as a cultural institution and as an expression of man's many capabilities. Examines the growth and change in scientific ideas as a function of human purposes, and of the intellectual and cultural preoccupations of the particular settings in which these ideas were developed and refined.
Solid foundation	Science content which may facilitate future science instruction

THE STUDY

As argued, there is an increased interest in monitoring large-scale tests on computers (e.g. Martin, 2009) and there are far-reaching plans within the OECD organization for launching computer-based science tests in 2015. In relation to this development, the discussion has highlighted issues about managing computer-based science tests and the fact that this test mode raises new questions about how science can be represented and constituted. Researchers have emphasized these issues in relation to a gender perspective. The results from the CBAS test 2006 indicate significant gender differences compared to the PISA main study, which accentuates the question of what is tested and how students' scientific literacy could be assessed and understood. As mentioned, the PISA framework applies three science competencies for approaching the task of assessing scientific literacy. However, we found it fruitful and possible to use Roberts' (1982, 1998, 2011) framework of curriculum emphases to accomplish a complementary description of what is being tested, and thereafter, relate the results to girls' and boys' achievements. The aim of this article is to explore how girls' and boys' performances are affected by the way science is represented in the CBAS test. Thus, the analysis does not include a comparison of how the different items are distributed in the two tests, but focuses more on how science is constituted in the CBAS test and then relates these results to the gender differences.

The research question for this study is as follows:

- In what ways are girls' and boys' achievements affected by the science content and how science is represented in the context of the Computer-Based Assessment in Science test?

Methodological Considerations

The CBAS test was conducted along with the PISA study in 2006. The students who participated in CBAS did so within a few days after taking the PISA main study. As described earlier, the CBAS test consisted of different items and comprised different science content than the PISA test. The CBAS test was administrated by a test team, which set up a local network of standardized laptops, on which the students took the test. The study presented in this article, includes only those students who answered

all of the CBAS items. Therefore, in total, 3,095 students were included in the analysis and distributed as 837 Danish, 781 Icelandic, and 1477 Korean students.

As mentioned, each item in the CBAS test was designed to test one of three science competencies: explaining phenomena scientifically, identifying scientific issues, or using scientific evidence (OECD, 2007). The items sought to test the competencies in relation to either knowledge of science (physical systems, living systems, and earth and space systems) or knowledge about science (scientific enquiry, scientific explanations, and science and technology). This means that knowledge of science was assessed from the respondents' abilities to explain phenomena scientifically, and their knowledge about science was tested from their abilities to identify scientific issues and use scientific evidence. This study first explored whether there were significant gender differences in the CBAS test scores for the different science competencies used in the OECD framework. However, the analysis revealed no significant variation; the boys performed significantly better in relation to all science competencies. In relation to this, it is important to remember that it is not possible to identify a corresponding difference in the PISA paper-and-pencil-test. One possible explanation for these results is the one proposed by Lau (2009), who analyzed the categorization of the items in the PISA test and found that items designed to assess ability to identify scientific issues or use scientific evidence also contained significant elements of explaining phenomena scientifically. According to Lau, this could mean that knowledge of science can constitute a hurdle for assessing knowledge about science and thereby could risk making this part less valid.

In order to circumvent use of the PISA scientific framework, we instead chose to analyze the CBAS items from the perspective of curriculum emphases (Roberts, 1982, 1998, 2011). We argue that it is possible to interpret written items in tests as representations or expressions of similar emphases. This implies that each item may agree with one or several of the emphases: *everyday coping*; *structure of science*; *science, technology, and decisions*; *scientific skill development*; *correct explanations*; *self as explainer*; and *solid foundation*. The advantage of this categorization compared to the PISA framework is that the analysis may be more detailed and content-related, but it also recognizes science from a more comprehensive perspective. The comprehensive aim of this procedure is to complement the PISA framework in order to reveal possible hidden patterns in the material. Our expectations are that it is possible to explain the emerged gender difference by using this framework.

The categorization of the items provided an overview of the distribution according to the different emphases. The preliminary results of the categorization showed that three emphases were overrepresented: correct explanations, scientific skill development, and everyday coping; the following analyses involve these categories. As the next step, the sum of correct answers was calculated for each respondent and within each category, and also the mean values of the sums of correct answers. For each of the three categories, mean values of the sum of correct answers for girls and for boys were calculated in each of the participating countries. The groups were then compared by conducting an ANOVA analysis. One prerequisite for conducting ANOVA is that there must be no difference between the values of least square Mean (LS Mean) and Mean. This prerequisite was achieved and the groups could be compared in an ANOVA analysis. The aim of the analysis was to determine whether there were significant gender differences within each country for each category of items and to come closer to the study's research question.

RESULTS

Categorization of Items

In the first step of data analysis, all items in the CBAS test were categorized into one or several curriculum emphases. From the analysis, 17 items were identified as belonging to the category *correct explanations*, because these items "only" required knowledge of scientific facts and products to provide

a correct answer (Roberts, 1982, 1998). Nine items were categorized as *scientific skill development*, because the items required understanding of scientific processes or experimental situations in order to solve the problems. Ten items were categorized as *everyday coping*, which imply the use of science knowledge in an everyday perspective. However, eight of these items were identified as also belonging in an additional curriculum emphasis. Table 2 below lists the number of items in each category.

Table 2. Number of items in each curriculum emphases.

Number of items	Every-day coping	Structure of science	Science, technology and decisions	Scientific skill development	Correct explanation	Self as explainer	Solid foundation
17	.	.	.	.	X	.	.
2	.	.	.	.	X	.	X
9	.	.	.	X	.	.	.
1	.	.	.	X	X	.	.
2	.	X	.	.	.	.	.
2	.	X	.	.	X	.	.
2	.	X	.	X	.	.	.
2	X	.	.	.	.	.	.
5	X	.	.	.	X	.	.
3	X	.	.	X	.	.	.

As Table 2 shows, the most common curriculum emphasis is *correct explanations*. One example of an item that was categorized into this emphasis is “Bicycle” (question 5). In the item, the students are shown a clip from a movie in which a boy is riding a bike and skids to a halt. The students are requested to consider what part of the bicycle is likely to have the highest temperature just as the bicycle comes to a stop, by choosing between four image alternatives. The alternatives are the hub, the rubber pads along the rim, the chain, or the part of the wheel that is in contact with the ground. In order to solve this item, the student only needs to know the correct explanation—that the friction between the wheel and the ground causes heat. The items belonging to this category focus on the correct answer and do not demand more extensive explanations or understandings of the scientific phenomenon (Roberts, 2011).

The next category, *scientific skill development*, comprises nine items. One item, categorized into this emphasis, is “Plant growth” (question 35) which is an interactive task and contains a plant growth simulation. The students are requested to find the most optimal cultivation conditions for the plants by varying the factors of soil acidity and temperature. To solve the problem, students need to vary one factor at a time and thus, approach the task from a classic, scientific experiment point of view. “Fish farm” (questions 36 and 37) and “Nuclear power” (question 13) are other examples of multiple choice items that were considered to belong to the category, *scientific skill development*. Simulation of a scientific investigation is thus an important feature in all these items.

The final emphasis, *everyday coping*, consists of ten items focusing on science as an important means for handling everyday situations. The focus in this emphasis is on ways in which it is possible to act in daily life by using knowledge from the disciplines. For example, this means knowing the following: how to manage practical situations involving electricity in the home, how to protect oneself from stroke of lightning, what substances could be poisonous, or interpreting the declaration of nutrients in food. One example of an item in this group is “Litter” (question 40) in which a person is about

to put a plastic bottle in the recycle bin. He presses the bottle together in order to minimize the size and then turns the cork on. The students are asked to explain why he turned the cork on afterward, by choosing between four alternatives.

The three emphases—*correct explanations*, *scientific skill development*, and *everyday coping*—are frequently represented among the items in the CBAS test. According to Table 2, the emphasis, *structure of science* could constitute an additional group for further analysis. However, two of the included items are already part of the emphasis, *scientific skill development*, which means that there are only four items left. This was considered a too-few-to-merit-further analysis. When it comes to the remaining emphases, *solid foundation*, *self as explainer*, or *science*, *technology and decision*, these were sparsely represented or were not present at all. From the perspective of Roberts' emphases, it is possible to argue that the CBAS test portrays a rather limited view of what science may be. We will however return to this issue in the discussion. Table 3 consists of a compilation of categorized items used in the further analysis.

Table 3. Items categorized into the different emphases.

Category	Names of items
Correct explanations	Chicken and egg 1, Bicycles 1, Bicycles 2, Ping-pong ball water, Ping-pong ball air, Tower crane 1, Tower crane 2, X-rays, Seaside steps, Bean leaves, Earthquake 2, Eagle, Echolocation 1, Fish farm 3, Litter 1, Washing glasses 1, Spinifex 1.
Scientific skill development	Chicken and egg 2, Bicycle 3, Nuclear power, Plant growth 1, Plant growth 2, Fish farm 1, Fish farm 2, Washing glasses 2
Everyday coping	Car safety, Assembly line 1, Vanishing krill 1, Vanishing krill 2, Earthquake 1, Drug test 1, Drug test 2, Litter 2.

Students' Performances Related to Curriculum Emphases

From the original data file, it was not possible to easily derive the number of correct answers on specific items for each student. In order to solve this problem, their answers were recoded and the value 1 was assigned for a correct answer and 0 for a wrong or missing answer. This made it possible to calculate a sum of correct answers and mean values on items related to the different curriculum emphases for each student through using an independent t-test, Kruskal-Wallis' test. This first step of the analysis revealed a quite evenly distributed result of correct answers, but the students performed slightly better on items categorized as belonging to *correct explanations* (53% correct answers) compared to *scientific skill development* (45% correct answers) and *everyday coping* (44% correct answers). However, our research question concerned possible differences in girls' and boys' achievements related to the scientific content and context in the test. Table 4 shows the mean value of correct answers for girls and boys within each category, regardless of country. Furthermore, Table 4 shows the calculated difference between the mean values.

The results show small differences between the mean values of girls' and boys' achievements in solving problems related to the category of *scientific skill development* (gender difference – 0.267) and *everyday coping* (gender difference – 0.311). This means that the analysis suggests only minor gender differences on items regarding the relation between science and everyday life. However, when it comes to the emphasis, *correct explanations*, significant gender differences exist: the boys outperform the girls (gender difference – 0.928). As mentioned, items categorized into this emphasis tend to focus on correct explanations without demanding deeper scientific reasoning, which could imply that these items do not necessarily acknowledge the context. We will return to this issue in the discussion section.

Table 4. Mean value of the sum of correct answers within each category but independent of country.

	Mean value within the category correct explanations (17 items)	Mean value within the category scientific skill development (9 items)	Mean value within the category everyday coping (10 items)
Girls	8,5369	3,9459	4,2749
Boys	9,465	4,2129	4,586
Difference	-0,928	-0,267	-0,311

Regarding the results on items related to the emphasis, *correct explanations*, gender differences exist. One possible conclusion is that these kinds of items disfavor girls. However, another possibility is that the gender differences could be related to differences in the school systems of the participating countries. Because the CBAS study comprises data from Denmark, Iceland, and South Korea, it was feasible to compare and analyze gender differences in these three countries. The next step was to determine whether there were significant gender differences within each country for each category of items. We used a one-way ANOVA analysis because the statistic mean values and least square means for the effect of interaction were equal for all categories included in this analysis. Therefore, it was possible to compare the mean values of the girls' and boys' achievements within each emphasis and in each country. Regarding the emphasis, *correct explanations*, the results revealed a significant difference between the achievements of Danish girls and boys ($p < 0.0001$). The same pattern becomes explicit for Icelandic ($p < 0.0001$) and Korean students ($p < 0.0001$). Accordingly, the analysis showed that the previously mentioned gender difference is valid for all three participating countries.

Regarding the emphases, *scientific skill development* and *everyday coping*, the image is more ambiguous. The gender differences are significant within these emphases when it comes to the results of Danish and Icelandic students. However, no corresponding gender differences exist for Korean students. From a Danish and Icelandic perspective, the results imply that there are significant differences, independent of the focus of the item. The significance levels regarding the emphasis, *everyday coping*, are not as strong for the results of the Icelandic students ($p = 0.0373$) compared to the Danish ($p < 0.0001$). Regarding these results, an important question is what significance the different item foci had in relation to the fact that the test was conducted on computers? The significance level of each comparison is presented in Table 5.

Table 5. The significance levels of least square means for the interaction effect of country and gender.

Comparison between the results of girls and boys in:	correct explanations	scientific skill development	everyday coping
Denmark	$< 0,0001$	$< 0,0001$	$< 0,0001$
Iceland	$< 0,0001$	0,0099	0,0373
Korea	$< 0,0001$	0,3349	0,2294

Results in Relation to What the Students State About the CBAS Test

When taking the CBAS test, the students answered four questions about how they perceived doing science tests on computers and about computer-based tests compared to paper-and-pencil-based tests. The results show minor gender differences, which could contribute to explain the differences on the knowledge test. The students were asked whether they enjoyed doing the computer-based science test. Icelandic and Korean boys strongly agreed with the statement "I enjoyed doing the computer-based test" more than the girls did in either country, and these results were statistically significant.

There was a similar, but smaller, difference among the Danish students, and that result was not statistically significant.

The students were also asked to consider how much effort they put into the PISA test compared to the CBAS test, and the results show a gender difference among the Danish and Icelandic students. To a higher extent than the boys, the girls stated that they spent more effort on the PISA test than on the CBAS test. On the other hand, the boys stated, to a higher extent than the girls, that they spent more effort on the CBAS test, and these differences were statistically significant. Both effort and enjoyment could thus principally provide possible explanations to the emerged gender differences on the CBAS test. In order to investigate the possible role of these two variables as confounding variables, we conducted a generalized linear analysis (ANOVA) where girls' and boys' performances on items related to the three categories *correct explanation*, *scientific skill development* and *everyday coping* was related to the variables effort and enjoyment. The results on the analyses showed however significant gender differences also after adjusting for both effort and enjoyment. Consequently, differences in girls' and boys' effort and enjoyment do not play a statistically significant role for the presented results.

DISCUSSION

This article originates from an interest in understanding the consequences of the increasing use of computer-based tests within the field of large-scale assessments of students' scientific literacy. The paper has focused on discussing how the tendency of the transition from paper-and-pencil-based tests toward computer-based tests may influence girls' and boys' achievements. The main aim was to explore in what ways students' achievements are affected by the content and the context of the CBAS items. Through comparing the results from these analyses to the results from earlier studies and the PISA main study, some significant patterns are evident. One such pattern is that the computer-based test seems to imply new or enhanced gender differences and biases, which is confirmed in part by others (OECD, 2010; Halldórsson et al., 2009; Sørensen & Andersen, 2009). But what do the gender differences consist of and what significance do the representation of scientific content and context of the items have for girls' and boys' performances on the computer-based test?

In order to come closer to this issue we have chosen to analyze the CBAS items through categorizing them into the framework of curriculum emphases (Roberts, 1982, 1998, 2011). Our intention has been to approach a richer description of how science can be represented and constituted, compared to the PISA framework of scientific competencies. We argue that Robert's curriculum emphases may provide a more nuanced and detailed portrayal of the complex concept of scientific literacy. The results of the analyses point to different gender patterns depending on the content and the context of the items. This means, for example, that boys in general tend to perform significantly better on items that are intended to measure knowledge of scientific facts and scientific products. The gender differences were less pronounced for items that aimed to evaluate students' knowledge of scientific processes or knowledge useful for handling everyday life. These findings are in line with other studies concluding that girls tend to depend more on having a relevant context for solving scientific problems than boys (e.g. Murphy & Whitelegg, 2006). A possible risk related to these results is that a transition from paper-and-pencil-based tests toward computer-based tests leads to a situation where items focusing on correct explanations seem to be over-represented, which evidently disfavors girls. An important conclusion for future scientific literacy tests is to consider the ways science could be represented in the tests and how these choices could have an impact on the results from a gender perspective.

Other studies within the area of science education suggest that additional factors must be considered from a gender perspective when constructing computer-based tests, such as the reading load (Halldórsson et al., 2009) and the representation of different science domains (Sørensen & Andersen, 2009). Another factor is the format of the items; LaFontaine and Monseur (2009) argue that girls

have an advantage in open-ended questions, whereas boys perform better on multiple choice questions. However, a more detailed image appears when interpreting the separate national results in this study. From a Danish perspective, significant gender differences appear to exist, independent of the test mode as well as the content and context of the CBAS items. On the other hand, according to the study of McKelvie, Halldórsson, and Björnsson (2008), the gender differences of the Danish students' results are enhanced in the CBAS test compared to the PISA main test. Regarding the Icelandic and the Korean students, the test mode seems to play a decisive role. No significant gender differences in general exist on the PISA main study, whereas the results from the CBAS study point to significant differences.

Therefore, it is possible to assume that when converting PISA surveys from paper-and-pencil-tests to computer-based tests, there arises an obvious risk of enhanced gender differences, which implies increased gender biases. This change of test modes could be explained from the fact that the computer per se accomplishes gender differences, because girls and boys experience different engagements in these situations (Volman et al., 2005; Colley & Comber, 2003). Furthermore, the analysis in this study shows that boys expressed a stronger engagement than girls when taking a computer-based test. There is another possible explanation. When conducting the test on computers, the items' context and content change considerably. For example, from the results of this study, a majority of the CBAS items aim to evaluate students' knowledge of scientific facts and products, which convey a specific image of science. Another example is that the explicit aim of reducing the reading load in the computer-based science test may reduce the contextualization of the scientific problem. These changes risk an overly narrow view of what scientific literacy may be, and also to a situation where girls and boys experience the tests differently (e.g. Murphy & Whitelegg, 2006). This means that the concept of scientific literacy is differently represented and constituted, depending on the test mode. Another consequence is the difficulty in reaching equivalence between the tests and thereby having limited possibilities for comparing the results. One could argue that it is not the computer per se that causes the emerged gender differences, but rather how science is conveyed and represented on a computer-based test.

On the basis of this article and the findings of the study, there is a need for further research on how to integrate different test modes for the same test and ways to explore how the respondents experience different examination tasks. This research would include ways to identify how girls, boys, high and low achievers, and students who take the test in a foreign language consider and approach the test items as well as what impacts previous computer experiences, such as playing games/doing simulations and chatting/communications. This could involve studying how the different test modes can be used and what the test modes could tell us about students' knowledge in and about science. Results from such tests could provide us with valuable knowledge, including how students approach the tests and in what ways the tests represent students' actual knowledge in and about science. Moreover, the results could increase the validity of future tests.

ACKNOWLEDGEMENTS

This research was funded by the Swedish Research Council (Dnr 721-2008-4717).

REFERENCES

- American Association for the Advancement of Science. (1990). *Science for all Americans*. New York: Oxford University Press.
- Bennett, R., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing complex problem-solving performances. *Assessment in Education*, 10, 347–360.
- Bridgeman, B. (2009). Experiences from large-scale computer-based testing in the USA. In F. Scheuermann & J. Björnsson (Eds.) *The transition to computer-based assessment and implications for large-scale testing* (pp. 39–44). Luxembourg: European Communities.
- Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2002). *Effects of screen size, screen resolution, and display rate on computer-based test performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Chua, S. L., Chen, D. T., & Wong, A. F. L. (1999). Computer anxiety and its correlates: a meta-analysis. *Computers in Human Behaviour*, 15(5), 609–623.
- Colley, A., & Comber, C. (2003). Age and gender differences in computer use and attitudes among secondary school students: What has changed? *Educational Research*, 45(2), 155–165.
- Dooley, P. (2008). Language testing and technology: problems of transition to a new era. *ReCALL*, 20(1), 21–34.
- Dyson, M. C., & Haselgrove, M. (2001). The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies*, 54, 585–612.
- Goldberg, A., & Pedulla, J. (2002). Performance differences according to test mode and computer familiarity on a practice graduate record exam. *Educational and Psychological Measurement*, 62(6), 1053–1067.
- Halldórsson, A., McKelvie, P., & Björnsson, J. (2009). Are Icelandic boys really better on computerized tests than conventional ones? In F. Scheuermann & J. Björnsson (Eds.) *The transition to computer-based assessment—New approaches to skill assessment and implications for large scale testing* (pp. 178–193). Luxembourg: European Communities.
- Hendley, D., Stables, S., & Stables, A. (1996). Pupils' subject preferences at Key Stage 3 in South Wales. *Educational Studies*, 22, 177–187.
- Horky, N., Bennett, R., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NEAP. *The Journal of Technology, Learning, and Assessment*, 5(2), 1–50.
- IEA. (1997). *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Report*. Boston: TIMSS and PIRLS International Study Center.
- IEA. (2008). *TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Boston: TIMSS and PIRLS International Study Center.
- Jakobsson, A., Säljö, R., & Mäkelä, Å. (2009). Conceptions of knowledge in research on students. Understanding of the Greenhouse Effect: Methodological positions and their consequences for representations of knowing. *Science Education*, 93(6) 978–995.
- Jones, G., Howe, A. & Rua, M. (2000). Gender differences in students' experiences, interests, and attitudes towards science and scientists. *Science Education*, 84, 180–192.
- LaFontain, D., & Monseur, C. (2009). Gender gap in comparative studies of reading comprehension: to what extent do the test characteristics make a difference? *European Educational Research Journal*, 8(1), 69–79.
- Lau, Kwok Chi. (2009). A critical examination of PISA's assessment on scientific literacy. *International Journal of Science and Mathematics Education*, 7, 1061–2088.
- Lee, M-K. (2009). CBAS in Korea: Experiences, results and challenges. In F. Scheuermann & J. Björnsson (Eds.) *The transition to computer-based assessment—New approaches to skill assessment and implications for large-scale testing* (pp. 187–193). Luxembourg: European Communities.
- Leeson, H. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, 6(1), 1–24.

- Martin, R. (2009). Utilising the potential of computer-delivered surveys in assessing scientific literacy. In F. Scheuermann & J. Björnsson (Eds.) *The transition to computer-based assessment—New approaches to skill assessment and implications for large-scale testing* (pp. 172–177). Luxembourg: European Communities.
- McDonlad, A. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers and Education*, 39, 299–312.
- McKelvie, P., Halldórsson, A. M., & Björnsson, J. K. (2008). *PISA CBAS analysis and results—Science performance on paper-and-pencil and electronic tests*. Reykjavik: Educational Testing Institute.
- Murphy, P., & Whitelegg, E. (2006). *Girls in the physics classroom: a review of the research on the participation of girls in physics*. London: The Open University.
- OECD. (2001). *Knowledge and skills for life. First results from the OECD program for international student assessment 2000*. OECD Publishing.
- OECD. (2004). *Problem solving for tomorrow's world. First measures of cross-curricular competencies from PISA 2003*. OECD Publishing.
- OECD. (2007). *Assessing scientific, reading and mathematical literacy: a framework for PISA 2006*. OECD Publishing.
- OECD. (2009). *Equally prepared for life? How 15-year-old boys and girls perform in school*. OECD Publishing.
- OECD. (2010). *PISA computer-based assessment of student skills in science*. OECD Publishing.
- Osborne, J., & Collins, S. (2001). Pupils' views of the role and value of the science curriculum: a focus group study. *International Journal of Science Education*, 23(5), 441–467.
- Osborne, J., Simons, S., & Collins, S. (2003). Attitudes towards science: a review of the literature and its implications. *International Journal of Science Education*, 25(9), 1049–1079.
- Revuelta, J., Ximénez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement*, 63, 791–808.
- Roberts, D. A. (1982). Developing the concept of “Curriculum emphases” in science education. *Science Education*, 66(2), 243–260.
- Roberts, D. A. (1998). Analysing school science courses: The concept of companion meaning. In D. Roberts and L. Östman (Eds.) *Problems of meaning in science curriculum* (pp. 5–12). Columbia University, New York: Teachers College Press.
- Roberts, D. A. (2011). Competing visions of scientific literacy: The influence of science curriculum policy image. In C. Linder et al. (Eds.) *Exploring the landscape of scientific literacy* (pp. 11–27). New York: Routledge.
- Sjøberg, S. (2000). Interesting all children in ‘science for all’. In R. Miller, J. Leach, & J. Osborne (Eds.) *Improving science education* (pp. 165–186). Buckingham: Open University Press.
- Sørensen, H., & Andersen, A.M. (2009). How did Danish students solve the PISA CBAS items? Right and wrong answers from a gender perspective. In F. Scheuermann & J. Björnsson (Eds.) *The transition to computer-based assessment—New approaches to skill assessment and implications for large-scale testing* (pp. 194–201). European Communities.
- Turmo, A., & Lie, S. (2006, November). *PISA's computer-based assessment in science (CBAS)—a gender equity perspective*. Paper presented at the AEA-E Annual Conference 2006 Assessment and Equity. Napoly, Italy.
- Warrington, M., & Younger, M. (2000). The other side of the gender gap. *Gender and Education*, 12(4), 493–508.
- Volman, M., Eck, E., Heemskerk, I., & Kuiper, E. (2005). New technologies, new differences. Gender and ethnic differences in pupils' use of ICT in primary and secondary education. *Computers and Education*, 45, 35–55.