Chapter 2

2

Administration of the CBAS test

The computer-based assessment of science was field tested in thirteen PISA countries in 2005 and the main study was carried out in three of them in 2006, Denmark, Iceland and Korea. Rigorous sampling was used in all countries and all students also took the regular PISA test. The computer-based test had appropriate test characteristics and sampled a number of different aspects of science knowledge and skills.

FIELD TRIAL

CBAS was administered in the PISA 2006 Field Trial in 13 PISA-participating countries. Preliminary analyses were performed on the data from 12 of these countries to determine the final item selection and final method of test administration for the Main Study which was undertaken later in 2006. Only three countries (Denmark, Iceland and Korea) continued through to the Main Study phase of CBAS and the data from these countries are the focus of this report.

SCHOOL AND STUDENT SAMPLING

A subsample of 100 schools was selected to participate in CBAS from the main PISA 2006 school sample test in Iceland. From these schools, clusters of 5 to 45 PISA-eligible students were sampled from the PISA student sample. All schools and students selected for CBAS had already participated in the paper-and-pencil PISA 2006 assessment.

It is important to note that the sample considered in the present analyses includes all students that participated in the CBAS test session as well as all PISA-participating students from the schools that had at least one student participating in CBAS. For Iceland, the original CBAS sample was drawn with 1 104 students out of which 784 students participated (71% response rate). However these analyses include data for an additional 2 782 students who participated in the paper-and-pencil assessment of science, attended a CBAS-participating school but did not respond to the CBAS test. To give achievement scores on the CBAS test for these students, plausible values on the CBAS scale were statistically imputed based on the students' PISA paper-and-pencil achievement and background information. A total of 3 566 students are therefore included in the CBAS analyses for Iceland, which is very close to the total number of students participating in the paper-and-pencil PISA 2006 (3 789). As a result, it is fairly certain that the Icelandic sample for CBAS is representative of the population of 15-year-old students in the country.

In Denmark and Korea, approximately half of the PISA sampled schools were sampled for CBAS and, again, all students participating in CBAS or PISA from these schools were included in the final data file. The final numbers of included students were: 837 CBAS-responding students out of 1 254 students included in the CBAS analyses for Denmark and 1 475 CBAS-responding students out of 2 650 students included in the analyses for Korea.

Sample weighting

To account for any biases in selection of schools and students, the PISA data are weighted using a balanced repeated replication method. This accounts for, for example, any over- or under- representation of geographical areas within countries. More information about the weighting techniques in PISA can be found in the Data Analysis Manual (OECD, 2005a) or in the PISA 2006 Technical Report (OECD, 2009).

Gender distribution in the weighted sample

In Denmark and Iceland the CBAS sample was approximately equally constituted of males and females. Table 2 shows however, that there was a greater number of males than females in Korea. The greater number of males for Korea is of concern given that Field Trial results indicate a strong gender difference in achievement in favour of males. This should be kept in mind when interpreting cross-country achievement comparisons.

Table 2.
Proportion of females and males in the sample analysed in this report

	CBAS sample	
Country	Females	Male
Denmark	52%	48%
Iceland	50%	50%
Korea	44%	56%

ADMINISTRATION OF TEST SESSIONS

CBAS sessions took place either on the same day as the PISA paper-and-pencil assessment of students' reading, mathematics and science performance, or very shortly thereafter. Test administration was standardised so that all students performed the test on the same type of laptop, using the same software and in a similar testing environment.

Up to five students participated in each test session under the guidance of one Test Administrator. The computer-based science items were presented to students on laptop computers through CBAS software specially designed for this purpose. This was a fixed-form test where the same 45 items were presented to all students in one of either two orders. The order of items was split from the middle point of the second form so as to reduce fatigue effects on the items occurring later in the test.

The software allowed students to move between items as they wished and to return to questions (changing their answers if necessary) up until 55 minutes had elapsed since the beginning of the test, at which point the Test Administrator stopped the session. This allowed for just over one minute per question. If a student finished early the items remained on the screen until the completion of the 55 minute test session.

Following the cognitive items, four questionnaire items were presented and students had five minutes to respond to these. In total therefore, test sessions were one hour long.

Technical specifications

All laptops used for student testing were required to comply with a number of minimum specifications as given in Table 3 below.

Table 3.
Minimum hardware specifications for student laptops used for CBAS

Component	Minimum required
Central processing unit	1.6 GHz Intel Pentium® M Processor
Memory	512 MB of RAM
Hard disk	40 GB
Display	14.1" XGA (this exact screen size must be used)
Network	Must be equipped with wireless capability (Centrino technology) – IEEE 802.11 b/g
Optical drive	Computers must be equipped with a DVD-ROM drive for installing the software required for test administration. Also, the test administrator's computer must be equipped with a CD-RW drive for writing test session result data to CDs.
Operating System	Windows® XP Professional
Pointing device	Symmetrical external, optical mouse
Listening device	External stereo headphones

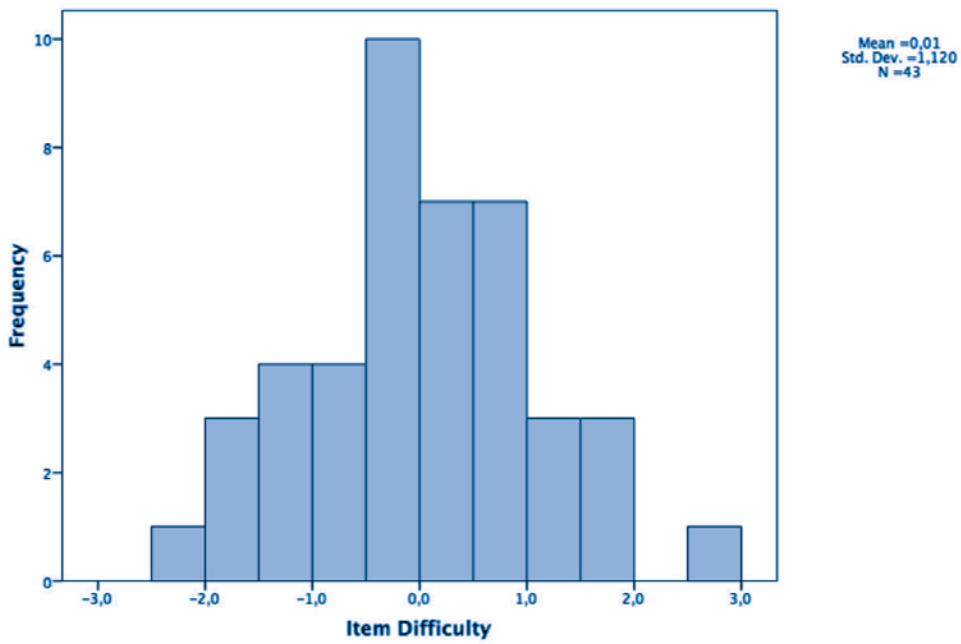
Cognitive item information

In total, 45 items with multimedia extensions (animations using flash software, video footage or photos) were presented to students. The final analyses are performed on 42 items as two items (Spinifex Q2 and Assembly Line) were dropped prior to the analyses and two items were combined into one (Bicycle Q1 and Q2) as they were considered to be assessing the same knowledge. Two additional items were set to 'not administered' for Icelandic students (Tablet Fizz Q2 which shows video footage of a vitamin tablet dissolving in water which was judged as an unfamiliar concept for Icelandic students and Fish Farm Q2 because one of the specific terms in the item was not translatable into Icelandic). All item designs were either multiple choice or complex multiple choice involving, for example, a number of Yes/No responses for the answers offered. A small number of complex multiple choice items asked the students to place items in a specific order or position in a given diagram.

Item order and missing responses

The two items that were removed from the database prior to the analyses for all countries may have been affected by their late presentation in both forms: one was the final question on Form 1 and the 23rd item on Form 2 (and had the highest combined order rank score) and the other was the second to last question on Form 2 (40th out of 45 in order rank). This indicates that fatigue may have played a substantial role in the way students responded to these items which eventually saw them deleted. Missing responses also increased towards the end of the test, however, overall the proportion of missing responses was very low for all countries (less than 1%). Further examination of gender differences in missing responses will be presented in a later section of this report.

Figure 5.
Distribution of item difficulty for final CBAS items



Item difficulty

Item difficulties were calculated and Figure 5 above shows that the items were approximately evenly distributed across the item difficulty scale from -3 to 3 with the mean item difficulty at zero, indicating good coverage of all potential competency levels. More information about the calculation of PISA item difficulties can be found in the 2006 PISA Technical Report (OECD, forthcoming). Percentage correct per CBAS item was also calculated and ranged from 13% to 94% with an average percentage correct per item of 60%.[1] Percentage correct per item was strongly associated with item difficulty from the model at 0.90 indicating that percentage correct per item is also an adequate measure of performance for specific analysis purposes.

Item difficulty across genders

After the Field Trial, to determine the final selection of Main Study items, item difficulties were calculated per gender in a gender differential item functioning (DIF) analysis. Two of the items identified as easier for females were maintained in the Main Study (Ping Pong Ball Q1 and Washing Glasses Q3) and three items that were identified as easier for males were maintained in the Main Study (Bicycle Q2 and Q3 and Designed for Flight). Performance on these items will be examined in closer detail in the results section of this report.

Science framework classification

In the same way that paper-and-pencil science items were classified into domains of science, CBAS items were categorised according to one of the three scientific competencies of PISA (Explaining Phenomena Scientifically, Identifying Scientific Issues or Using Scientific Evidence) and according to what sort of knowledge was being assessed. They were either considered to be assessing:

1. Knowledge of Science (*i.e.* knowledge about the natural world, about the topics of science) and one of the sub domains of knowledge: Physical systems, Living systems, Earth and Space systems or Science and Technology, or,

2. Knowledge about Science (*i.e.* knowledge about science itself and the methodologies of science) and one of the sub-categories: Scientific Enquiry or Scientific Explanations.

The final CBAS items selected for the Main Study were relatively evenly distributed across the types of scientific knowledge and sub domains of science and were similar in distribution to the paper-and-pencil science items. Slight differences were found in the proportions of items assessing the *explaining phenomena scientifically, identifying scientific issues* and *using scientific evidence* between tests in that the CBAS test had more items assessing (presumably because the computer medium facilitated items involving scientific trials and repetition more so than did the paper-and-pencil medium). In terms of domains, the CBAS test had slightly more technology items and items involving the physical sciences whereas the paper-and-pencil test had more items assessing living systems. However, as the distributions across categories in Table 4 show, these differences are relatively slight. Overall, it seems that the paper-and-pencil test and the computer-based test were assessing similar areas of scientific competencies and knowledge.

Table 4.
Distribution of paper-and-pencil science and CBAS items across domains of science

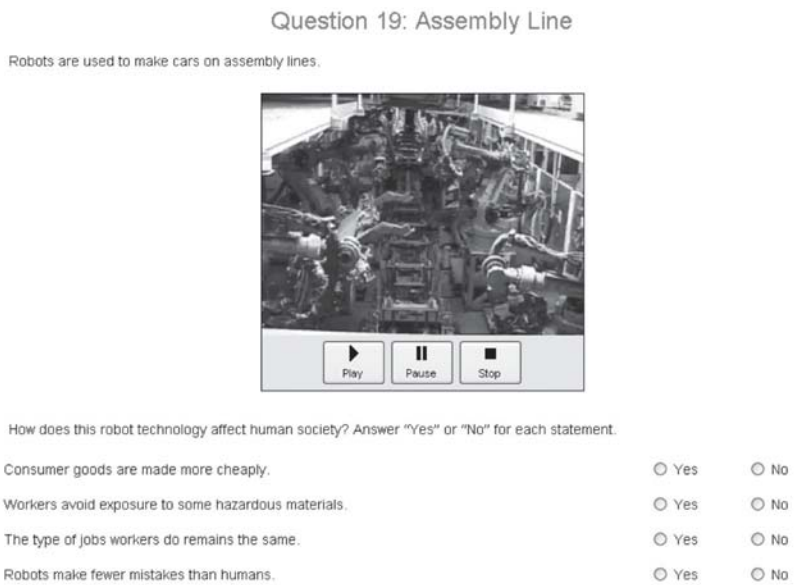
			Paper-and-pencil Science	CBAS
Scientific Competency		Explaining Phenomena Scientifically	49%	40%
		Identifying Scientific Issues	22%	24%
		Using Scientific Evidence	29%	36%
Scientific Domain	Knowledge of Science	Earth and Space Systems	11%	12%
		Living Systems	23%	17%
		Physical Systems	16%	21%
		Technology Systems	7%	10%
	Knowledge about Science	Scientific Enquiry	23%	19%
		Scientific Explanation	19%	21%

Interactivity

As the computer-based items differ markedly from the paper-and-pencil items in terms of how much the student can interact with the item (for example, the possibility of moving levers to adjust levels in experimental trials or dragging and dropping the answer into the correct location in the diagram), it may be useful to consider the effects of the interactivity of the items on performance, in particular across genders.

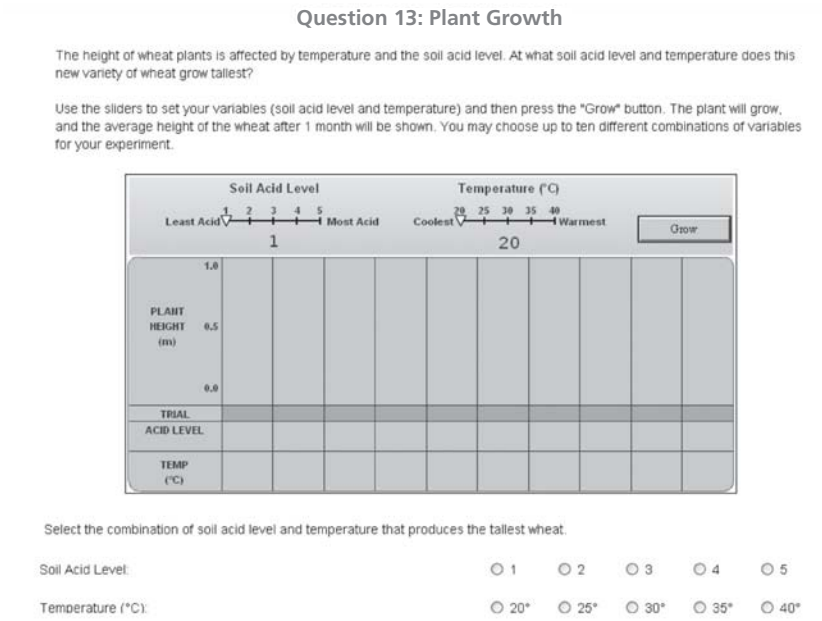
A panel of three independent judges rated all CBAS items into three groups according to the level of interactivity (low, medium and high) based on the types of activities the student had to perform with the item and based on how much the student needed to engage with the audiovisual material to answer the question. An example of a low interactivity item is the "Assembly Line" item in Figure 6 that shows a short video of an automated car assembly line and asks a question related to the role of robots in society.

Figure 6.
Sample unit: Assembly Line



Here, the video footage serves as contextual information to the item but does not provide the answer. In fact, this question could be answered correctly without the student watching the video footage and is therefore considered to be of low interactivity. In contrast, the following item in Figure 7, "Plant Growth", where the student is required to move buttons up and down a scale, performing experimental trials on optimal temperature and soil acidity levels for growing wheat, was considered as highly interactive.

Figure 7.
Sample unit of highly interactive item: Plant Growth



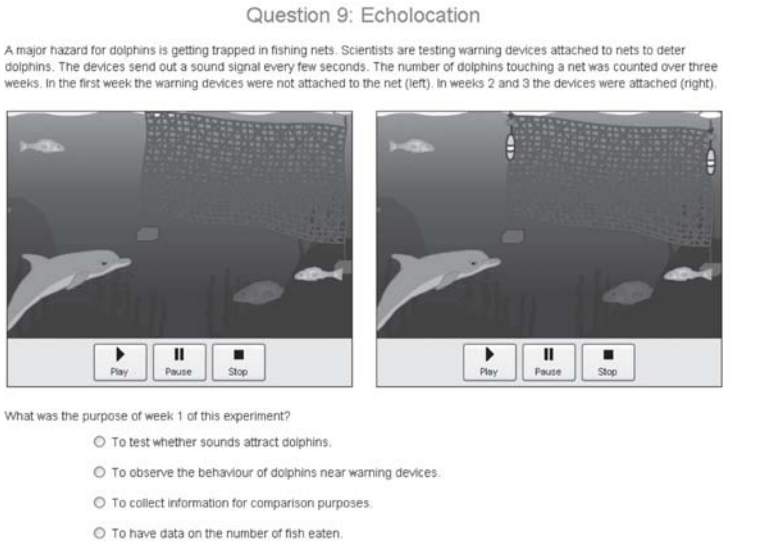
Overall, 14 items were classified as high interactivity, 13 as medium and 16 as involving low interactivity. Achievement on these items will be compared in the results section.

Reading load

Word counts for each CBAS item were recorded according to the number of words embedded in the image of each stimulus in both the question stem and multiple choice response options. Based on these figures, the CBAS items were divided into three groups according to reading load: low, medium and high. Eleven items were considered to be of a high reading load, for example as shown in Figure 8.

Figure 8.

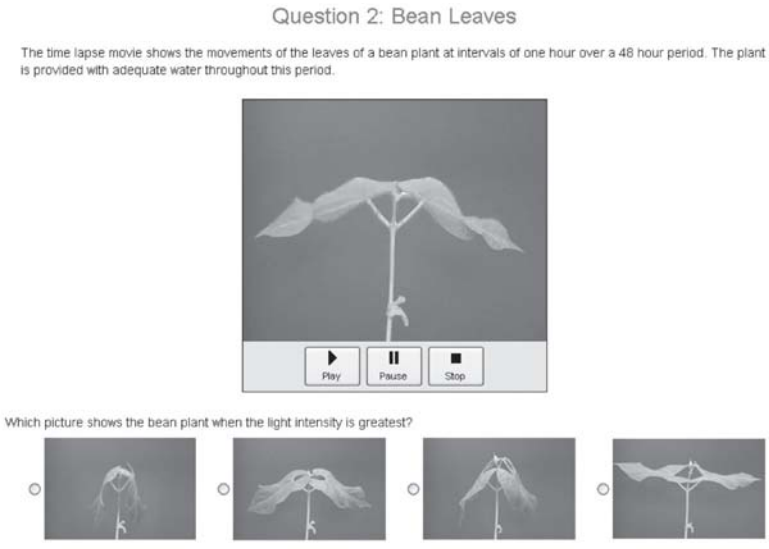
Sample item showing high reading load item: Echolocation



Fourteen items were classified as medium reading load and eighteen items (including the item in Figure 9 below) were classified as low reading load.

Figure 9.

Sample item showing low reading load item: Bean Leaves



CBAS questionnaire items

After the cognitive items the students were asked to respond to several short questions to investigate the effects of enjoyment, motivation and effort on performance. Students were asked to use a four-point Likert scale to rate how much they enjoyed the computer-based and paper-and-pencil tests, and whether they would do a similar test where the answers were provided "just for fun" (assessing motivation). The PISA Effort Thermometer was also used where students were asked to imagine an actual situation that was highly important to them personally, so that they would try their very best and put as much effort as they could to do well. They were told that in this situation they would mark the highest value on the effort thermometer (10) and then they were asked to report: how much effort they put into doing the CBAS test compared to the situation they had just imagined; and how much effort they would have invested if their marks from CBAS had counted in their school marks. This questionnaire item (identical to the item used in the PISA paper-and-pencil test) is displayed below in Figure 10.

Figure 10.
PISA Effort Thermometer

How much effort did you invest?

Please try to imagine an actual situation (at school or in some other context) that is highly important to you personally, so that you would try your very best and put in as much effort as you could to do well.

In this situation you would mark the highest value on the "effort thermometer", as shown below:	Compared to the situation you have just imagined, how much effort did you put into doing this test?	How much effort would you have invested if your marks from the test were going to be counted in your school marks?
☐ 10	☐ 10	☐ 10
☐ 9	☐ 9	☐ 9
☐ 8	☐ 8	☐ 8
☐ 7	☐ 7	☐ 7
☐ 6	☐ 6	☐ 6
☐ 5	☐ 5	☐ 5
☐ 4	☐ 4	☐ 4
☐ 3	☐ 3	☐ 3
☐ 2	☐ 2	☐ 2
☐ 1	☐ 1	☐ 1

In addition, students were asked which test they put more effort into between the CBAS test and the PISA paper test (assessing relative effort) and they were finally asked what type of test they would prefer between a two hour paper-and-pencil test, one hour of each type of test and two hours of computer-based testing.

Notes

1. Percentage correct for all PISA Science items was calculated across the three CBAS participating countries, ranging from 9% to 91% with an average percentage correct across items at 46% which is lower than that for CBAS. Overall, percentage correct per paper-and-pencil science item was correlated with the item difficulties at 0.992 indicating that percentage correct is also an adequate indication of performance.