ANNEX A.2
PISA 2012 COMPUTER-BASED ASSESSMENT LOG FILES

Chapters 4, 6 and 7 make use of variables derived from the log files for the PISA 2012 assessment of digital reading. The log files for all digital reading items share a common structure:

- The first four variables uniquely identify students (3-letter country code, 6-digits national centre code, school identifier, student identifier), and can be used to link these data to the public-use PISA files available on the website of the Australian Council for Educational Research: http://pisa2012.acer.edu.au/downloadsCBA.php.

- The fifth variable identifies the test form; the same information is also contained in the main PISA student database.

- For every student, a sixth variable (event_number) uniquely identifies the observation.

- Variable "time" corresponds to the time when the event was written to the log file. This is measured in seconds, and initialised to "0" at the start of the test session.

- Variables "event", "event_value" and "event_detail", together, can be used to describe the activity of students or the system during the test. The raw logfiles were cleaned so that the only events included are those that are needed to reconstruct the sequence of what students do to solve a task. In digital reading files, "event" can take only six values: "START_ITEM", "END_ITEM", "click" or "dblclick" (for clicks/double clicks on the browser space or answer space), "change" and "focus". For "change" events, "event_value" specifies which answer option is selected after a valid "click" in the answer space. For "click" and "dblclick" events in the browser space, "event_detail" contains information on the page clicked (the origin page) and on the link selected, which is usually enough to identify the destination page. "Focus" events are only included for a few items where the value of event_detail in "click"/"dblclick" events does not uniquely identify the destination page; they contain, in general, information about the page displayed after "click"/"dblclick" event.

For released items, log files that correspond to the above description are available for download on the OECD "PISA products" website: http://www.oecd.org/pisa/pisaproducts/

Cleaning operations

These raw files were treated to extract the navigation sequence and derive navigation indices in the following steps:

1. Each student's log was sorted by time and event_number. In a few cases where two logs existed for the same student (two "START_ITEM" events), only one log –the one containing the longest sequence of events between "START_ITEM" and "END_ITEM"– was kept. In

general, the log that was dropped did not contain any information at all about students' behaviour (i.e. only system-generated events). Events recorded before the start of the selected sequence, or after the end of the sequence, were dropped.

2. The "origin" and "destination" page for each "click" or "dblclick" event was extracted by recoding information in the event_value and event_detail variables. In items with "focus" events, event_value was similarly recoded. This step ensures that each page is assigned a unique code (see example in Figure 7.8).

3. The navigation sequence was extracted (in a variable called "navigation") from the sequence of origin, destination and (if applicable) focus pages; in most cases, the destination page coincided with the origin page for the next "click" or "dblclick" event. In such cases, to avoid counting two page visits when there was only one, the navigation variable coincided with the origin page. However in some cases, particularly when navigation ended, the destination page had to be added to the sequence, and this was done by expanding the log with a new observation (carrying the same identifiers and time stamp as the previous one) and writing the destination page code as the value of the "navigation" variable for this new observation. Figure A.2 shows an example of a navigation sequence extracted from the log files for Task 2 in unit *SERAING* (see Chapter 7).

Figure A.2   Extracting the navigation sequence from digital reading log files

Original log file information:

event	time	number	event_value	event_detail
START_ITEM	1107.3	1	NULL	NULL
focus	1119.9	2	noID	seraing_home.php?tok=&link=275
click	1120.0	3	noID	seraing_home.php?tok=&link=275
focus	1139.1	4	noID	ccc_home.php?tok=&link=217
click	1139.2	5	noID	ccc_home.php?tok=&link=217
click	1144.3	6	noID	ccc_programme_by_date.php?tok=&link=16
focus	1144.4	7	noID	ccc_programme_by_date_november.php?tok=&link=72
focus	1150.1	8	noID	item2.php
click	1150.2	9	era002q2o4	item2.php
change	1150.2	10	era002q2o4	NULL
focus	1152.9	11	noID	[…]
click	1153.0	12	NextButtonPForm	[…]
END_ITEM	1153.9	13	NULL	NULL
click	1153.9	14	noID	[…]

Extracted navigation behaviour:

event	time	navigation	answer
START_ITEM	1107.3		
click	1120.0	E002P01	
click	1139.2	E002P03	
click	1144.3	E002P04	
focus	1144.4	E002P06	
change	1150.2		era002q2o4
END_ITEM	1153.9		

The navigation sequence was further treated to identify relevant and non-relevant pages, and construct navigation indices, as described in Chapter 4 and Chapter 7. A database containing the navigation measures used in Chapter 4 and 6 is available for download at http://www.oecd.org/pisa/pisaproducts/.

Sample for log-file data analyses

Log files for digital reading items are available for all students who participated in the PISA computer-based assessment and were assigned to a test form containing digital reading tasks. They are not available for students whose scores in digital reading were imputed.

The total number of students included in analyses based on digital reading log files is 58 823; this corresponds to 93.4% of all students included in the PISA sample who were assigned a test form containing digital reading tasks (the latter being a representative sample of 15-year-old students in each country). The missing students, it can be assumed, are mostly those who did not attend the computer-based test session or were prevented by a technical failure from taking the computer-based test.

For analyses based on single items or units as in Chapter 7, the sample size is typically only about half as large, as any individual item was included only in 6 of the 12 test forms containing digital reading tasks.