OECD Skills Studies

Beyond Proficiency

USING LOG FILES TO UNDERSTAND RESPONDENT BEHAVIOUR IN THE SURVEY OF ADULT SKILLS

OECD

OECD Skills Studies

Beyond Proficiency

USING LOG FILES TO UNDERSTAND RESPONDENT BEHAVIOUR IN THE SURVEY OF ADULT SKILLS

OECD
BETTER POLICIES FOR BETTER LIVES

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document, as well as any data and any map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Please cite this publication as:
OECD (2019), *Beyond Proficiency: Using Log Files to Understand Respondent Behaviour in the Survey of Adult Skills*, OECD Skills Studies, OECD Publishing, Paris.
https://doi.org/10.1787/0b1414ed-en

ISBN 978-92-64-59082-3 (print)
ISBN 978-92-64-80280-3 (pdf)

OECD Skills Studies
ISSN 2307-8723 (print)
ISSN 2307-8731 (online)

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Photo credits:
© aleksandr-mansurov-ru/iStockphoto
© Don Pablo/Shutterstock
© Jamie Grill/Getty Images
© Jaroslav Machacek/Shutterstock
© Konstantin Chagin/Shutterstock
© Lightspring/Shutterstock
© momentimages/Tetra Images/Inmagine LTD
© Ocean/Corbis

Corrigenda to OECD publications may be found on line at: *www.oecd.org/about/publishing/corrigenda.htm*.

Foreword

Digital technologies are revolutionising many aspects of our everyday life, and skills assessments are no exception. The Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC) (hereafter referred to as "PIAAC") was the first large-scale international assessment fully designed to be primarily delivered on a computer.

This choice of computer-based delivery was motivated by several considerations. First, an assessment of information-processing skills in the 21st century needed to test adults' capacity to access and interpret information in digital formats. Second, assessment tasks delivered by computer can be highly interactive. This makes it possible to refine the measurement of traditional information-processing skills, as well as to develop measures of innovative testing domains, such as problem-solving in technology-rich environments or adaptive problem-solving. Finally, computer delivery provides other benefits, such as increased efficiency and improvements in data quality (e.g. automatic scoring of answers, lower loss of data and more complex test designs and in the management of survey administration).

Another important consequence of computer-based delivery is that the testing platform can record information about all the interactions between test takers and the computer. This information is stored in log-files and is also known as "process data".

By providing a way to observe how test-takers approach and try to solve the tasks presented to them, process data have the potential to substantially enrich the information we get and, therefore, the lessons we learn from skills assessments. Process data can be used to further refine the measurement of skills traditionally assessed and to enlarge the set of indicators we obtain from assessments. They can be used to proxy unobservable traits, such as motivation and perseverance, to better understand the relationship between these attitudes and performance, and also to better interpret and contextualise the results of large-scale assessments and the differences that we normally observe across countries or socio-demographic groups.

This report offers a roadmap for readers interested in knowing more about process data and how they can be used for research purposes and to inform policy making. It describes currently available process data from PIAAC and provides examples of the analysis that can be undertaken with them.

At the same time, the report acknowledges that research on process data is still in its infancy. For the moment, log files are largely an unintended by-product of computer-based administration. Not all information that we would like to have in them has been recorded, and the available information is often cumbersome to extract and, more importantly, to interpret.

But the path before us is now clearly traced. The analysis of process data will increasingly inform the process of test development through a better understanding of the strategies and behaviour of test takers. Process data will increasingly be used to design better assessments. In an iterative process, assessments will be increasingly designed to better exploit the fact that we now have the tools to observe not only whether or not test takers are able to solve a task presented to them, but also how they arrived at the solution and where they went right and where they went wrong.

If research on process data fulfils its promise, large-scale assessments will no longer only be used as a tool to describe where OECD countries stand in terms of the skills of their adult and student populations, but also as a tool that will teach them how they can improve.

Andreas Schleicher,
Director for Education and Skills and Special Advisor on Education Policy to the Secretary-General,
OECD

Acknowledgements

Development of this report was guided by Andreas Schleicher and overseen by the PIAAC Board of Participating Countries. The report was led by William Thorn and drafted by Marco Paccagnella and François Keslair, with statistical analysis by François Keslair. Susan Copeland edited the report, and Jennifer Cannon and Sophie Limoges co-ordinated its production.

Table of contents

Tables

Figures

Follow OECD Publications on:

- *http://twitter.com/OECD_Pubs*
- *http://www.facebook.com/OECDPublications*
- *http://www.linkedin.com/groups/OECD-Publications-4645871*
- *http://www.youtube.com/oecdilibrary*
- OECD Alerts *http://www.oecd.org/oecddirect/*

This book has...

StatLinks

A service that delivers Excel® files from the printed page!

Look for the *StatLinks* at the bottom of the tables or graphs in this book. To download the matching Excel® spreadsheet, just type the link into your Internet browser, starting with the *http://dx.doi.org* prefix, or click on the link from the e-book edition.

Executive summary

Computer-based administration of large-scale assessments makes it possible to collect a much richer set of information on test takers than pencil and paper tests. In principle, it is possible to record all interactions between the computer user interface on which the test is taken and a server.

This information about the actions undertaken in the course of the assessment can help policy makers, researchers and educators to better understand the cognitive strategies used by respondents and the underlying causes of low and high performance, and thus to design appropriate interventions.

The information contained in log files (also denoted as process data) can also be used to investigate aspects of respondents' ability, attitudes and behaviour, over and above the cognitive constructs that test items are designed to measure. For example, timing information can be used as proxies of test-takers' motivation, engagement and perseverance. As performance in a test is always the combined outcome of the ability of the respondent and the effort exerted in the course of the assessment, information on the motivation and engagement of respondents is essential for interpreting differences in observed performance, especially when respondents do not have any stakes in the assessment.

The analysis and interpretation of process data are, however, not straightforward. As log files are records of the interaction between respondents and items, interpretation of the information contained in log files is necessarily item-dependent. Moreover, existing log files contain only a subset of the respondent-computer interactions, and the choice of which information to record was usually not informed by considerations about the usefulness of the data for subsequent analysis. Finally, many of the actions that a respondent undertakes while solving an assessment item cannot be recorded in log files.

This report, based on data from the Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC), focuses on the analysis of timing indicators. These have the advantage of being available for all items and being easier to interpret in a consistent way across different items. The analysis concentrates on three indicators: 1) time on task (the total time spent on an item by the respondent); 2) time to first interaction (the time elapsed between the moment the item is presented to the respondent and the moment at which he/she first interacts with the testing platform); and 3) time since last action (the time elapsed between the respondent's last interaction with the platform and the moment at which he/she moves on to the next item). The analysis is limited to the domains of literacy and numeracy because it is only in these domains that timing indicators can be safely generalised across multiple items. The domain of Problem Solving in Technology-Rich Environments provides richer information, because the items it contains are much more interactive, but interpretation of that information becomes then largely dependent on the content and context of specific items.

A first important finding of this report relates to the unexpectedly large cross-country differences in the amount of time respondents spent on the PIAAC assessment. Overall time spent on the assessment is positively correlated with average performance, and negatively correlated with the incidence of missing answers.

Large differences are also found at the individual level. Time spent on the assessment tends to increase with the age and the education level of respondents, despite the fact that older individuals also display a higher propensity to skip items. Gender differences are small. Respondents reporting greater familiarity with information and communications technology (ICT) tend to complete the assessment more rapidly, but the difference disappears after controlling for other observable characteristics. Familiarity with ICT is also associated with a shorter time to first interaction and a longer time since last action. Nonetheless, large differences persist between individuals with similar socio-demographic characteristics.

The time spent on different items is closely related to intrinsic item characteristics, most notably item difficulty. Respondents devoted a significantly smaller amount of time to items administered in the second half of the assessment. This was accompanied by an increase in the proportion of missing answers and a decrease in performance. Respondents tend to spend the most time on items that are challenging but feasible, while spending little time on items that, in relation to their estimated proficiency, are either very easy or very difficult.

Timing information can be used to construct indicators of disengagement. Respondents are considered as disengaged with an item if they spend too little time on it. In such situations, it can be assumed that the respondent has not even devoted the effort necessary to understand the item and has skipped it without even evaluating his/her chances of answering the item correctly. The incidence of disengagement varies substantially across countries. Disengagement is more likely to be observed in items presented in the second half of the assessment, consistent with the analysis of time allocation to different items. Adults with low levels of education and adults who are less familiar with ICT are more likely to become disengaged in the course of the assessment.

Research using log files is still in its infancy. PIAAC was the first large-scale international assessment delivered primarily on computers. The information available from PIAAC has been used in a number of analyses. It has contributed to the understanding of what can be drawn from this type of data, as well as aided in the exploration of substantive issues, such as test-engagement and respondents' cognitive strategies.

By capitalising on the lessons learned from these data and the results of this report, future large-scale assessments will likely be able to improve their design and maximise the research potential of log files. The information contained in current log files will be useful in improving the design of new items. Test developers will strive to design interactive items that will enrich the content of future log files, greatly enhancing their analytical potential. It will be particularly important to prespecify theoretical constructs or competing theoretical hypotheses that could be measured or tested using the information recorded in log files. It should also be made clearer whether the purpose of log files is to better measure the underlying cognitive constructs, or whether they can be used to proxy for other dimensions of respondents' skills, such as personality traits or attitudes.

Large-scale assessments have been often criticised for not taking into account the effort and motivation of test takers, and for being silent about policy actions that can help improve individual skills. Log files carry the analytical promises to improve large-scale assessments on both dimensions.

Chapter 1. Overview

This report describes and analyses the recently released dataset of information extracted from the log files generated during the Survey of Adult Skills cognitive assessment. It explores the potential and shortcomings of these data, as well as pitfalls to avoid when working with them. This chapter explains the value and limitations of log files and discusses the information available from the Survey of Adult Skills log files, with particular focus on timing indicators.

This report describes the content and characteristics of process data generated in the course of the Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC) (hereafter referred to as "PIAAC") and stored in log files, with examples of how these recently released data might be analysed. The potential of process data to provide information relevant to improving cognitive assessments and as a window into test-takers' behaviour has been known for at least 30 years (Bunderson, Inouye and Olsen, 1989[1]), but until recently there has been little progress in their analysis. This is partly due to the complexity of the data and the lack of well documented data sets accessible to social science researchers in readily usable formats.

The public release of PIAAC log files, along with documentation and dedicated software to import them, follows the release of similar data from the OECD Programme for International Student Assessment (PISA). The goal is to contribute to these recent advances.

This chapter explains the value and limitations of process data and discusses the information available from PIAAC log files, focussing on timing indicators. Chapter 2 provides background on what log files are and how they complement traditional proficiency scores, describing specific features of the PIAAC log files and how design of the assessment affects interpretation of the information they contain. Chapter 3 presents a descriptive analysis of the indicators that can be extracted from the PIAAC log files, focussing on timing indicators. Chapter 4 examines how respondents allocate time to the different tasks they face in the course of the assessment. Chapter 5 discusses how the information contained in the log files can be used to construct indicators of test disengagement.

The value of log files…

Computer-based administration is increasingly the norm for large-scale assessments. This has been made possible by technological developments and the increasing familiarity of test-takers with computers and digital devices. Computer-based administration makes it more efficient to administer, manage and monitor surveys, and it also reduces the risk of human error.

More importantly, for the purposes of this report, computer-based administration makes it possible to collect a richer set of information on test-takers. In principle, it is possible to capture a complete record of communication between the user interface and the server. This means that it is possible to observe not only a respondent's final answer to a specific assessment item, but also all interactions with the testing platform as he/she answers the question. Moreover, as all events recorded are associated with a timestamp, it is possible to compute the amount of time elapsed between these events.

Information about respondents' actions in the course of the assessment is potentially useful to understand their cognitive strategies. In this sense, log files can be seen as a "window into students' minds" (Greiff, Wüstenberg and Avvisati, 2015[2]). They can help policy makers, researchers and educators to better understand the underlying causes of low and high performance, and thus to design appropriate interventions.

Over and above the cognitive constructs that test items are designed to measure, the information contained in log files can be used to investigate aspects of respondents' ability, attitudes and behaviour. For example, information on the amount of time respondents devote to the different items of the assessment has been used to compute various indicators of test-takers' attitudes, such as motivation, engagement and perseverance. Performance in a test is always the combined outcome of the ability of the respondent and the effort exerted in the course of the assessment. In low-stakes assessments, such as PIAAC and PISA, information on the motivation and engagement of respondents is essential for interpreting differences in observed performance.

The analysis of log files, therefore, offers considerable promise in terms of enriching the information obtained in large-scale assessments. In particular, it will help to develop a more nuanced and more accurate picture of respondents' skills. These insights will help to improve the design of assessments and, ultimately, more effective training and learning programmes.

… and their limitations

The analytical potential of log files has only recently begun to be fully appreciated and exploited, although it was anticipated 30 years ago by Bunderson, Inouye and Olsen (1989[1]). Existing log files are usually unformatted files that need to be decrypted before being used for statistical analysis. They are not easy to analyse and should be seen as a useful but incidental by-product of the introduction of computer-based test delivery.

As log files are records of the interactions between respondents and survey items, interpretation of the information contained in the files is necessarily item-dependent. This is further complicated by the fact that existing log files contain only a subset of the respondent-computer interactions and the choice of which information to record was usually not informed by considerations about the usefulness of the data for subsequent analysis. Moreover, many of the actions that respondents undertake while solving an assessment item cannot be recorded in log files (e.g. notes taken on a piece of paper or mental reasoning). Some of this information could be collected by using devices such as webcams or eye-tracking devices, but this has not yet been done on a large scale.

Several conclusions flow from the above. First, researchers using log files need to have detailed knowledge of the characteristics of test items, including response formats, context and possibly content. Second, the item-dependent nature of test-taker/item interactions means that the same indicator can be interpreted differently for different items, so generalisations across multiple items are not straightforward. Third, the analytical utility of log files could be increased if users of the data were involved in: 1) defining the information to be captured in the files; and 2) deciding what derived variables or indicators should be included in user-accessible files for public and scientific use. In future, it will be important to consider the potential of log files to help understand cognitive processes and test-taker behaviour in the process of item design.

Information available from PIAAC log files

This report focuses on the analysis of timing indicators, which are available for all literacy and numeracy items and are easy to interpret consistently across different items. Other recent research papers based on the analysis of data from log files have examined the processes of solution of test items. These analyses tend to be highly item-specific, due to the individual nature of interactions between test-takers and specific items in the case of complex interactive items, such as those in the assessment of Problem Solving in Technology-Rich Environments (PSTRE). Recent attempts to analyse log files from PSTRE items using techniques borrowed from text mining and natural language processing include He and von Davier (2015[3]; 2016[4]) and He, Borgonovi and Paccagnella (forthcoming[5]).

In this report, the analysis concentrates on three indicators: 1) time on task (the total time spent on an item by the respondent); 2) time to first interaction (the time elapsed between the moment when an item is presented to the respondent and the moment at which he/she first interacts with the testing platform; and 3) time since last action (the time elapsed between the respondent's last interaction with the platform, typically inserting the answer, and the moment at which he/she and moves on to the next item).

Differences in timing indicators across countries and respondents

A first important finding of this report is the cross-country variation in the amount of time respondents spent on the PIAAC assessment. Respondents in Norway, Germany, Finland, and Austria took the longest time to complete the literacy and numeracy assessment (about 50 minutes on average). In Spain, Italy, Slovak Republic, England / Northern Ireland (United Kingdom) and Ireland, respondents spent about 40 minutes on average. A similar picture emerges when looking at the other timing indicators.

At the country level, the overall time spent on the assessment is positively correlated with average performance and negatively correlated with the incidence of missing answers.

At the individual level, time spent on the assessment tends to increase with the age and education level of respondents, despite the fact that older individuals also display a higher propensity to skip items. Gender differences are relatively small, with women spending about one minute less than men to complete the literacy and numeracy assessment.

Respondents reporting greater familiarity with information and communications technology (ICT) tend to complete the assessment more rapidly than others, but the difference disappears after controlling for other observable characteristics. Familiarity with ICT is also associated with a shorter time to first interaction and a longer time since last action.

How respondents allocate time to different items

The time spent on different items is closely related to the intrinsic characteristics of items, most notably item difficulty.

Respondents devoted a significantly smaller amount of time to items administered in the second half of the assessment. This was accompanied by an increase in the proportion of missing answers and a decrease in performance, suggesting that the decrease in time on task is probably due to fatigue or disengagement.

Respondents appear to allocate time to items in a rational way. They tend to spend the most time on items that are challenging but feasible (for which the ex ante individual probability of giving a correct answer is close to 50%), while spending little time on items that, in relation to their estimated proficiency, are very easy or very difficult.

The analysis also shows that spending more time on an item increases the probability of giving a correct answer, although at a declining rate.

Log files can be used to capture respondents' disengagement

Timing information can be used to construct indicators of disengagement. Respondents can be considered disengaged with an item if they spend too little time on it (on the basis of item-specific time thresholds). In such situations, it can be assumed that the respondent has not even devoted the effort necessary to understand the item and has skipped it without trying to determine if he/she was in a position to give a correct answer.

Disengagement may occur because PIAAC is a low-stakes assessment, and respondents do not have a strong incentive to perform at their best during the test. In assessments such as PIAAC or PISA, disengagement is an undesirable phenomenon, because it can introduce variation in estimated proficiency that is unrelated to the cognitive skills that the surveys intend to measure.

At the same time, disengagement may be associated with respondents' attitudes or intrinsic motivation, which may well be related to important outcomes in real life. A joint analysis of disengagement and actual performance helps to better interpret the results of the survey and to perform more meaningful comparisons across different countries or different socio-demographic groups.

Disengagement varies across countries and socio-demographic groups

The incidence of disengagement varies substantially across countries. In Finland, the Netherlands and Norway, less than 10% of respondents are disengaged in relation to at least 10% of the items, compared to more than 20% in France, Ireland, Poland and the Slovak Republic, and more than 30% in Italy.

Disengagement is more likely to be observed on items presented in the second module of the assessment. This is consistent with the analysis of time allocation to different items. Adults with low levels of education and adults who are less familiar with ICT are more likely to become disengaged in the course of the assessment.

Moving forward

Research using log files is still in its infancy. PIAAC was the first large-scale international assessment delivered primarily on computers, and the information available from the PIAAC log files has already been used in a number of analyses. It has contributed to understanding what can be drawn from this type of data and aided in exploring substantive issues, such as test-engagement and respondents' cognitive strategies. However, current PIAAC log files are, to a large extent, an accidental by-product of the computer-testing platform. Neither the items nor the information stored in log files were designed with a view to maximising the analytical potential of the information collected. As a result, analysis of log files is often cumbersome and item-specific, and the information they contain often lends itself to multiple interpretations.

The release of PIAAC log files has sparked a lot of interest on the part of both researchers and policy makers, and the LogDataAnalyzer (an instrument for processing data in PIAAC) has greatly contributed by facilitating access to the data.[1] By capitalising on the lessons learned from these data and the results of this report, future large-scale assessments will likely be able to improve their design to maximise the research potential of log files.

Item design plays a crucial role in maximising the potential of log files. More interactive items, for instance, offer more possibilities to observe and record a variety of respondent-computer interactions. For the data to be interpretable without ambiguity, it is important to prespecify theoretical constructs or competing theoretical hypotheses that log files will be able to measure or test. In particular, it should be made clear whether the information recorded in log files is used to better measure the underlying cognitive construct (such as proficiency in literacy, numeracy, or problem solving), or whether it can be used as a proxy for other dimensions of respondents' skills (which might include personality traits or attitudes).

Some improvements are relatively easy to achieve. Even if item design remains constant, potentially useful information that is currently not available could be recorded in future. For instance, it would be useful to track the input in text fields. Even without specifying the exact content, information on the insertion or deletion of characters would provide useful insights on the approaches followed by test-takers. Similarly, for multiple-choice items, it would be useful to track how many times respondents have checked a box (and which one) and whether they changed their mind before confirming the final answer.

It would be also possible to rethink the derived variables to be released in public-use files. For example, the analysis presented in this report shows that "time to first interaction" is largely dependent on item content, which limits its usefulness. On the other hand, it would be helpful to add "item position" to the public database to facilitate analysis. Prior to deciding on the content of the PIAAC public-use files for the second cycle of the study, it would be valuable to have experts review the current variables and suggest new ones, where relevant.

Note

[1] For more information, see Annex A and https://tba.dipf.de/en/projects/logdataanalyzer.

References

Bunderson, C., D. Inouye and J. Olsen (1989), "The four generations of computerized educational measurement.", in *Educational Measurement, 3rd ed.*, American Council on Education. [1]

Greiff, S., S. Wüstenberg and F. Avvisati (2015), "Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving", *Computers & Education*, Vol. 91, pp. 92-105, http://dx.doi.org/10.1016/J.COMPEDU.2015.10.018. [2]

He, Q., F. Borgonovi and M. Paccagnella (forthcoming), "Using process data to understand adults' problem-solving behaviours in PIAAC: Identifying generalised patterns across multiple tasks with sequence mining"*, OECD Education Working Papers*, OECD Publishing, Paris. [5]

He, Q. and M. Von Davier (2015), "Identifying feature sequences from process data in problem-solving items with n-grams", in van der Ark, L. et al. (eds.), *Quantitative Psychology Research The 79th Annual Meeting of the Psychometric Society*, Springer, New York, NY, http://dx.doi.org/10.1007/978-3-319-19977-1_13. [3]

He, Q. and M. von Davier (2016), "Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment", in Rosen, Y., S. Ferrara and M. Mosharraf (eds.), *Handbook of Research on Technology Tools for Real-World Skill Development*, http://dx.doi.org/10.4018/978-1-4666-9441-5.ch029. [4]

Chapter 2. What log files are and why they are useful

This chapter explains what log files are and how they complement traditional proficiency scores. It describes the features of the log files generated in the Survey of Adult Skills cognitive assessment and explores the impact of assessment design on interpretation of information contained in the files.

What are log files?

Log files are the traces of the communication events between a user interface and a server. They are unformatted, produced on a large scale and not designed to be interpreted. The primary goal of log files is to serve as a means of communication between the interface and the server and a way to store information within a software application. The structure and contents of log files are generally created by software developers, not survey scientists, so their content is typically determined by the functionality of the computer interface, rather than the needs and interests of researchers and analysts.

The availability of log files for studies like the Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC) (hereafter referred to as "PIAAC"), or the Programme for International Student Assessment (PISA) can thus be seen as a by-product of technological innovations that increasingly allow these types of assessments to be administered on a computer-based platform, replacing more traditional paper-based assessments.

PIAAC was the first large-scale international assessment to be primarily designed for computer-based administration (Kirsch and Lennon, 2017[1]). The advantages of computer-based administration are manifold. It automates error-prone tasks, such as questionnaire branching (when delivery of questions depends on how participants answered previous questions) and manual and costly encoding of handwritten answers into a formatted dataset, and also allows automatic scoring of items.

Moreover, the technology is helpful for implementation of adaptive testing. In an adaptive test, the sequence of items (particularly their difficulty) is targeted to the expected proficiency of respondents, which is inferred from responses to previous questions. The use of such an algorithm maximises the efficiency of information that can be extracted from answers to a particular item, by avoiding easy questions for individuals who have proven that they can correctly answer more difficult items. The implications of adaptive testing for the analysis of log files are discussed more extensively later in this chapter.

The PIAAC testing application had to fulfil two important technical requirements: 1) flexibility with respect to operating systems, hardware and character fonts (most notably non-western alphabets) to enable successful implementation of the platform in all participating countries; and 2) the possibility of adapting the system to PIAAC-specific features. The web-based architecture and open-license status of the TAO (*testing assisté par ordinateur*) platform satisfied these requirements. Chapter 9 of the *Technical Report of the Survey of Adult Skills (PIAAC)* provides further details on TAO and a detailed overview of the software framework to which log files belong (OECD, 2013[2]).

Like other similar platforms, TAO is structured around two main components: a user interface and a server. Survey participants interact with a user interface similar to a web browser. It displays test items sequentially, displaying embedded buttons, text entry boxes, checkboxes and selection in a single pane, together with the item stimulus contents and questions. The interface sends all relevant participants' inputs (including final answers) to the server. The server saves this information, adapts the display dynamically according to participants' actions within each item and directs the user towards the next item or question. As the test is adaptive, this allocation follows predefined rules that assign items according to past answers and participants' characteristics.

Log files are the traces of the communications between the user interface and the server. Importantly, all these records come with a precise timestamp that allows reconstruction of the complete chronology of respondents' interactions with the test application over the course of the assessment.

Log files were created as a means to communicate and store information within a software application. They are comparable to the kind of web server traces that constitute the basic elements of what is now commonly known as big data. From the point of view of a user of survey data, the files are unformatted, not designed to be interpreted and produced on a large scale. They are typically stored in xml format, where (as in html) relevant information is embedded in series of tags. Since each interview creates several log files, the total number of files made available can quickly become very large. The current release of PIAAC log files is based on more than 200 000 files, covering about 60 000 respondents. Each of these files mixes embedded tags, question-specific and user-dependant information, the latter accounting for just a fraction of their total size.

The opacity of log-file data has limited their use by education researchers, who are generally unfamiliar with xml files. This is why the public release of the PIAAC log files was accompanied by the release of the LogDataAnalyzer (LDA). The LDA was developed to turn log files into a user-friendly dataset. It also features a few statistical tools to make it easy to explore the distribution of each variable extracted (Figure 2.1).

It is important to note that log files record only some interactions between the respondent and the computer. They do not capture many potentially important aspects of respondents' behaviour. For example, interactions that are dealt with within the user interface itself and do not need to be transferred to the server (such as text scrolling or mouse movement) are not captured. But they could, in principle, provide valuable information about cognitive and non-cognitive processes followed by respondents in solving the items.

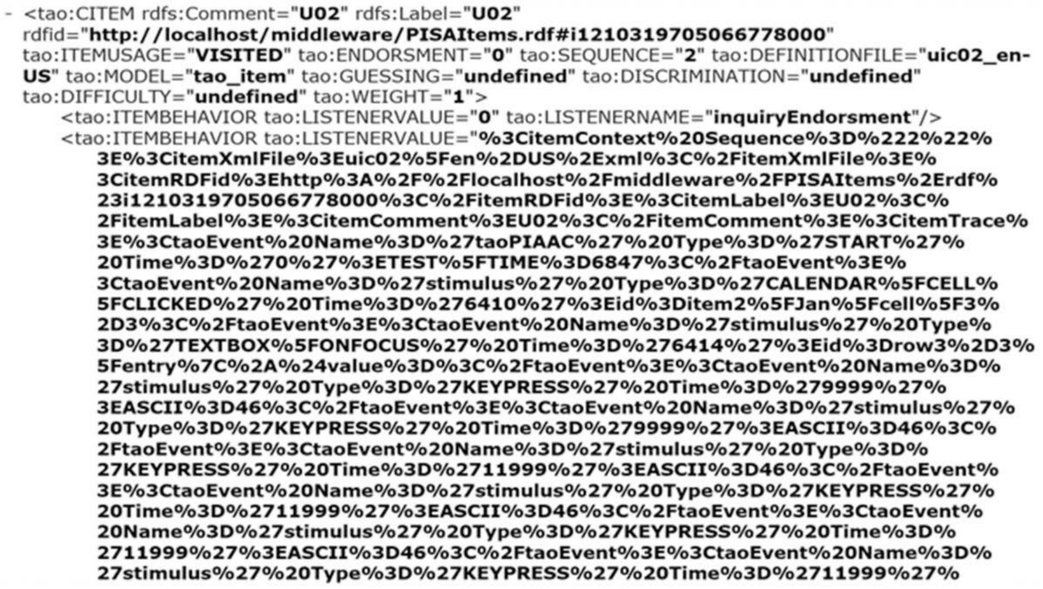
Moreover, participants' actions are obviously not limited to interactions with the computer. Between each recorded interaction, participants may be reading the stimulus, thinking or simply remaining idle. They could also take other actions to solve the item, such as writing on a piece of paper or using a calculator. All these actions are potentially very valuable to help understand the behaviour of test-takers and the cognitive processes they follow to solve the tasks, but they are missing from the information contained in log files (Maddox, 2018[3]; Maddox et al., 2018[4]).

Why are log files valuable?

While log files are essentially traces or records of the transfer of information between components of the testing software, they are valuable because they provide information on how respondents processed their answers. All actions on the part of test-takers that cause a change in the interface need to be communicated to the server and are hence recorded in the log file along with a timestamp. These include not only test-takers' responses to tasks, but also some (although not all) intermediate actions preceding input of the final answer.

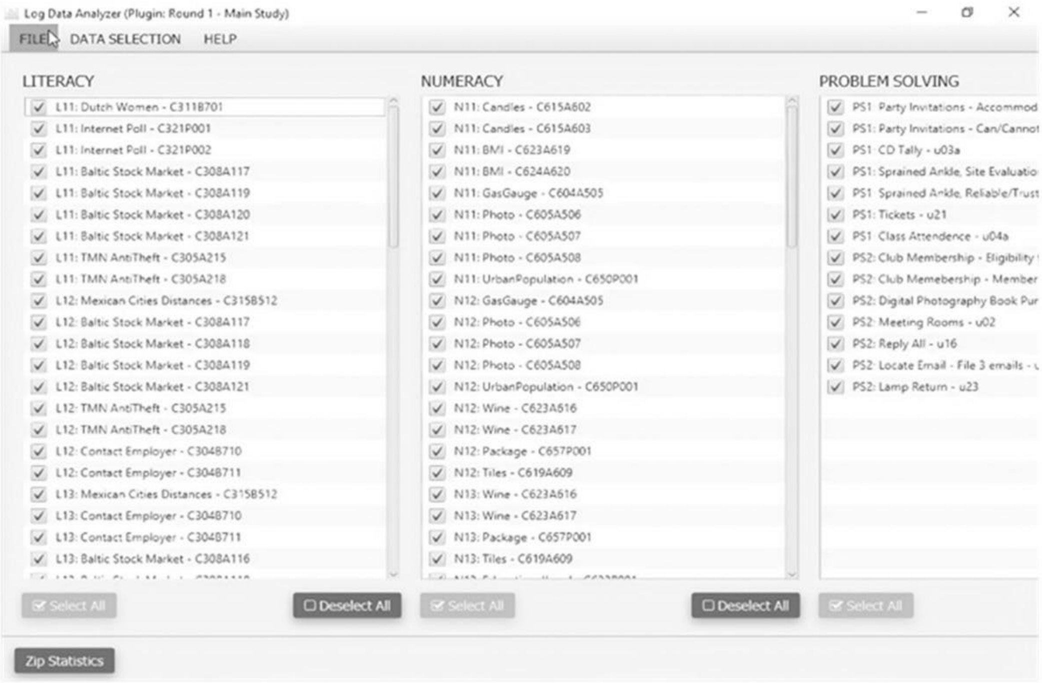
Figure 2.1. PIAAC LogDataAnalyzer

Panel A: Excerpt, PIAAC log file (screenshot)



```
- <tao:CITEM rdfs:Comment="U02" rdfs:Label="U02"
  rdfid="http://localhost/middleware/PISAItems.rdf#i1210319705066778000"
  tao:ITEMUSAGE="VISITED" tao:ENDORSMENT="0" tao:SEQUENCE="2" tao:DEFINITIONFILE="uic02_en-
  US" tao:MODEL="tao_item" tao:GUESSING="undefined" tao:DISCRIMINATION="undefined"
  tao:DIFFICULTY="undefined" tao:WEIGHT="1">
      <tao:ITEMBEHAVIOR tao:LISTENERVALUE="0" tao:LISTENERNAME="inquiryEndorsment"/>
      <tao:ITEMBEHAVIOR tao:LISTENERVALUE="%3CitemContext%20Sequence%3D%222%22%
          3E%3CitemXmlFile%3Euic02%5Fen%2DUS%2Exml%3C%2FitemXmlFile%3E%
          3CitemRDFid%3Ehttp%3A%2F%2Flocalhost%2Fmiddleware%2FPISAItems%2Erdf%
          23i1210319705066778000%3C%2FitemRDFid%3E%3CitemLabel%3EU02%3C%
          2FitemLabel%3E%3CitemComment%3EU02%3C%2FitemComment%3E%3CitemTrace%
          3E%3CtaoEvent%20Name%3D%27taoPIAAC%27%20Type%3D%27START%27%
          20Time%3D%270%27%3ETEST%5FTIME%3D6847%3C%2FtaoEvent%3E%
          3CtaoEvent%20Name%3D%27stimulus%27%20Type%3D%27CALENDAR%5FCELL%
          5FCLICKED%27%20Time%3D%276410%27%3Eid%3Ditem2%5FJan%5Fcell%5F3%
          2D3%3C%2FtaoEvent%3E%3CtaoEvent%20Name%3D%27stimulus%27%20Type%
          3D%27TEXTBOX%5FONFOCUS%27%20Time%3D%276414%27%3Eid%3Drow3%2D3%
          5Fentry%7C%2A%24value%3D%3C%2FtaoEvent%3E%3CtaoEvent%20Name%3D%
          27stimulus%27%20Type%3D%27KEYPRESS%27%20Time%3D%279999%27%
          3EASCII%3D46%3C%2FtaoEvent%3E%3CtaoEvent%20Name%3D%27stimulus%27%
          20Type%3D%27KEYPRESS%27%20Time%3D%279999%27%3EASCII%3D46%3C%
          2FtaoEvent%3E%3CtaoEvent%20Name%3D%27stimulus%27%20Type%3D%
          27KEYPRESS%27%20Time%3D%2711999%27%3EASCII%3D46%3C%2FtaoEvent%
          3E%3CtaoEvent%20Name%3D%27stimulus%27%20Type%3D%27KEYPRESS%27%
          20Time%3D%2711999%27%3EASCII%3D46%3C%2FtaoEvent%3E%3CtaoEvent%
          20Name%3D%27stimulus%27%20Type%3D%27KEYPRESS%27%20Time%3D%
          2711999%27%3EASCII%3D46%3C%2FtaoEvent%3E%3CtaoEvent%20Name%3D%
          27stimulus%27%20Type%3D%27KEYPRESS%27%20Time%3D%2711999%27%
```

Panel B: LDA main screen



LITERACY

☑ L11: Dutch Women - C311B701
☑ L11: Internet Poll - C321P001
☑ L11: Internet Poll - C321P002
☑ L11: Baltic Stock Market - C308A117
☑ L11: Baltic Stock Market - C308A119
☑ L11: Baltic Stock Market - C308A120
☑ L11: Baltic Stock Market - C308A121
☑ L11: TMN AntiTheft - C305A215
☑ L11: TMN AntiTheft - C305A218
☑ L12: Mexican Cities Distances - C315B512
☑ L12: Baltic Stock Market - C308A117
☑ L12: Baltic Stock Market - C308A118
☑ L12: Baltic Stock Market - C308A119
☑ L12: Baltic Stock Market - C308A121
☑ L12: TMN AntiTheft - C305A215
☑ L12: TMN AntiTheft - C305A218
☑ L12: Contact Employer - C304B710
☑ L12: Contact Employer - C304B711
☑ L13: Mexican Cities Distances - C315B512
☑ L13: Contact Employer - C304B710
☑ L13: Contact Employer - C304B711
☑ L13: Baltic Stock Market - C308A116

NUMERACY

☑ N11: Candles - C615A602
☑ N11: Candles - C615A603
☑ N11: BMI - C623A619
☑ N11: BMI - C624A620
☑ N11: GasGauge - C604A505
☑ N11: Photo - C605A506
☑ N11: Photo - C605A507
☑ N11: Photo - C605A508
☑ N11: UrbanPopulation - C650P001
☑ N12: GasGauge - C604A505
☑ N12: Photo - C605A506
☑ N12: Photo - C605A507
☑ N12: Photo - C605A508
☑ N12: UrbanPopulation - C650P001
☑ N12: Wine - C623A616
☑ N12: Wine - C623A617
☑ N12: Package - C657P001
☑ N12: Tiles - C619A609
☑ N13: Wine - C623A616
☑ N13: Wine - C623A617
☑ N13: Package - C657P001
☑ N13: Tiles - C619A609

PROBLEM SOLVING

☑ PS1: Party Invitations - Accommod
☑ PS1: Party Invitations - Can/Cannot
☑ PS1: CD Tally - u03a
☑ PS1: Sprained Ankle, Site Evaluatio
☑ PS1: Sprained Ankle, Reliable/Trust
☑ PS1: Tickets - u21
☑ PS1: Class Attendence - u04a
☑ PS2: Club Membership - Eligibility
☑ PS2: Club Memebership - Member
☑ PS2: Digital Photography Book Pur
☑ PS2: Meeting Rooms - u02
☑ PS2: Reply All - u16
☑ PS2: Locate Email - File 3 emails - u
☑ PS2: Lamp Return - u23

Such actions causing a change in the interface typically include switches between the simulated web pages forming the test stimulus and the highlighting of text. As a result, it is possible to construct indicators that give information on how participants interact with each item before delivering the answer. Such information would be nearly impossible to collect systematically in the paper-based version of assessment, but it is an outcome of technical features of the testing application in computer-based assessments.

The type and amount of information recorded in the log files depend on the choices made by software developers. A first important lesson of this report is that the value of log files is maximised when researchers co-operate with software developers to identify the variables that should be recorded in the log files and when test developers incorporate the possible availability of log files in designing assessment items.

The information contained in log files can then be seen as clues on how participants process answers. This is potentially valuable to users of survey data (statisticians, education researchers and policy makers), as well as survey designers and managers.

Policy makers, researchers and test developers

In PIAAC and similar international surveys, measurement is outcome-based. The psychometric models used to estimate the proficiency of respondents take into account successes and failures and use them to build a continuous proficiency scale.

However, the performance of respondents on any given test results from a combination of individual cognitive and non-cognitive traits that are not measured. The Reader's Companion for PIAAC (OECD, 2016[5]) acknowledges that: " […] the focus of the Survey of Adult Skills is less on the mastery of certain content […] and a set of cognitive strategies than on the ability to draw on this content and these strategies to successfully perform information-processing tasks in a variety of real-world situations." Importantly, this ability also includes the attitude of participants and their disposition to do well. However, although cognitive strategies, attitude and context provide a useful grid to describe test-taking behaviour, they are still ill-defined.

Up to now, efforts to understand skills acquisition with PIAAC have built on linking PIAAC proficiency scores with personal characteristics collected in the background questionnaire. PIAAC proficiency scales are principally designed to rank and compare survey participants for the purpose of studying the distribution of skills, both as a whole and across groups. These comparisons have provided essential results, most notably a quantification of the relationships between education, age and skills. These relationships identify good and bad performers and, consequently, which characteristics are associated with higher skill levels.

However, such analysis cannot account for differences in the cognitive processes deployed during the assessment. PIAAC proficiency scores are well suited to identify populations lacking skills, but they are not able to characterise the reasons behind a given performance, such as cognitive strategies, knowledge or attitudes, thus hindering development of policies to target these issues.

This limits the extent to which skill acquisition can be studied with PIAAC. Meaningful changes in skills always proceed from variations in knowledge, cognitive strategies or attitudes. Teachers, parents or peers do not teach skills per se, but influence attitudes and transfer knowledge or cognitive strategies.

Indeed, a large part of the research undertaken using log files or other forms of process data has tried to infer some measures of non-cognitive skills from participants' behaviour during the assessment (Goldhammer et al., 2016[6]; Anghel and Balart, 2017[7]) (see Chapter 5 for a full discussion and analysis of disengagement).

Because log files describe how participants interact with each cognitive item, they can be used to analyse the cognitive and non-cognitive resources deployed by respondents during the assessment. However, making inferences on cognitive resources on the basis of the data from log files is not an easy task. It demands a certain amount of ingenuity on the part of the analyst, and the analysis may deliver only partial results.

Since the content of individual items provides the context for interpretation of the log-file variables, item-specific analysis is a simple and useful way to take advantage of log files. The meaning of each log-file variable is, in the end, always item-dependent. Focusing on a single item makes the analysis easier and more robust. But that is at the cost of lower generalisability of the results, because it is often difficult to extract information from log files that consistently measures the same underlying construct across different items. For example, a participant who managed to solve a given item in 30 seconds could be considered either slow or fast, depending on the item.

Item-specific analysis generally consists of going beyond an interest in the difference between correct and incorrect answers. In some cases, cognitive strategies can be observed. Greiff, Wüstenberg and Avvisati (2015[8]) provide an excellent example of the promises offered by log-file data. They study a PISA 2012 item on climate control, extracted from the Complex-Problem-Solving domain. This item requires an understanding of how a multi-parameter system works. It is generally solved by following a strategy described as vary-one-thing-at-a-time. The implementation of the strategy can be identified through the log files, without taking the final answer into account. As a result, it is possible to classify respondents based on the strategy they followed, irrespective of whether they ended up giving the right answer.

Such analysis can be extended to a set of items, as long as their content is homogeneous enough to define common strategies or features. OECD (2015[9]) analyses web navigation strategies in a subset of PISA 2012 digital reading items. It distinguishes between task-oriented and exploratory navigation. In the case of exploratory navigation, while pupils may still find the correct answer, their browsing activity features visits to irrelevant web pages. Thanks to this distinction, participating countries can be classified according to the efficiency of web navigation rather than according to digital reading scores (i.e. in terms of attitude and cognitive strategies rather than outcome-based skills). A recent attempt to identify consistent indicators across PIAAC items measuring ability in Problem Solving in Technology-Rich Environments (PSTRE) is found in He, Borgonovi and Paccagnella (forthcoming[10]).

The more diverse the set of items, the less specific the analysis will be. At the same time, the conclusions will be more wide-ranging, opening the door to measuring attitudes or behavioural traits. However, a few indicators do lend themselves to consistent analysis of all items. The most straightforward example of such an indicator is probably time on task, which has been used, for instance, by Goldhammer et al. (2016[6]) to infer attitudes such as item disengagement among survey participants.

Measurements based on log-file variables are attractive, because they reflect actual behaviour, although only in the specific context of the PIAAC assessment. In this sense, they could be a useful complement to more traditional measures of behavioural traits. Psychometric measurements of individual behavioural traits, such as the widely used Big-Five scales or the readiness-to-learn scale available in PIAAC, generally rely on self-assessment. Although these measurements do not have the disadvantage of being context-specific, they are prone to other biases that are not present in the case of log-file data (e.g. lack of sincerity or differences in how respondents interpret the questions).

Log files can also help to better understand why some items display insufficient psychometric properties. The probability of successfully completing an item should be related only to the respondent's underlying proficiency. Non-construct-related factors, such as culture or gender, should not affect item difficulty. This is a particularly challenging constraint in international assessments, where it is not rare to find that the relative difficulty of an item is not the same in all countries.[1] In such cases, the item is said to lack measurement invariance. However, the available statistical procedures are only able to detect the presence or the absence of measurement invariance; they are silent on the underlying reasons behind the failure of an item to satisfy the invariance condition. Typically, the lack of invariance is due to some features of the item content or to translation errors. Information on respondents' behaviour contained in log files can be a useful complement to provide test developers with a better understanding of why a certain group of respondents finds a certain item more or less difficult.

Survey design and management

For survey designers and managers, log files have proved to be useful in improving data quality in several ways. In PIAAC and PISA, log files have been used to detect data falsification. By allowing a comparison of the processes leading to a response, log files represent powerful tools in the prevention and detection of data falsification in low-stakes assessments. In contrast with high-stakes assessments, such as exams, the most important source of falsification in an international survey such as PIAAC is not the participants themselves, but those involved in survey administration: interviewers, survey contractors or national managers.

Yamamoto and Lennon (2018[11]) highlight how log-file data, in particular timing data, can be used to detect cases of fabricated data. Interviewers who want to minimise effort can fill in questionnaires and assessment answers themselves, but doing that in a way that is consistent with the timing and response patters of real respondents would be cumbersome, and the amount of effort would likely offset the benefits. Survey managers who wish to inflate country performance could do so by replicating the response profiles of high achievers, if they have access to the master datasets. In doing so, however, they will also duplicate the associated timestamps (which are precise to the millisecond). Although identical answer profiles are plausible, identical timing profiles are not. Log files are particularly valuable for this purpose, because they are difficult to edit. In principle, it is possible to fabricate log files and create plausible profiles, but this would require much more sophisticated knowledge and far more time and effort than simply editing final datasets by copying and pasting respondent records.

The use of log files for the management of data quality has been taken further through their integration into dashboard software. Dashboards are tools designed to help survey managers monitor the progress of data collection. Mohadjer and Edwards (2018[12]) document the use of dashboards during the data collection phase of Round 3 of PIAAC in the United States. These dashboards were connected to interviewers' computers and used the log files throughout their generation during interviews in order to track interviewer activity. Thanks to this system, it was possible to detect suspicious cases during data collection (such as interviews taking place at unlikely times or assessments with improbably short completion times), identify mistakes or falsification in a timely manner and take corrective action. By increasing the chances of detection of such behaviour in close to real time, the integration of dashboards and log files can greatly reduce the incentives for falsification and effort reduction on the part of interviewers and survey administrators.

Content and characteristics of PIAAC log files

The interpretation of variables derived from log files depends to a large extent on the content of test items: the tasks test-takers must carry out, the questions they must answer and the nature of the item stimulus. Inferences made regarding the cognitive strategies of test-takers on the basis of information in log files only makes sense in light of the content and format of items.

To make sure that potential respondents do not have access to the items and the correct answers, in most cases, test items are treated as confidential and are not accessible to researchers. The log-data documentation helps external users to access the contents of items that are already public. Some confidential items, including all items from the assessment of PSTRE, are also available upon submission of a detailed research proposal. The PIAAC technical report (Chapter 2) provides definitions for all three domains and describes the different context categories, the different types of tasks and the various dimensions that contribute to item difficulty (OECD, 2013[2]). However, while the technical report is a helpful resource to understand how diverse cognitive items can be, it does not give any item-specific details.

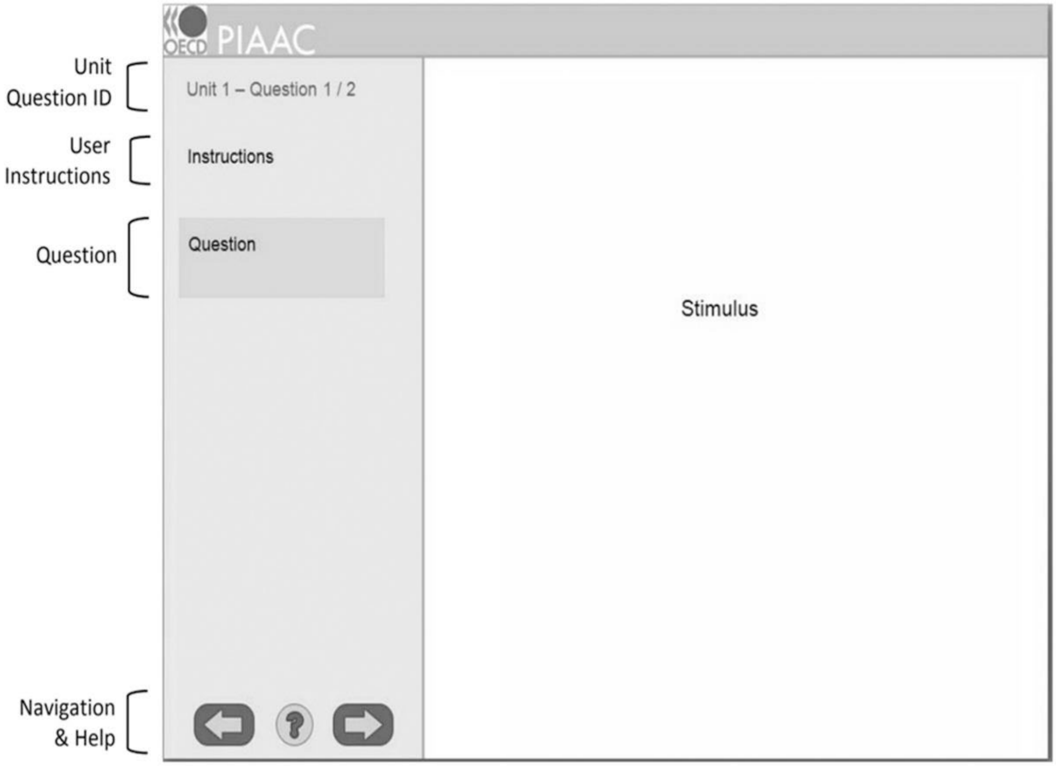
In the end, the type of information that recorded in the log files is a function of the interaction between the characteristics of each item and the characteristics of the digital assessment platform. Generally, the more complex the item stimulus, the more variables will be available. In principle, dynamic items, whose elements change in response to the actions of test-takers (e.g. manipulating values through the use of sliders or radio buttons) or become accessible only through the action of the test-taker (e.g. accessing a simulated web page by clicking on a hyperlink) will allow collection of more variables, as all changes in the user interface require some exchange of information between the server and the user interface.

It is important to keep in mind that assessment items have mainly been designed with the objective of estimating a proficiency score based on the final answer provided. Consequently, they often do not lend themselves to analysing the process through which the respondent has arrived at a specific answer. For example, it is not always possible or straightforward to unambiguously observe or define a variety of theoretically-grounded cognitive strategies that a respondent might choose to follow in trying to solve the items. This will depend on the design of the item and/or on the amount of information that ends up being recorded in the log file. By their nature, PSTRE items and, to a lesser extent, multipage literacy items lend themselves to this kind of analysis.

The user interface that is common to all PIAAC items is divided into two parts (Figure 2.2). The left panel features navigation buttons, presents the item and states the question or describes the task. Clicking on the right-hand arrow terminates the current unit and opens a new one. The right panel consists of a flexible stimulus frame in which graphical representations, text, a website or application environment can be displayed.

The features of the stimuli vary according to the domain (Table 2.1). All numeracy items contain either charts or print text. Literacy items include stimuli based on printed text, charts or web environments. Web environments can include one or several web pages. Compared to literacy and numeracy items, PSTRE items feature a wider range of stimuli, including web environments, e-mail environments and combinations of e-mail/spreadsheet/web environments.

Figure 2.2. PIAAC user interface



Source: OECD (2016[13]), *Technical Report of the Survey of Adult Skills, (Second Edition)*, www.oecd.org/skills/piaac/PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf.

Table 2.1. Types of stimuli in PIAAC items

	Literacy	Numeracy	PSTRE
Print text / chart	27	49	0
Web environment	22	0	4
E-mail environment	0	0	4
Multiple environments	0	0	6

Note: Multiple environments include spreadsheets, e-mail and web.

StatLink http://dx.doi.org/10.1787/888933959377

Response types define the format and range of possible answers. Numeracy and literacy items have different response types, and this will affect the interpretation of final answers. Response types can be classified as follows:

- Stimulus choice or left-panel choice, which features a limited number of precoded answers that may or may not be mutually exclusive.
- Stimulus clicking, which requires the participant to click on a graphical element in the stimulus (a cell in a table, a link).
- Stimulus highlighting, which targets a string or strings of text. In the clicking and highlighting response modes, a correct response is defined in terms of a range of

response actions (e.g. the minimum and maximum amount of text that can be highlighted for an answer to be correct).

- Left-panel numeric entry, which requires the participant to provide the answer in the form of a number. The range of possible incorrect answers will thus depend on response mode.

Different item formats may also provide different incentives to respond in the first place. For example, it could be the case that respondents are more likely to provide answers to multiple-choice items, as this permits guessing (there is in fact no penalty for providing a wrong answer). The effort to provide an answer may be greater and the expected benefits lower in items with a more open response format (such as the input of a number or highlighting a portion of text).

It may seem surprising that PSTRE items do not feature a response mode. In fact, PSTRE items typically require a participant to perform a task, not to answer a question. The correct response to PSTRE items generally involves the participant reaching an appropriate stage in the stimulus. PSTRE items are not framed as questions but as tasks. For example, several items ask respondents to select objects among a list to verify some criteria.

Table 2.2 shows the number of items by response mode. Most numeracy items require a numeric entry, while most literacy items require the highlighting of strings of text in the stimulus. In each domain, only a few items feature a multiple-choice response format.

Table 2.2. Distribution of response modes

Response type	Literacy	Numeracy
Left-panel numeric entry	3	31
Left-panel choice	0	5
Stimulus clicking	8	11
Stimulus highlighting	31	0
Stimulus choice	7	2

Note: Left-panel choice and stimulus choice are both multiple-choice items. In the former, respondents select the answer in the left panel; in the latter, they select the answer below the stimulus.

StatLink http://dx.doi.org/10.1787/888933959396

Table 2.3 lists the different variables extracted from the log files by the LogDataAnalyzer and the number of items to which they relate. Time on task, time to first interaction, number of helps and number of cancel actions are the only variables available for all items. In most cases, cancel actions and helps are very rare events. Final answers cannot be defined for PSTRE items.

Time to first interaction is a generic variable that has a very different interpretation depending on the nature of an item. In the simple static items, the first interaction will also be the final interaction, the selection or input of an answer. For more dynamic items the first interaction will be the first change in the stimulus.

Time since last action represents the time elapsed between the action of providing a final answer and the time at which the test-taker passes to the next item. Although this variable is present for all numeracy and literacy items, it does not capture exactly the same information for all items. Answer interactions are transferred immediately to the server in all response modes other than left-panel numeric entry. In that case, the content of the text field is transferred only when the item is terminated (i.e. when the test-taker moves to the next item). As a result, for all items with a numeric entry response, time since last action is defined as zero and provides no useful information. Most numeracy items are in this category.

Table 2.3. Variables extracted from log files

	Numeracy	Literacy	PSTRE
Final response	49	49	0
Time on task	49	49	14
Time to first interaction	49	49	14
Time since last action	49	49	0
including validation	18	45	0
Number using cancel button	49	49	14
Number using help menu	49	49	14
Number of highlight events	0	31	0
Number of page revisits	0	15	5
Number of page visits	0	15	5
Number of different pages visited	0	15	5
Sequence of visited web pages	0	15	5
Time-sequence of spent time on web pages	0	15	5
Number of created e-mails	0	0	6
Number of different e-mail views	0	0	6
Number of revisited e-mails	0	0	6
Sequence of viewed e-mails	0	0	6
Sequence of switching environments	0	0	6
Number of switching environments	0	0	6

StatLink http://dx.doi.org/10.1787/888933959415

The other variables record changes in the testing environment that result from participants' actions. Four variables were generated for e-mail environments: number of created e-mails; number of e-mail views; number of revisited e-mails; and sequence of visited e-mails. In items containing a web environment with several web pages, the LogDataAnalyzer extracts five different variables: sequence of web pages; time-sequence of web pages; number of page visits; number of page revisits; and number of different pages visited. Finally, a series of variables describes the sequence of switching environments.

The construction of a chronology of respondents' interactions with the test application is possible only for items containing web environments with several web pages and/or e-mail environments. This is true for a good proportion of literacy items (if they feature several web pages) and most PSTRE items. As numeracy items are all displayed in a much simpler environment, it is not possible to construct a similar chronology.

Although these variables cover most of the information available in log files, the documentation also includes details about all the various events that can be extracted from them. For every type of event, a short description is presented, along with the xml code that stands for the event and a few examples. Guidelines about the structure of log files complete these descriptions.

Log-file data are publicly available for 16 countries. They include data recorded from the cognitive instruments only.

Other features of test design relevant to analysis of log-file data

When analysing data from the PIAAC log files, it is important to consider two features of PIAAC: the routing of respondents in the computer-based branch and the adaptive nature of the assessment. These features are designed to maximise the efficiency of PIAAC and respond to the main objective of the study, which is to estimate the distribution of proficiency of the target population in the most efficient way. However, both of these features have consequences for secondary analysis of data at the individual level.

Routing of respondents

According to the PIAAC design, not all respondents were routed to the computer-based branch of the assessment (Figure 2.3). Log-file data are obviously not available for respondents that were routed in the paper-based branch of the assessment. It follows that the log-file data are not representative of the entire PIAAC target population, but are only available for a selected sample.

The allocation of respondents to the paper-based assessment followed a two-stage process. First, respondents who declared no prior computer experience, or who failed a simple test of information communications technology (ICT), were automatically directed to the paper-based assessment. In addition, respondents who passed the ICT assessment were offered the possibility of opting out of the computer-based route and choosing to take the assessment on paper. As a result, the population for which log-file data are available (equivalent to the population assigned to the computer-based assessment) is a sub-group within the PIAAC target population that: 1) had some computer experience; 2) accepted the computer-based assessment; and 3) passed the core ICT test. There is considerable variation in the proportion of the population that meets these criteria across countries (Table 2.4).

In all countries, log-file data are available for a majority of the overall sample, and in most of them, by a large margin. The lowest proportions are in Estonia, Italy and the Slovak Republic (60% or below), but the proportion exceeds 75% in Belgium (Flanders), Denmark, England / Northern Ireland (United Kingdom), Finland, the Netherlands, Norway and the United States.

Table 2.4. Proportion of respondents that took the computer-based assessment

	Proportion of sample covered	Number of cases
Austria	0.746	3 827
Belgium (Flanders)	0.755	4 125
Denmark	0.824	6 036
England / Northern Ireland (United Kingdom)	0.806	7 163
Estonia	0.531	4 053
Finland	0.815	4 454
France	0.692	4 836
Germany	0.825	4 509
Ireland	0.678	4 055
Italy	0.605	2 797
Netherlands	0.874	4 521
Norway	0.836	4 286
Poland	0.635	5 951
Slovak Republic	0.609	3 487
Spain	0.640	3 873
United States	0.810	4 060

StatLink http://dx.doi.org/10.1787/888933960156

Adaptive nature of assessment design

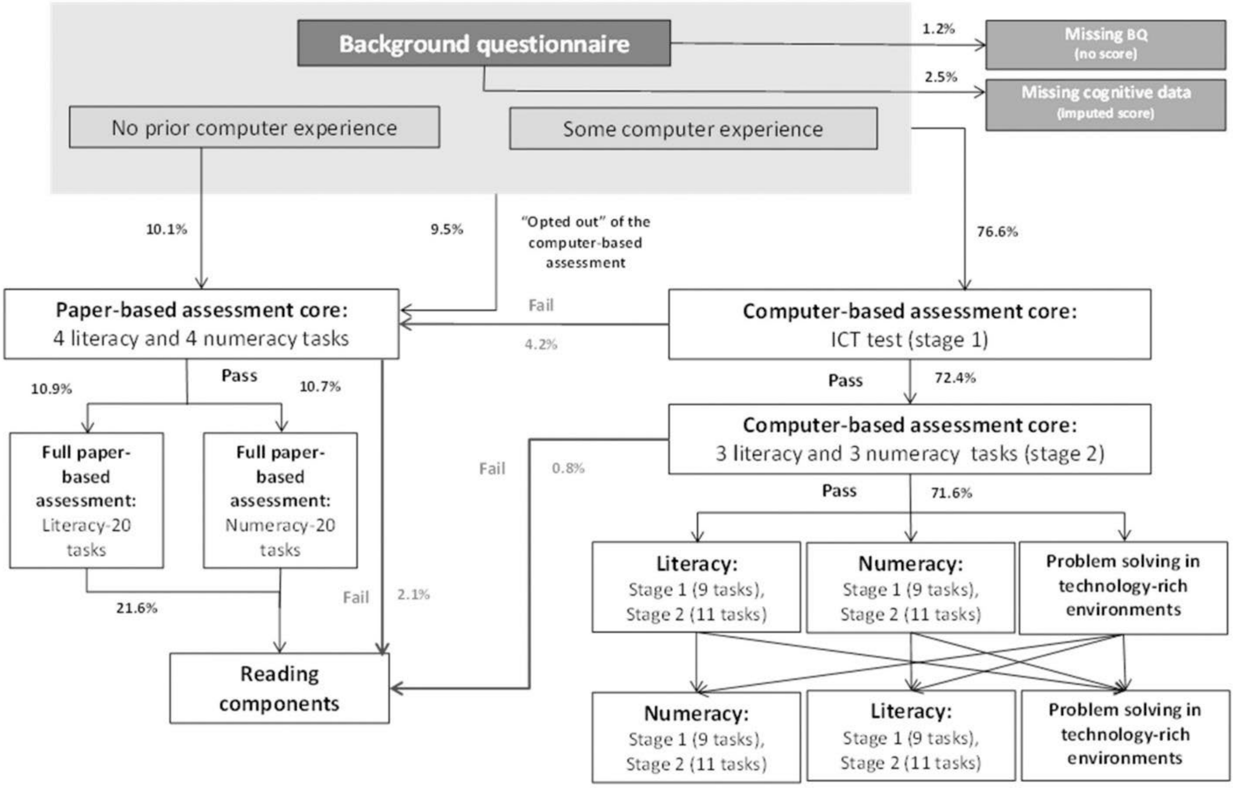
In PIAAC, items were grouped in booklets, with each individual test-taker answering items from a selection of all booklets. The population answering any specific item is then, strictly speaking, not representative of anything. The allocation of booklets to test-takers followed several sequential steps (Figure 2.3).

Test-takers taking the computer-based version of PIAAC were initially randomly allocated to a literacy, numeracy or PSTRE module. Participants assigned to literacy or numeracy would obtain first-stage and second-stage booklets. Booklets varied in difficulty, and allocation of the booklets to participants was only conditionally random. Allocation to the first-stage booklet was determined by a set of background variables that were assumed to be correlated with proficiency, such as age and education. Allocation to the second-stage booklet was based on the same background variables and on performance on the first-stage booklet. Knowledge of the characteristics that drove the allocation of respondents to different booklets is therefore essential to any kind of analysis that aims to investigate behaviour at the item level.

After the first module, participants were allocated to a second module, with the restriction that no respondent could take a second literacy or numeracy module (it was, however, possible to be assigned a second PSTRE module). Allocation of literacy and numeracy booklets in the second module followed the same rules as in the first module.

To some degree, the representativeness of the population answering each item was traded off for more efficient measurement of proficiency at the level of the overall target population. The more successful a participant was (according to background characteristics and answers to previous items), the more likely he/she was to get a booklet with more difficult items. This is not an issue for PSTRE. In that domain, the limited size of the item pool did not allow using an adaptive design, with the consequence that all respondents who were (randomly) allocated to PSTRE modules took exactly the same items.

Figure 2.3. PIAAC assessment design


Background questionnaire
1.2%
Missing BQ (no score)
2.5%
Missing cognitive data (imputed score)
No prior computer experience
Some computer experience
10.1%
9.5%
"Opted out" of the computer-based assessment
76.6%
Paper-based assessment core: 4 literacy and 4 numeracy tasks
Fail
4.2%
Computer-based assessment core: ICT test (stage 1)
10.9%
Pass
10.7%
Pass
72.4%
Computer-based assessment core: 3 literacy and 3 numeracy tasks (stage 2)
Full paper-based assessment: Literacy-20 tasks
Full paper-based assessment: Numeracy-20 tasks
Fail
0.8%
Pass
71.6%
Literacy: Stage 1 (9 tasks), Stage 2 (11 tasks)
Numeracy: Stage 1 (9 tasks), Stage 2 (11 tasks)
Problem solving in technology-rich environments
21.6%
Fail
2.1%
Reading components
Numeracy: Stage 1 (9 tasks), Stage 2 (11 tasks)
Literacy: Stage 1 (9 tasks), Stage 2 (11 tasks)
Problem solving in technology-rich environments

Source: OECD (2016[5]), *The Survey of Adult Skills: Reader's Companion, Second Edition,* OECD Skills Studies, https://dx.doi.org/10.1787/9789264258075-en.

A consequence of the adaptive nature of the literacy and numeracy assessment is that the subsamples of participants who answer a given literacy or numeracy item are generally not comparable. The share of respondents assigned to any given item ranges from 10% to 40% of the overall sample. As the test is adaptive, good performers tend to be assigned more difficult items. Individual averages over all assigned items are thus not particularly informative. For instance, two participants with a similar proportion of correct answers might actually end up with very different scores because they were assigned to booklets of different difficulty. Raw comparisons of statistics on different items could be misleading, because they are not computed on a similar population. As allocation is at the booklet level, analysis should focus on the population assigned a given booklet and study and compare the items it contains.

Conclusions

Log files have the potential to significantly enrich the information derived from large-scale assessments. In particular, they are likely to help deliver a more nuanced, multifaceted and, ultimately, more realistic picture of the skills possessed by respondents. They also have the potential to provide important insights that would help to design more effective training and learning programmes.

However, the research on log files is still in its infancy. PIAAC is the first large-scale assessment that has allowed a serious analysis of log files, but the PIAAC log files are, to a large extent, a by-product of the fact that PIAAC is a computer-based assessment. As a result, their analysis is often cumbersome, and the information they contain often lends itself to multiple interpretations.

Reaping the full benefits of log files will require specifically designing the assessment items, the delivery platform and the hardware and software infrastructure to capture well defined and theory-based alternative cognitive strategies that respondents may follow when approaching assessment tasks. Similar points are made by Bunderson, Inouye and Olsen (1989[14]). The fourth generation of their agenda for computer-assisted assessment, which they call “intelligent measurement”, aims to provide explanations for individual performance and advice to learners and teachers. The huge progress made in the last few years is a clear sign that we are finally embarking upon a generation of intelligent measurement.

Note

[1] See Chapter 12 of the PIAAC Technical Report (OECD, 2013[2]) for an illustration of the statistical procedures used to detect the psychometric properties of assessment items.

References

Anghel, B. and P. Balart (2017), "Non-cognitive skills and individual earnings: new evidence from PIAAC", *SERIEs*, Vol. 8/4, pp. 417-473, http://dx.doi.org/10.1007/s13209-017-0165-x. [7]

Bunderson, C., D. Inouye and J. Olsen (1989), "The four generations of computerized educational measurement.", in *Educational measurement, 3rd ed.*, American Council on Education. [14]

Goldhammer, F. et al. (2016), "Test-taking engagement in PIAAC"*, OECD Education Working Papers*, No. 133, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jlzfl6fhxs2-en. [6]

Greiff, S., S. Wüstenberg and F. Avvisati (2015), "Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving", *Computers & Education*, Vol. 91, pp. 92-105, http://dx.doi.org/10.1016/J.COMPEDU.2015.10.018. [8]

He, Q., F. Borgonovi and M. Paccagnella (forthcoming), "Using process data to understand adults' problem-solving behaviours in PIAAC: Identifying generalised patterns across multiple tasks with sequence mining"*, OECD Education Working Papers*, OECD Publishing, Paris. [10]

Kirsch, I. and M. Lennon (2017), "PIAAC: a new design for a new era", *Large-scale Assessments in Education*, Vol. 5/1, p. 11, http://dx.doi.org/10.1186/s40536-017-0046-6. [1]

Maddox, B. (2018), "Interviewer-respondent interaction and rapport in PIAAC article information", *Quality Assurance in Education*, Vol. 26/2, pp. 182-195, http://dx.doi.org/10.1108/QAE-05-2017-0022. [3]

Maddox, B. et al. (2018), "Observing response processes with eye tracking in international large-scale assessments: Evidence from the OECD PIAAC assessment", *European Journal of Psychology of Education*, http://dx.doi.org/10.1007/s10212-018-0380-2. [4]

Mohadjer, L. and B. Edwards (2018), "Paradata and dashboards in PIAAC", *Quality Assurance in Education*, Vol. 26/2, pp. 263-277, http://dx.doi.org/10.1108/QAE-06-2017-0031. [12]

OECD (2016), *Technical Report of the Survey of Adult Skills (PIAAC) (Second Edition)*, OECD Publishing, Paris, http://www.oecd.org/skills/piaac/PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf. [13]

OECD (2016), *The Survey of Adult Skills: Reader's Companion, Second Edition*, OECD Skills Studies, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264258075-en. [5]

OECD (2015), “The importance of navigation in online reading: Think, then click”, in *Students, Computers and Learning: Making the Connection*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264239555-7-en. [9]

OECD (2013), *Technical Report of the Survey of Adult Skills (PIAAC)*, OECD Publishing, Paris, http://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf. [2]

Yamamoto, K. and M. Lennon (2018), “Quality assurance in education understanding and detecting data fabrication in large-scale assessments article information”, *Quality Assurance in Education*, Vol. 26/2, pp. 196-212, http://dx.doi.org/10.1108/QAE-07-2017-0038. [11]

Chapter 3. Timing indicators in the Survey of Adult Skills

This chapter provides an overview of the indicators extracted from log files, analysing differences across countries, cognitive domains and the socio-demographic characteristics of respondents. Older respondents and more educated adults tended to spend more time on the assessment, and time spent on the assessment was found to be positively correlated with performance. The data also show a positive relationship between timing indicators and literacy proficiency.

Basic facts on timing indicators

As discussed in Chapter 2, this report focuses on indicators that are available for all or most items and cognitive domains (literacy, numeracy, and Problem Solving in Technology-Rich Environments [PSTRE]) and, more importantly, on indicators for which the meaning and interpretation can be plausibly considered consistent across different domains.[1] This chapter closely examines four indicators, three concerning timing information and the fourth concerning missing answers.

The three timing indicators are: 1) overall time spent on an item (time on task); 2) time spent between the appearance of an item on the screen and the first action undertaken by the respondent on that item (time to first interaction); and 3) time spent between the last action undertaken by the respondent and final validation of the answer (time since last action). It should be noted that time since last action is not defined for either PSTRE items or items requiring numeric answers (see Chapter 1).

The fourth indicator analysed is the proportion of missing answers, defined as the share of items to which the respondent did not give an answer. All items skipped by respondents are taken into account, irrespective of the amount of time they took before moving to the next item.

All analysis in this chapter is conducted from an individual perspective. Timing and response information on different items is aggregated or averaged at the level of the individual respondent. In other words, the analysis focuses on the average behaviour of respondents in the assessment. A complementary approach followed in subsequent chapters is to take an item-level perspective, exploring, for example, whether respondents' behaviour changes according to the characteristics of specific items or over the course of the assessment.

Interpreting timing indicators

Timing indicators have no straightforward or obvious interpretation. The Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC) (hereafter referred to as "PIAAC") was not conceived as a timed assessment, meaning that respondents were allowed to spend as much time as they wanted trying to solve the items. Spending more time on the assessment could be interpreted as a sign of higher motivation, but it could also be interpreted as a sign of lower ability, to the extent that equally motivated but less able individuals would likely need to spend more time on an item to solve it.

It follows that, although speed was not part of the constructs that PIAAC aimed to assess, timing information should be jointly analysed with information on proficiency. However, two complications arise in this respect. First, performance and timing are measured simultaneously in PIAAC. Second, the adaptive and rotated design of the assessment implies that respondents were not all assigned the same items. Typically, respondents who demonstrated higher proficiency were assigned more difficult items, which could require more time (e.g. because of lengthier stimuli).

Another important thing to consider is possible heterogeneity across domains. PIAAC included three different domains: literacy, numeracy, and PSTRE. However, as illustrated in Figure 2.3, each respondent undertook only two domains. After the computer-based assessment core, respondents were randomly allocated to a first module of literacy, numeracy or problem solving. After this first module, they were again randomly allocated to a second module, with the restriction that respondents could not take two literacy or two numeracy modules (while it was possible to take two modules of problem solving). Each literacy and numeracy module included 20 items, while the PSTRE modules included 7 items each.

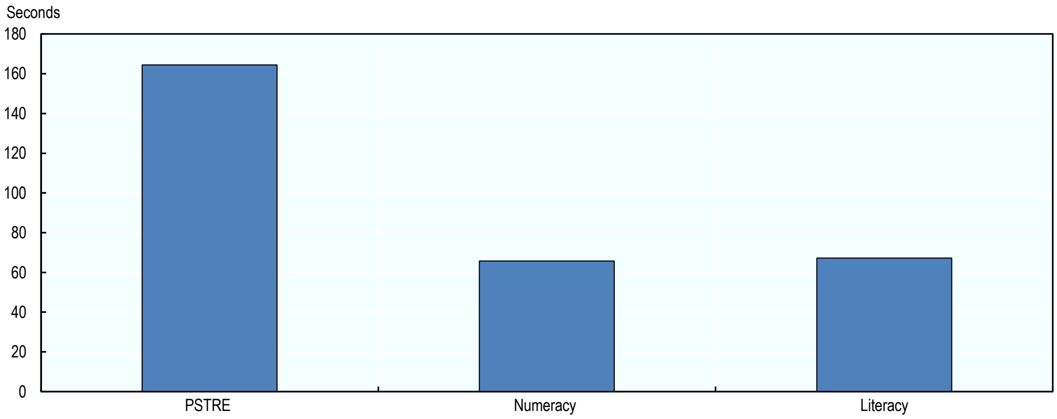
Furthermore, it must be kept in mind that the assessment of problem solving was optional. Among the countries for which process data are available, France, Italy and Spain did not administer the problem-solving assessment. As a result, all respondents in these countries took one literacy module and one numeracy module (in random order).

Modules were designed to be approximately the same length. As the PSTRE modules included only seven items, it follows that they took much more time to complete than literacy and numeracy items. Figure 3.1 shows the average time spent on literacy, numeracy and PSTRE items across all countries in which all three domains were administered. Respondents spent on average almost three minutes on each PSTRE item, and only slightly more than one minute on each literacy or numeracy item.

Differences in the time spent on different modules are smaller. In fact, as the literacy and numeracy modules included a larger number of items, respondents spent more time on those modules (about 22 minutes on average) than on PSTRE modules (19 minutes).

PSTRE items also stand out on a different dimension. On average, respondents did not provide an answer to 23% of the PSTRE items assigned to them. This was the case for only 5% of the literacy items and 3% of the numeracy items. This is partly due to the fact that the concept of "missing answer" is not well defined for PSTRE items, given that, in most cases, the aim was not to provide an answer but to reach a specific stage within a simulated environment.

Figure 3.1. Average time spent on items, in seconds



Note: Average time spent on items in different domains, in seconds. The average is taken over all countries that administered all three domains and therefore excludes France, Italy and Spain.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959434

In light of the particularities of the PSTRE assessment, the rest of the chapter focuses on timing indicators for literacy and numeracy items only. This makes it possible to retain in the sample the three countries that decided not to administer the PSTRE assessment

(Italy, France and Spain). In order to make more homogeneous comparisons, the analysis excludes individuals who were assigned either of the two PSTRE modules. This is particularly relevant because the PSTRE modules could affect the overall duration of the assessment, but also because undertaking the PSTRE assessment in the first module could have an effect on behaviour in the second module.

This choice has, of course, a cost in terms of sample size. In the countries that administered PSTRE, about 60% of respondents were directed to at least one PSTRE module. However, the remaining sample size is still reasonably large, averaging 2 300 respondents across the countries that administered PSTRE.

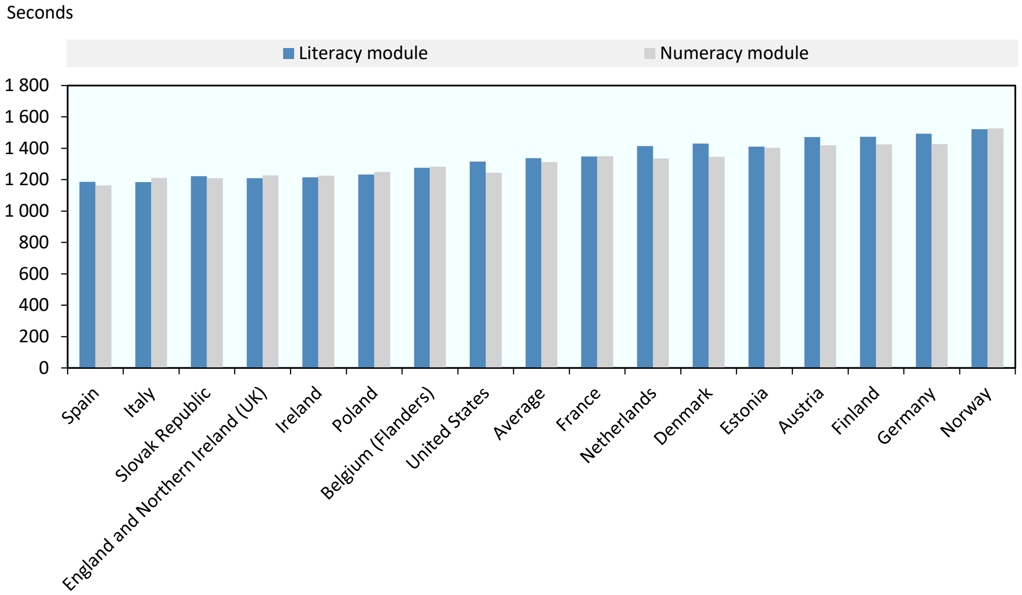
Cross-country differences in timing indicators

There is considerable cross-country variation in the time respondents spent on the literacy and numeracy assessments (Figure 3.2). Respondents in Norway spent almost 12 minutes more than those in Spain, a difference of about 30%.

Differences between domains were much smaller. In Figure 3.2, countries are ranked according to the overall time spent on the literacy and numeracy modules, but the ranking would be practically identical (with only a few minor switches) if it were based on either the literacy or the numeracy module.

In most countries, the literacy module took more time than the numeracy module, but the average difference is only 25 seconds. In England / Northern Ireland (United Kingdom), France, Ireland, Italy, Norway and Poland, respondents spent more time on the numeracy assessment, but the differences were again small, never exceeding 30 seconds.

Figure 3.2. Total time on task in literacy and numeracy

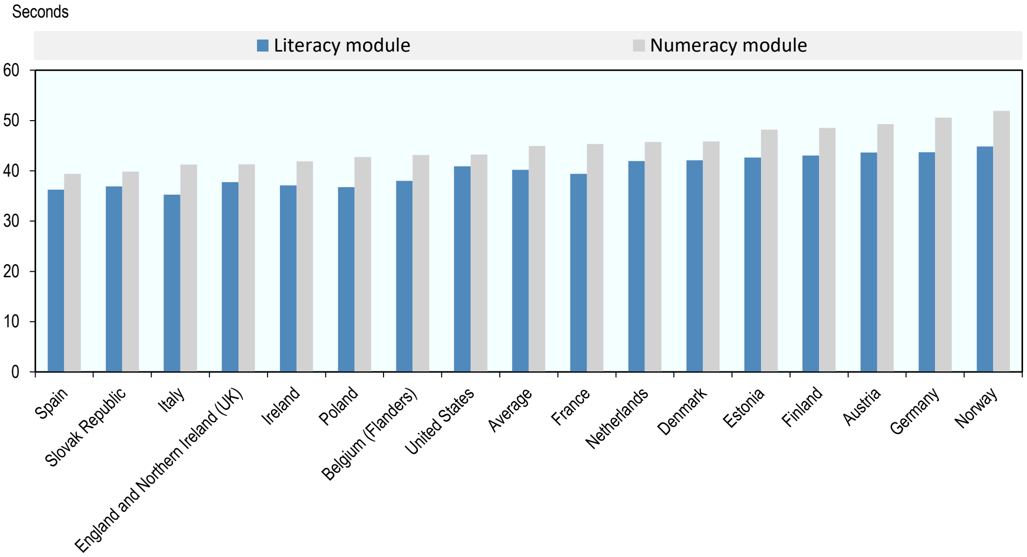


Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959453

A similar picture emerges when comparing the average time to first interaction (time elapsed between display of the item and respondent's first interaction with the item) (Figure 3.3).

Figure 3.3. Average time to first interaction in literacy and numeracy items



Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

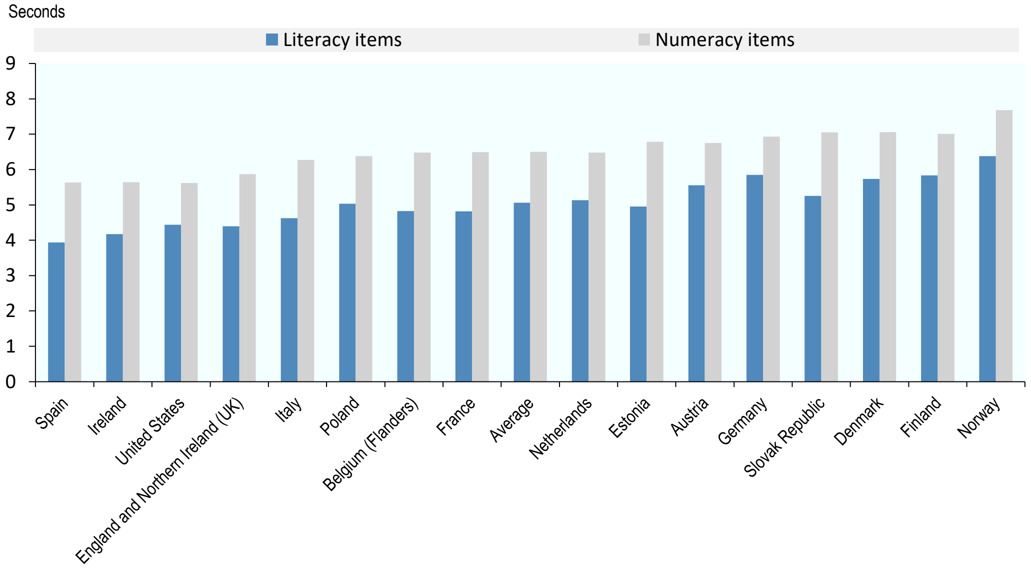
StatLink http://dx.doi.org/10.1787/888933959472

In Figure 3.3, countries are ranked according to time to first interaction on numeracy items, but the ranking would change only marginally if they were sorted according to contact time on literacy items. The ranking is consistent with Figure 3.2, with Austria, Finland, Germany and Norway at the top of the distribution and Italy, the Slovak Republic and Spain at the bottom.

The major difference between time to first interaction and overall time on task is that the time to first interaction is typically greater for numeracy items than for literacy items. In other words, respondents tend to spend more time on literacy items overall, but the time before their first observed interaction with items is greater in numeracy than in literacy. This is likely to be related to the different format of items in the two domains. Literacy items often require the respondent to move to a different page to examine the content of the item and give an answer, meaning that the first interaction (i.e. going to another page) is triggered very quickly. In most numeracy items, the respondent has immediate access to all the information needed to answer the question. The time of first contact may, therefore, often coincide with the time at which the respondent inputs an answer.

This is also true for time since last action, defined as time elapsed between the last interaction the respondent had with an item and the moment in which he/she decided to move to the following item (Figure 3.4). Before finally validating their answer, respondents appear to spend more time reflecting on their answers for numeracy items than for literacy items.

Figure 3.4. Average time since last action in literacy and numeracy items



Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

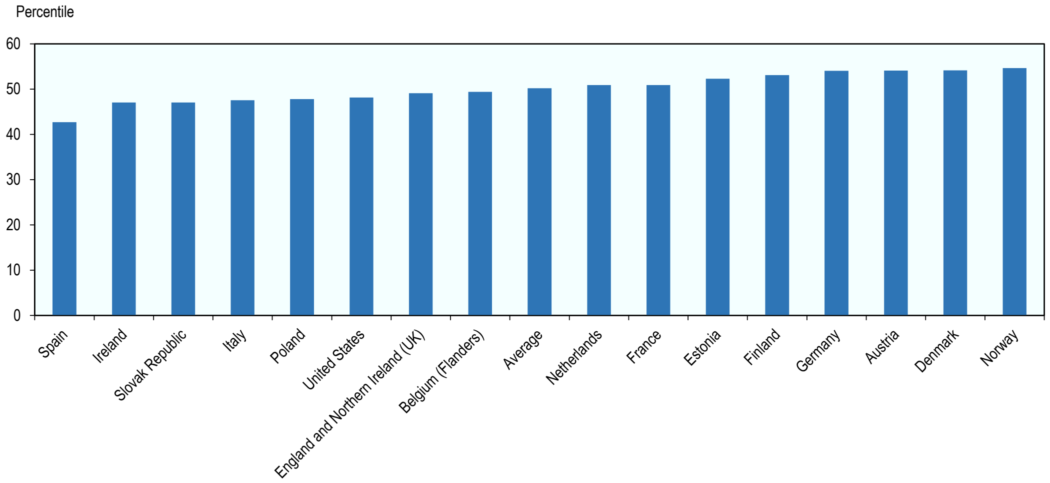
StatLink http://dx.doi.org/10.1787/888933959491

Ranking countries on the basis of time since last action delivers a slightly different picture than rankings based on the other indicators. Ireland and the United States are now at the bottom of the ranking, along with Spain, while Denmark and Finland are now at the top, together with Norway.

However, comparing countries on the basis of timing indicators based on absolute measures of elapsed time (in this case, seconds) is not necessarily appropriate. The adaptive nature of the assessment means that different respondents were assigned different items. In particular, respondents with higher estimated proficiency (based on their responses to items in the background questionnaire and their performance in the first stage of the assessment) were assigned more difficult items that could require more time to complete.

To make comparisons that control for differences in the items assigned to different individuals, it is possible to compute, for each individual and for each item to which he/she was assigned, a position (expressed in percentile) in the overall distribution of the timing indicator, and then to average these percentiles across items. For example, a respondent in a given country could be at the 40th percentile of the overall time on task distribution on one item and at the 60th percentile on the overall time on task distribution for a different item. In that case, the average position across the two items would be the 50th percentile (Figure 3.5).

Figure 3.5. Average position in the distribution of timing indicators



Note: Average percentile rank on the three timing indicators, across literacy and numeracy items. The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.

Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959510

In Figure 3.5, countries are ranked according to the average value of the three indicators. Overall, there is a close correlation between the values of different indicators within countries. Time since last action is the indicator that tends to deviate most from the general pattern observed, especially in countries such as Germany, the Slovak Republic and Spain.

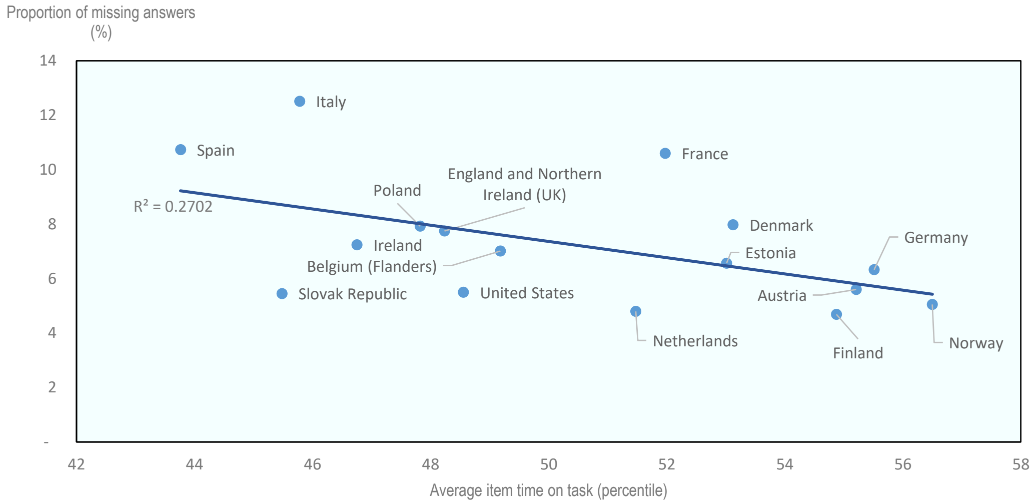
When it comes to ranking countries, there is little difference between using percentile-based indicators and indicators expressed in elapsed time in seconds. The partial exception is time since last action, where country rankings based on seconds vary more markedly from rankings based on percentiles. As shown in Chapter 4, time since last action is less related to item characteristics. Even in the case of time since last action, however, the correlation between the percentile-based indicator and the indicator expressed in elapsed time in seconds is very high, 0.78 for literacy items and 0.73 for numeracy items.

Timing indicators, missing answers and performance

A logical question following the analysis conducted so far is the relationship between the time spent on the assessment and the performance of respondents in literacy and numeracy. In performing such an analysis, however, a crucial element to take into account is the incidence of missing answers.

As explained above, no attempt is made in this chapter to identify the various reasons that lie behind a missing answer. In other words, no effort is made to distinguish between cases in which the respondent made the effort to solve an item (by spending time on it) and cases in which the respondent skipped the item without spending much time on it (due to lack of motivation or lack of confidence or because the test-taker felt that the item was out of reach and not worth spending time on). The fairly strong negative correlation between time spent on items and the proportion of items with missing answers suggests that the latter form of behaviour (i.e. skipping an item without spending too much time on it) is more prevalent (Figure 3.6).

Figure 3.6. Time spent on the assessment and incidence of missing answers



Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959529

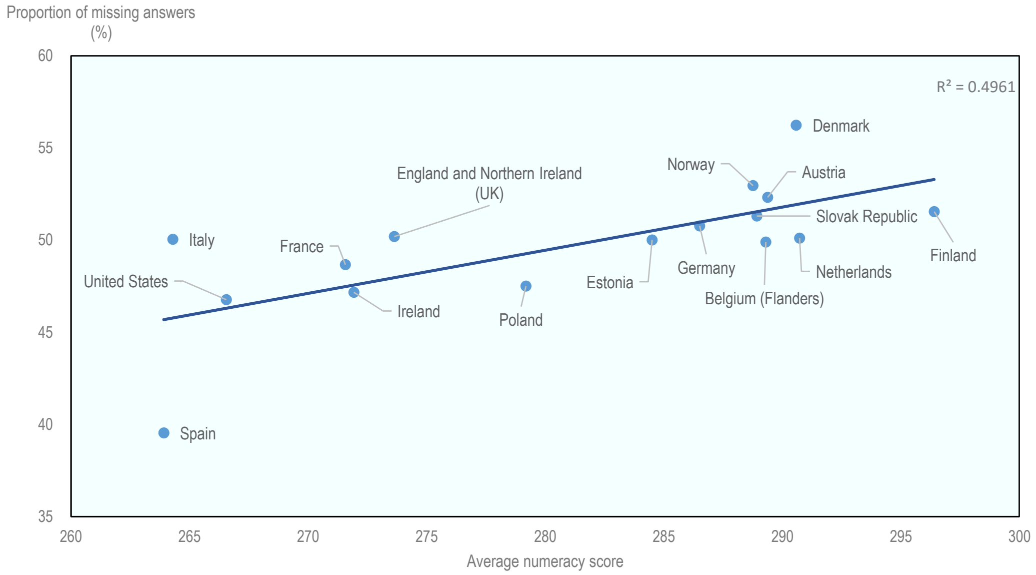
Countries in which respondents skipped a higher proportion of items are typically those in which respondents spent less overall time on the assessment. This result holds irrespective of the timing indicator used and irrespective of whether percentile-based measures are used, as opposed to raw time-based indicators.

Given the negative relationship between the incidence of missing answers and time spent on the assessment, it comes as no surprise that there is a positive relationship between overall performance and the various timing indicators. In other words, countries with higher literacy and numeracy scores are typically countries in which respondents spent more time on the assessment.

However, the relationship is not particularly strong, especially in the case of literacy. Among the various timing indicators, time since last action displays the strongest association with performance, in both assessment domains. For all indicators, the association is stronger for numeracy than for literacy.

To the extent that timing indicators capture some aspects of respondents' engagement and motivation, these results raise the question of how much impact the differences in motivation across countries have on performance on the assessment. These issues are explored more in depth in chapters 4 and 5.

Figure 3.7. Time since last action and performance in numeracy



Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959548

Socio-demographic distribution of timing indicators and missing answers

This section analyses differences according to the socio-demographic characteristics of respondents. It focuses on the subsample of individuals who only took the literacy and numeracy assessment, excluding respondents who took the PSTRE assessment for the reasons discussed above. In order to take into account country-specific factors associated with response time and patterns of missing answers, all statistics are computed at the country level and then averaged across countries.

The analysis groups respondents on the basis of standard socio-demographic characteristics, such as gender, age and education. Other distinctions are made according to three further criteria: 1) literacy and numeracy proficiency demonstrated during the assessment; 2) index of use of information and communications technology (ICT) in everyday life, to account for familiarity with the digital device on which the assessment was undertaken; and 3) time spent on the background questionnaire, to assess whether a consistent pattern emerges in terms of spending more or less time in both the direct assessment and in answering the background questionnaire.

The various subsections present raw differences, followed by adjusted differences estimated by means of regression analysis. This type of analysis, while still insufficient to provide evidence of a causal link between the characteristic under investigation and the behavioural outcome of interest, makes it possible to take into account a number of observable factors that are associated with the characteristic under investigation.

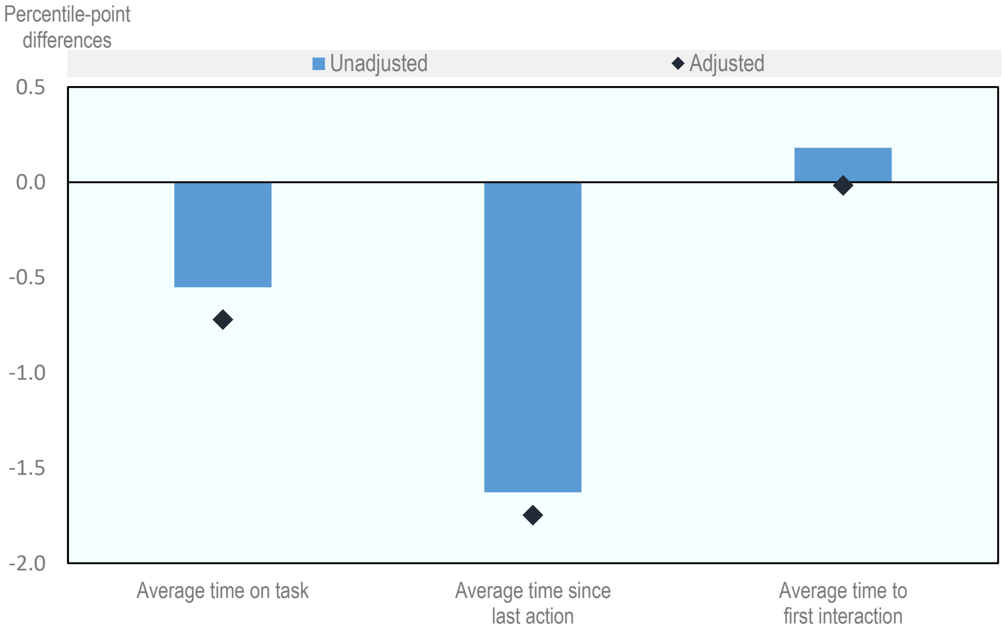
Gender differences

Gender differences in timing indicators are generally small, especially when compared to cross-country differences or to differences across groups defined by other socio-demographic characteristics. On average, women spent 41 seconds less than men on the literacy and numeracy assessment and were about 1 percentage point more likely to skip answering an item.

These results are robust to controlling for additional characteristics, such as age, education, employment status and familiarity with ICT. Estimated adjusted differences in time spent on the assessment increase to 53 seconds, while differences in the probability of skipping an answer remain at 1 percentage point.

Time to first interaction is virtually identical for women and men. Women spend slightly less time than men between the last interaction with the item and the final confirmation of their answer.

Figure 3.8. Gender differences in timing indicators



Note: The figure shows differences in timing indicators between men and women (men being the reference category). Adjusted differences account for age, education, employment status, familiarity with ICT, time spent on the background questionnaire, presence of another person and readiness to learn. The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

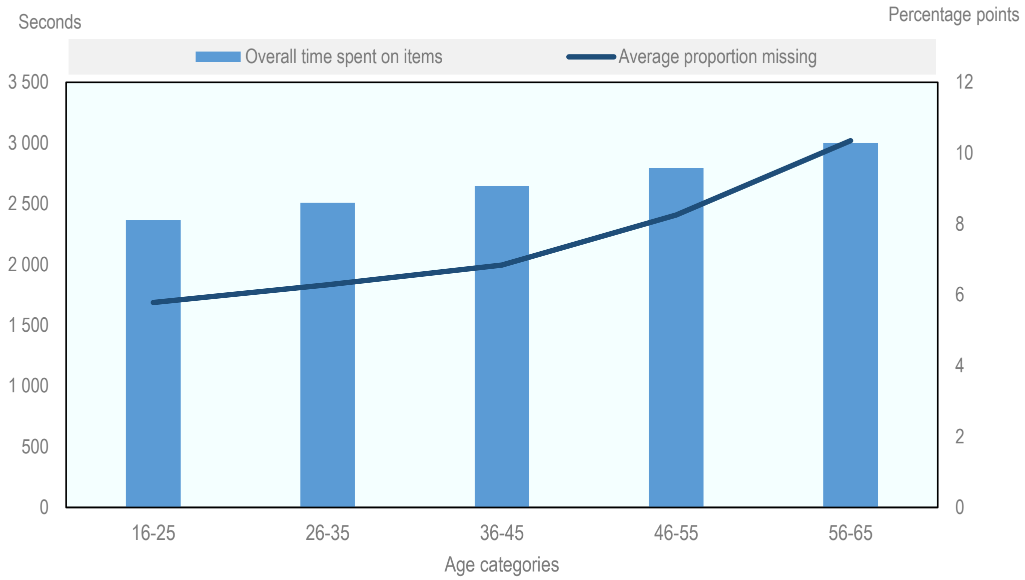
StatLink http://dx.doi.org/10.1787/888933959567

Age differences

Age differences in timing indicators are more pronounced than gender differences. All the timing indicators increase in a linear fashion with age. For instance, the overall time on task in the entire assessment is, on average, 39 minutes for respondents aged 16 to 25, increasing to almost 50 minutes for respondents aged 55 to 65.

A similar pattern is observed when looking at the proportion of items with missing answers, which is as low as 5% for the youngest respondents in the sample and as high as 10% for the oldest respondents (Figure 3.9). Importantly, these differences are not removed by controlling for other factors, such as education or familiarity with ICT.

Figure 3.9. Assessment time and missing answers in different age groups

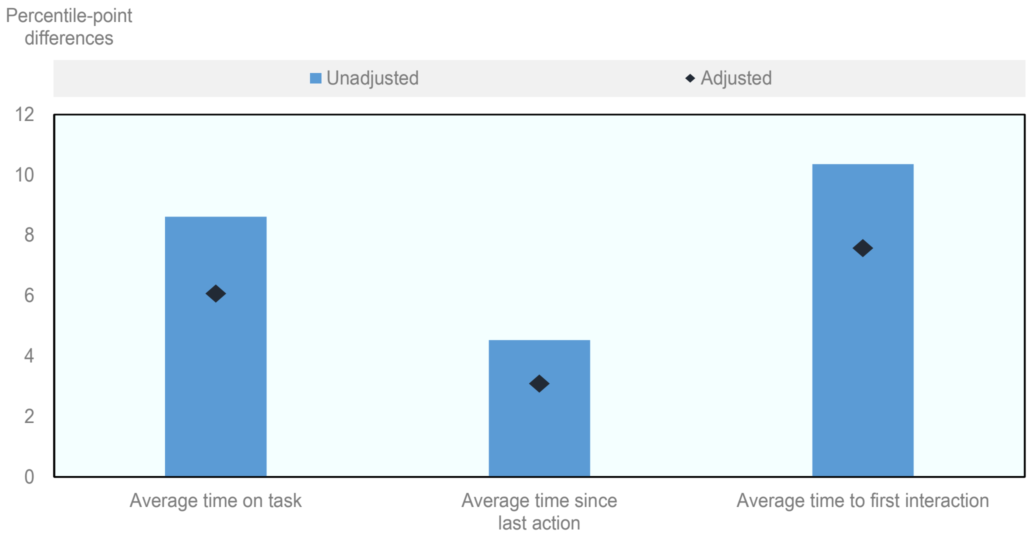


Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959586

A similar pattern emerges when analysing the other timing indicators, such as time to first interaction and time since last action. A summary measure of age effects (i.e. the difference between respondents aged 55-65 and respondents aged 25-34) is presented in Figure 3.10.

Figure 3.10. Age differences in timing indicators



Note: Difference between respondents aged 55-65 and 25-34 (age 25-34 being the reference category). Adjusted differences account for education, employment status, familiarity with ICT, gender, time spent on the background questionnaire, presence of another person and readiness to learn. The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959605

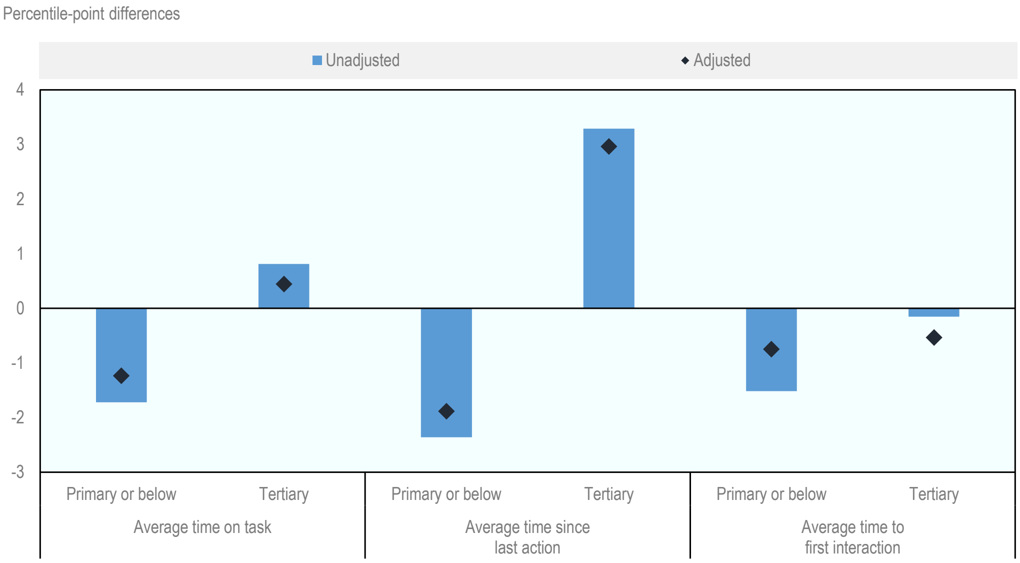
Education differences

In terms of overall time on task in the literacy and numeracy assessment, differences by highest level of education are also relatively large. Respondents whose highest attainment is an upper secondary degree took about 2.6 minutes more than respondents whose highest level of attainment is primary education or below, and respondents with a tertiary degree spent 1.7 minutes more than respondents with an upper secondary degree. The magnitude of these differences is only marginally reduced after taking account of the usual set of observable characteristics.

When looking at the share of missing answers, however, the pattern is reversed (contrary to what happens in the case of age-related differences). The share of items which were not answered was as high as 9% for respondents with primary education, 7% for respondents with upper secondary education and as low as 4% for respondents with tertiary education.

Figure 3.11 shows that tertiary-educated respondents tend to take slightly more time on the timing indicators analysed and primary-educated respondents a bit less time than the reference group of respondents with upper secondary education. Time to first interaction is a partial exception, as tertiary-educated adults do not differ from those with upper secondary education. But the differences are small in magnitude and, for the most part, they become even smaller after controlling for observable characteristics.

Figure 3.11. Differences in timing indicators, by level of education



Note: The reference category is respondents who attained upper secondary education. Adjusted differences account for age, employment status, familiarity with ICT, gender, time spent on the background questionnaire, presence of another person and readiness to learn. The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959624

Differences related to familiarity with ICT

A certain degree of familiarity with ICT devices and applications was a requirement for participation in the computer-based version of the PIAAC assessment. Adults who proved not to possess a minimum level of ICT proficiency were directed to the paper-based version of the assessment.

Familiarity with ICT was not part of the skills PIAAC intended to test as part of the literacy and numeracy assessment. In other words, a respondent's ICT skills are not supposed to exert a direct effect on his/her performance in the literacy and numeracy assessment. On the other hand, familiarity with ICT could well affect the speed at which respondents manage to solve the items (which is, by the way, an argument in favour of not incorporating timing information in the scoring of PIAAC).

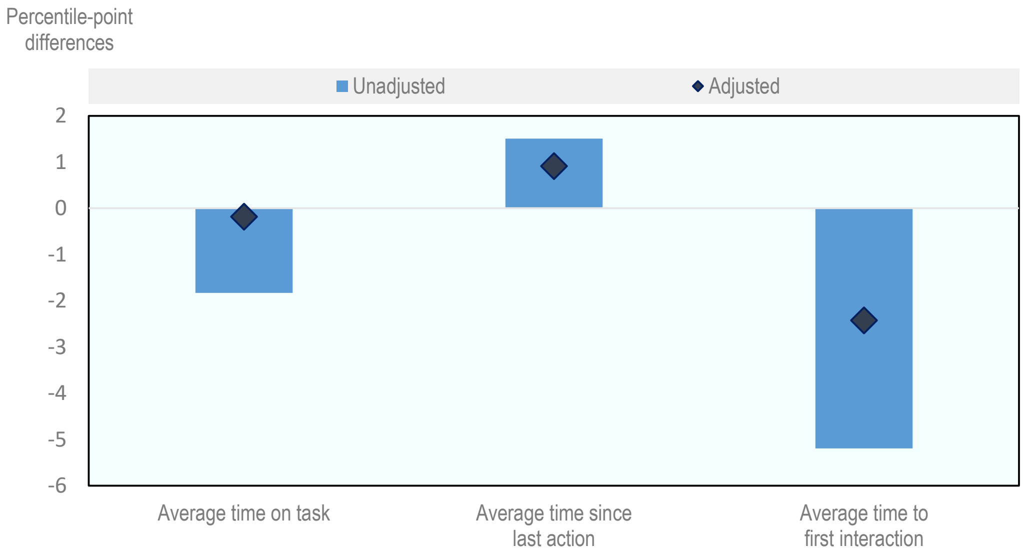
The PIAAC background questionnaire contained a number of questions on the frequency at which respondents undertake various activities with ICT. From such questions, it is possible to construct a scale of ICT use.

Figure 3.12 shows the gap in the various timing indicators according to the index of ICT use, comparing respondents who scored in the top quartile of the scale with respondents who scored in the bottom quartile.

On average, respondents who are more familiar with ICT spend less time on items, although the difference shrinks to zero once account is taken of differences in other observable characteristics. Similarly, familiarity with ICT is associated with less time to first interaction, although the gap is halved after controlling for observable characteristics.

However, respondents who are more familiar with ICT appear to take more time before confirming their final answer (time since last action), with an adjusted gap of about 1 percentile point.

Figure 3.12. Differences related to ICT familiarity



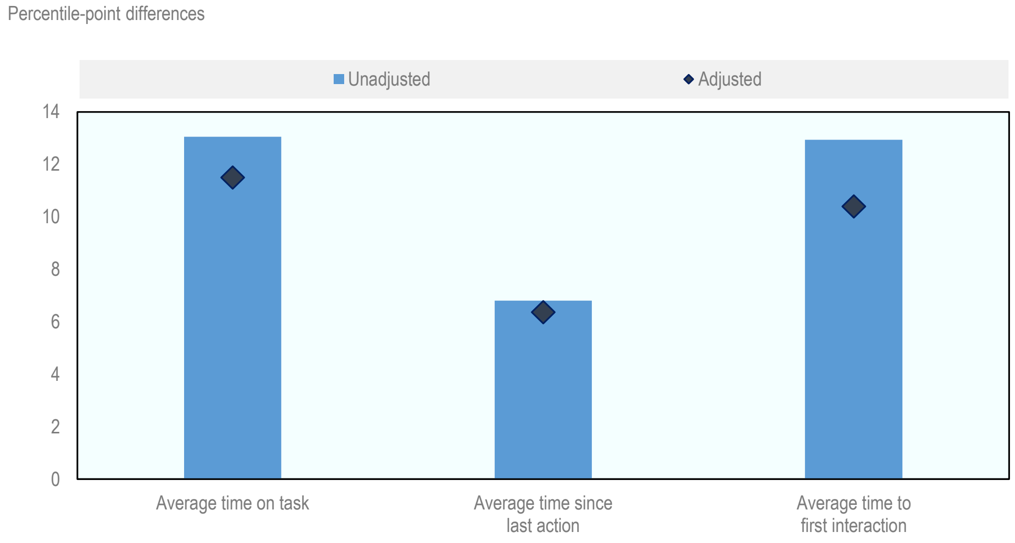
Note: Differences between respondents who are in the top and bottom quartile of the distribution of the index of familiarity with ICT. Respondents in the bottom quartile are the reference category. Adjusted differences account for age, employment status, gender, time spent on the background questionnaire, presence of another person and readiness to learn. The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959643

As timing indicators are also available for the background questionnaire, it is interesting to see to what extent there is a relationship between the time spent on these two components of the assessment.

An important difference between the background questionnaire and the assessment should be kept in mind. The background questionnaire was administered as a computer-assisted-personal-interview (CAPI). This means that an interviewer went through the various questions with the respondent and entered the answers on the computer, while the respondent was in control of the computer during the direct assessment. As a result, the timing indicators for the background questionnaire capture not only the speed at which the respondent gave his/her answers, but also the speed at which the interviewer asked the questions and filled in the answers.

Figure 3.13. Time spent on the background questionnaire and on the direct assessment



Note: Differences between respondents in the top and bottom quartile of the distribution of time spent in Section I of the background questionnaire. Respondents in the bottom quartile are the reference category. Adjusted differences account for age, employment status, familiarity with ICT, gender, presence of another person and readiness to learn. The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959662

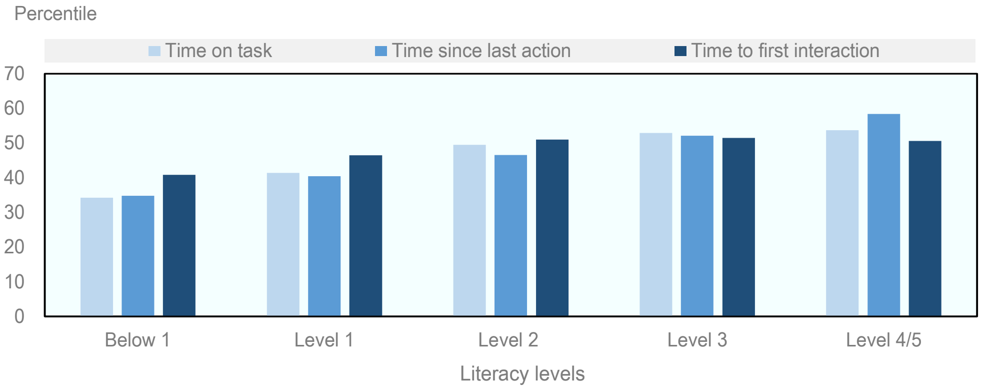
Figure 3.13 compares respondents in the top and bottom quartile of the distribution of time spent on Section I of the background questionnaire (the only section that was administered in its entirety to all respondents). The different indicators are related. Respondents who spent more time on Section I in the background questionnaire also tended to spend more time on the assessment. Adjusting for individual observable characteristics changes the size of the gap only marginally. It should be noted, however, that the relationship between time spent on the assessment and time spent on the background questionnaire is rather weak. Respondents separated by 50 percentile points in the distribution of time spent on Section I are separated on average by between 7 and 13 percentile points (depending on which indicator is analysed) in the distribution of time spent during the assessment.

Differences related to proficiency

The relationship between timing indicators and performance on the assessment (captured by the final proficiency scores) is difficult to interpret, because the two are intimately related. More skilled individuals are likely to need less time to solve assessment items, but the adaptive nature of the test implies that more skilled respondents are assigned more difficult (and probably longer) items. Moreover, causality could run in both directions, as spending more time on an item is likely to increase the chance of giving a correct answer.

While these issues are discussed and analysed more formally and in greater depth in Chapter 4, Figure 3.14 summarises the descriptive evidence on the relationship between timing indicators and literacy proficiency.

Figure 3.14. Timing indicators and literacy proficiency



Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[1]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959681

Consistent with the country-level analysis presented in Figure 3.7, performance in literacy is positively related to time spent on the assessment, irrespective of which indicator is analysed. Respondents scoring at Level 4 or Level 5 position themselves about 10 percentile points higher in the distribution of time to first interaction, and about 20 percentile points higher in the distribution of time on task and time since last action, than respondents scoring below Level 1.

Conclusions

This chapter has analysed three timing indicators (time on task, time since last action, and time to first interaction), as well as the share of missing answers, at both country and individual levels.

At the country level, the three indicators (with the partial exception of time since last action) deliver a rather consistent picture across assessment domains. In particular, respondents in countries like England/Northern Ireland (United Kingdom), Italy and Spain consistently spent much less time than respondents in Austria, Denmark, Finland, or Norway.

Not surprisingly, time spent on the assessment was found to be negatively correlated with the proportion of missing answers and positively correlated with performance on the assessment.

At the individual level, it was found that older respondents tended to spend more time on the assessment, even though they displayed a higher propensity to skip items without giving a response.

More educated adults typically spent more time on the assessment than those with less education and also displayed a higher propensity to provide answers to the items. However, the differences are rather small in magnitude, as are differences related to respondents' gender.

Respondents who have greater familiarity with ICT tended to take more time before confirming their final answer. On overall time on task, they tended to be faster, although the gap disappears after controlling for observable characteristics. Familiarity with ICT is also associated with faster time to first interaction.

Finally, the data show a positive relationship between the timing indicators and literacy proficiency, although it is particularly hard in this case to assess the direction of causality.

Note

[1] For different approaches that exploit the richness of interactions contained in PSTRE items using techniques borrowed from text-mining analysis, see He and von Davier (2015[2]), He and von Davier (2016[4]) and He, Borgonovi and Paccagnella (forthcoming[3]).

References

He, Q., F. Borgonovi and M. Paccagnella (forthcoming), "Using process data to understand adults' problem-solving behaviours in PIAAC: Identifying generalised patterns across multiple tasks with sequence mining"*, OECD Education Working Papers*, OECD Publishing, Paris. [3]

He, Q. and M. Von Davier (2015), "Identifying feature sequences from process data in problem-solving items with n-grams", in van der Ark, L. et al. (eds.), *Quantitative Psychology Research The 79th Annual Meeting of the Psychometric Society*, Springer, New York, NY, http://dx.doi.org/10.1007/978-3-319-19977-1_13. [2]

He, Q. and M. von Davier (2016), "Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment", in Y. Rosen, S. Ferrara and M. Mosharraf (eds.), *Handbook of Research on Technology Tools for Real-World Skill Development*, Hershey, PA: IGI Global, http://dx.doi.org/10.4018/978-1-4666-9441-5.ch029. [4]

OECD (2017), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, GESIS Data Archive, Cologne, http://dx.doi.org/10.4232/1.12955. [1]

Chapter 4. Allocation of time to different items in the Survey of Adult Skills

This chapter analyses disaggregated data at the respondent-item level to illustrate how respondents chose to allocate time to different items. Time spent on items was found to be strongly related to intrinsic characteristics of items, such as difficulty. Respondents devoted considerably less time to items administered in the second half of the assessment. This was accompanied by a decrease in performance (measured by the fraction of items answered correctly) and an increase in the proportion of missing answers. Respondents seem to allocate time to tasks rationally, spending less time on items that are both too difficult and too easy and more time on challenging items for which the probability of success is close to 50%. Spending more time on an item appears to increase the probability of giving a correct answer, although at declining rates.

Introduction

Chapter 3 analysed timing indicators for various groups of respondents related to the entire Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC) (hereafter referred to as "PIAAC") assessment. This chapter examines more disaggregated data at the respondent-item level, aiming to shed light on how respondents chose to allocate time to different items in the course of the assessment.

For this analysis, it is important to recognise that PIAAC was not designed as a timed assessment. Unlike other large-scale assessments such as the OECD Programme for International Student Assessment (PISA) and many high-stakes testing situations, there was no limit set on the amount of time that respondents could take to complete the assessment.

This feature of the design of the PIAAC assessment must be kept in mind when interpreting the results of the empirical analysis. The absence of an explicit time limit means that respondents who want to maximise their performance on the test are not subject to time constraints defined by the test protocol. In the abstract, they can take as much time as they need to maximise performance. The choice of how much time to allocate to different items becomes meaningful only when the analyst makes the assumption (reasonable in these circumstances) that time has a value for respondents (i.e. that doing the assessment is one of several alternative uses of their time). As a consequence, respondents in effect trade off the value they attach to their performance on the test with the value they attach to other uses of their time.

A second important aspect is that time on task represents an imperfect proxy of the effort exerted by respondents. This is because log files are silent about how respondents actually employ the time they spend on each item. While it is reasonable to think that the amount of time spent on an item is a good approximation of how much the respondent was engaged with the item, it is possible that the respondent spent a lot of time on a given item for other reasons (e.g. because he/she was interrupted, took a break or was distracted by different things or thoughts).

As noted in Chapter 2, log files provide no record of what respondents do in the course of the assessment and the many different ways they could show engagement with an item – see Maddox et al. (2018[1]) for a small scale observational study of respondents' behaviour during the PIAAC assessment.

Time on task, as measured in log files, is the result of the interaction of a variety of factors:

- respondents' cognitive ability
- respondents' engagement and motivation
- item characteristics
- external events (distractions or unforeseen events during the course of the assessment).

Each of these factors has a different relationship to time on task, and the relationship is often non-linear. In the case of cognitive ability, for instance, highly-skilled people are expected to solve items relatively rapidly. At the same time, low-skilled individuals are also expected to devote little time to difficult items, as they realise that they have low chances of solving them and will, therefore, skip them. Different item characteristics can affect the trade-offs that respondents face when deciding how much time to allocate to each item.

Item difficulty operates in a way similar to respondents' cognitive ability. Very easy items will generally take less time to solve. However, more difficult items might require more time to solve, or they might be so difficult that the optimal action is to devote as little time as possible to them – or to skip them entirely.

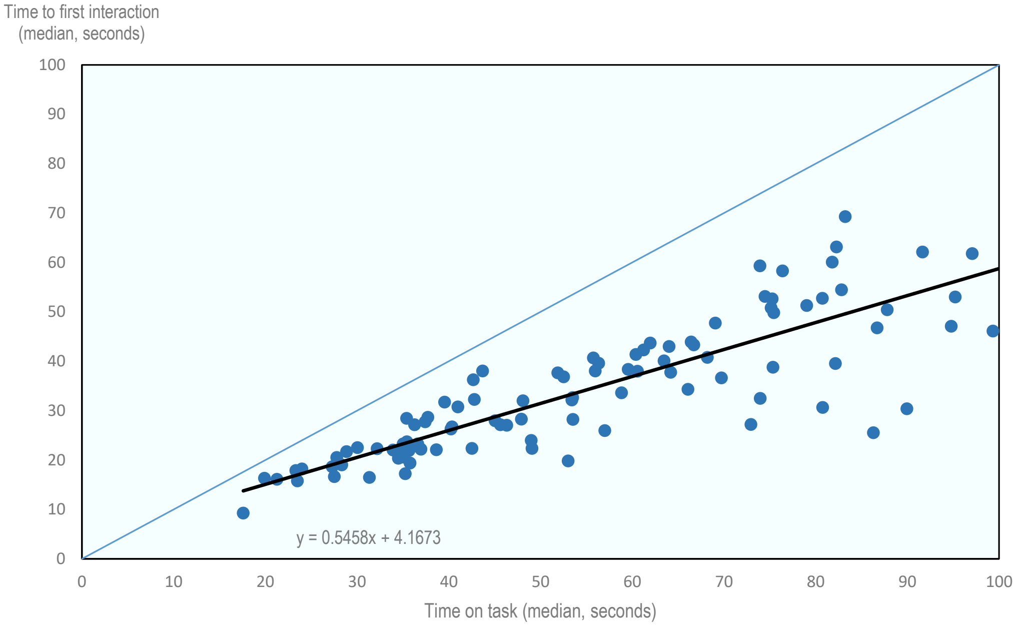
Item position can also have opposite effects on time on task. Respondents might become tired at later stages of the assessment and thus need more time to solve an item. But fatigue can lead to a decrease in motivation and could, therefore, reduce the time devoted to the item. A decrease in time on task at later stages of the assessment can, in principle, also be attributed to the fact that respondents learn test-taking strategies or become more familiar with the user interface. As a consequence, they become quicker to solve the items (although the latent ability that the assessment is meant to capture does not change).

It is hard to disentangle the separate impacts of all these factors, but the chapter provides some evidence in this respect, adopting two approaches. The first, at the item level, consists of relating the time spent on various items to a range of item characteristics. The second, at the individual level, looks more precisely at how respondents allocate time between items (this could be called a within-respondents / between-items analysis).

Timing indicators and item characteristics

The analysis in this section aggregates information at the item level, relating timing indicators with various item characteristics. In particular, it examines time on task (the overall amount of time spent on the item).

Figure 4.1. Correlation between time on task and time to first interaction

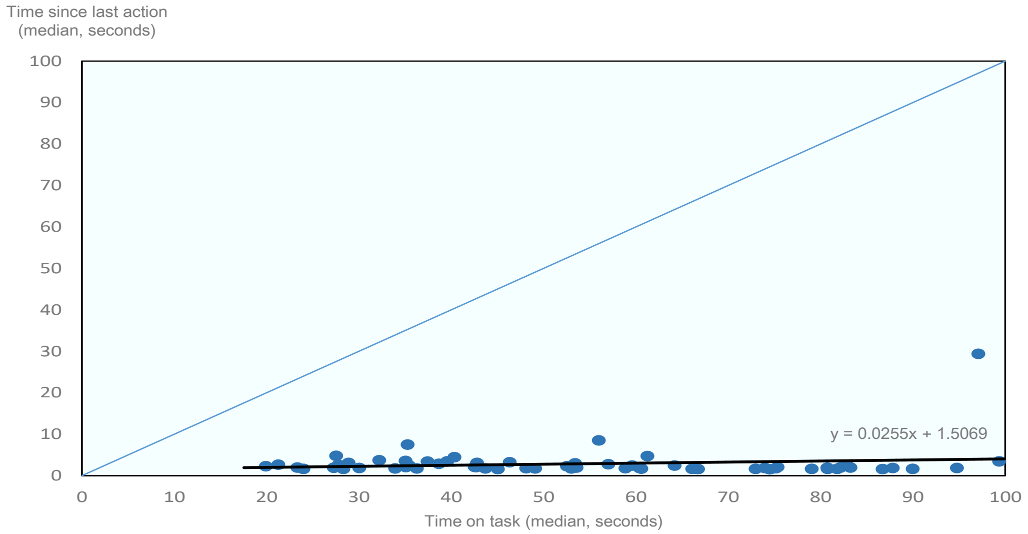


Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.

Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959700

Figure 4.2. Correlation between time on task and time since last action



Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959719

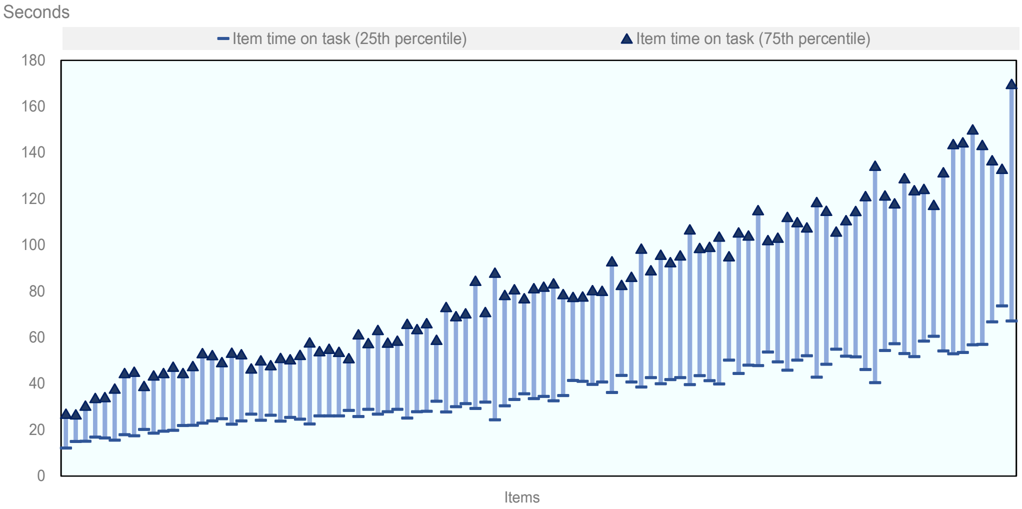
At the item level, time on task displays a strong correlation with time to first interaction (Figure 4.1), but a very weak correlation with time since last action (Figure 4.2). The main reason for this is that many items do not require multiple interactions between the respondent and the computer interface. Consequently, the first interaction and the last action essentially coincide. Time on task is, thus, the indicator that is easiest to interpret. It provides the best approximation of the effort respondents decide to allocate to an item.

Figure 4.3 shows the distribution of time on task for each literacy and numeracy item in the dataset. There is considerable variation both across and within items. For many items, the distribution of time on task is extremely compressed, with half of the respondents taking between 20 and 50 seconds. For other items, the distribution of time on task is much more dispersed, with the most rapid quarter of respondents spending at most 60 seconds (1 minute), and the slowest quarter spending almost 150 seconds (2.5 minutes).

Some of the within-item variation presented in Figure 4.3 is likely to be due to differences across countries in the average time spent on the assessment. Figure 4.4 shows this for a selection of items (and for a selection of countries, in order to preserve the readability of the graph). Respondents in Finland and Norway, for example, spend consistently more time on each item than respondents in Italy and Spain.

There are, however, no major differences between countries in the degree of variability of time spent on each item (Figure 4.5). Spain is the only outlier in this respect. In other words, for each item, the distribution of time on task appears to be shifted to the right or to the left depending on the country, but the spread of the distribution displays much less variation across countries.

Figure 4.3. The distribution of time on task, by item

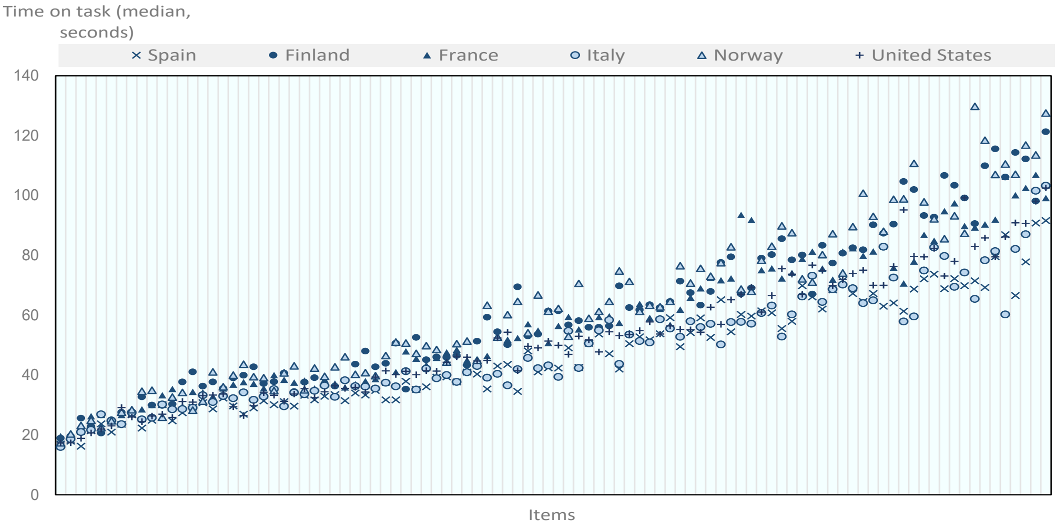


Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.

Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959738

Figure 4.4. Median time on task, by item and country

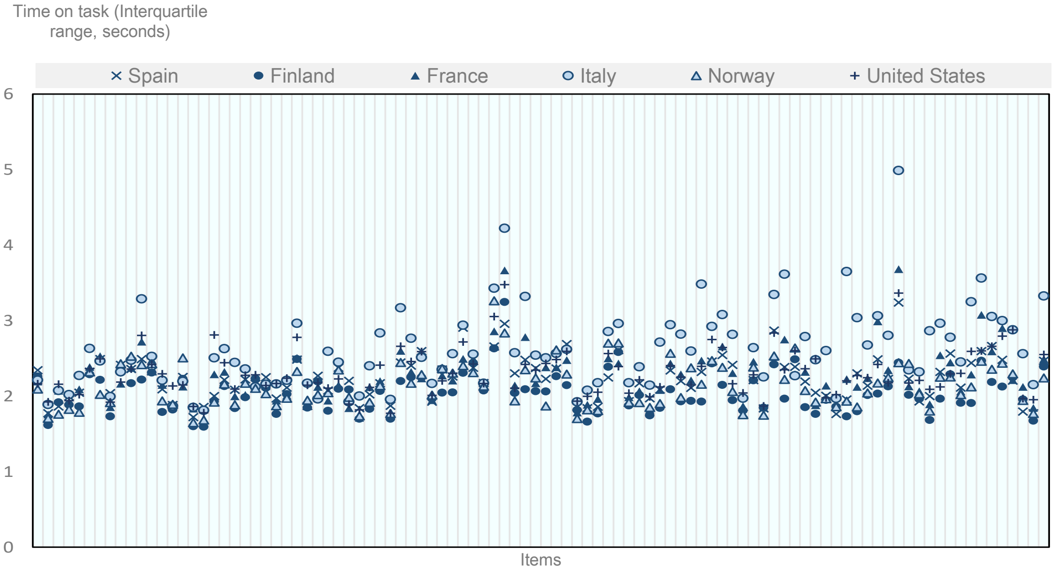


Note: Items are sorted by overall average time on task. The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.

Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959757

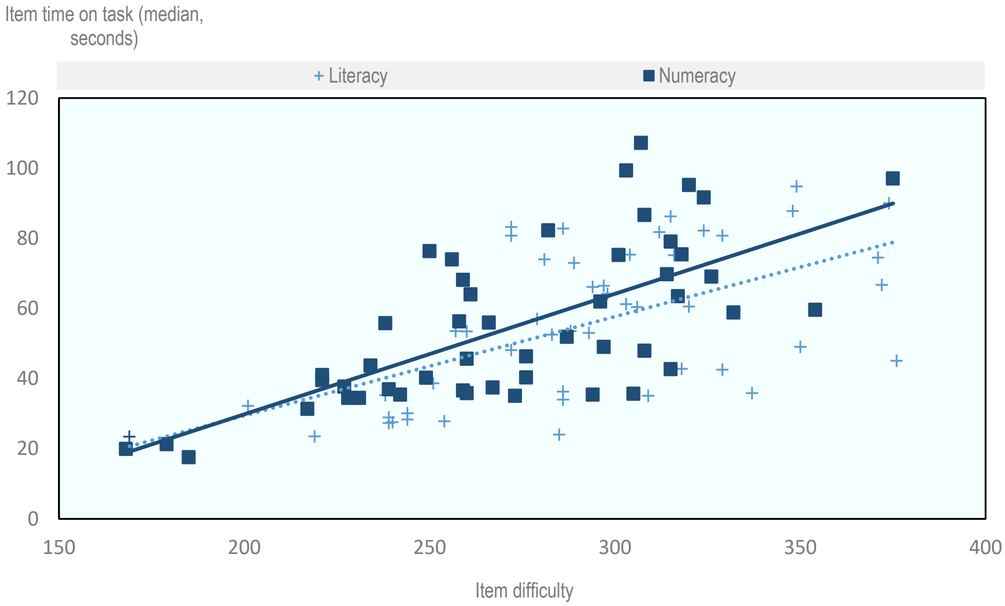
Figure 4.5. Interquartile range of time on task, by item and country



Note: Items are sorted by overall average time on task. The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959776

Figure 4.6. Time on task and item difficulty



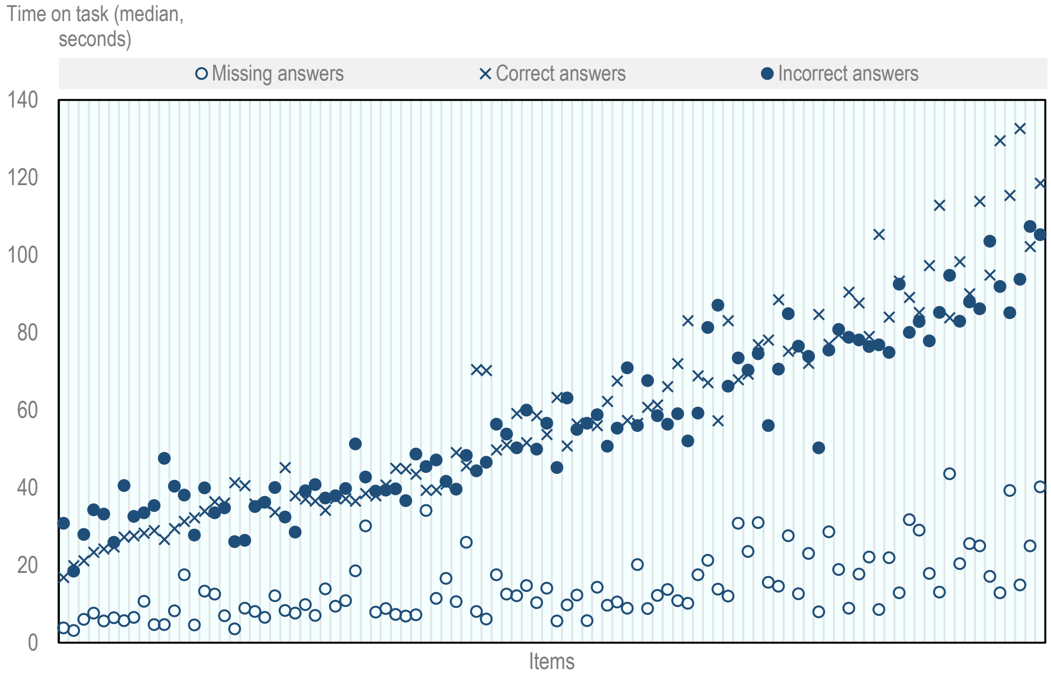
Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959795

An important factor that could explain between-item differences in time on task is item difficulty. Figure 4.6 clearly shows that time on task increases with item difficulty, for both literacy and numeracy items. This is partly because more difficult items involve more complex cognitive operations and more extensive stimulus materials.

It is also interesting to relate time on task to the final status of the item (i.e. whether the item was answered correctly, answered incorrectly or not answered at all). Figure 4.7 shows that, for any given item, respondents who gave the correct answer did not spend a significantly different amount of time than respondents who gave an incorrect answer. This is further indication that time on task is strongly related to intrinsic characteristics of the item. The variation in time on task is much more limited for items that were not answered, indicating that the time required to decide whether or not it is worth trying to solve the item does not increase as much with item difficulty as the time needed to actually solve the item.

Figure 4.7. Time on task, by answer category



Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.

Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959814

Finally, it is worth looking at time on task in relation to the position of the various items within the overall assessment. This provides a useful bridge to the following section, which examines in more detail individual behaviour in terms of allocation of time to items.

As explained in Chapter 2, PIAAC was designed as an adaptive test. One consequence is that the allocation of items to respondents was (conditionally) random. There are two main reasons why the same item could have been presented to respondents in a different position. First, respondents were randomly allocated to a literacy or numeracy module. Therefore, the same literacy item would be in a different position depending on whether the respondent took literacy as the first module or the second module. Second, within modules, respondents were allocated to different booklets. This allocation was only conditionally random, as the booklets varied in their level of difficulty and the allocation was based on observable characteristics of respondents that are likely to be correlated with ability. However, given that each item appears at most in two booklets, the main source of variation in the position of any given item is whether or not the item was part of the first or the second module.

In all countries, respondents tend to devote less time to the second module than to the first module. Table 4.1 illustrates this point with reference to time on task, but the same result is found when looking at other timing indicators.

Table 4.1. Time on task, by module

	First module	Second module	Difference	% Difference
Austria	1 529.6	1 358.7	-171.0	-11.2%
Denmark	1 486.2	1 293.7	-192.5	-13.0%
England/Northern Ireland (UK)	1 305.6	1 134.4	-171.2	-13.1%
Finland	1 538.9	1 363.9	-175.0	-11.4%
France	1 461.5	1 247.8	-213.7	-14.6%
Germany	1 550.7	1 365.8	-184.9	-11.9%
Ireland	1 328.9	1 110.7	-218.2	-16.4%
Italy	1 334.1	1 071.8	-262.3	-19.7%
Netherlands	1 437.5	1 310.6	-126.9	-8.8%
Norway	1 622.1	1 432.6	-189.5	-11.7%
Average	1 420.5	1 229.6	-190.8	-13.4%
Poland	1 335.4	1 144.9	-190.5	-14.3%
Slovak Republic	1 297.1	1 132.0	-165.2	-12.7%
Spain	1 273.6	1 075.1	-198.5	-15.6%
United States	1 385.5	1 173.0	-212.5	-15.3%

Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933960175

The decrease in the time devoted to the assessment is associated with a decline in performance. Items in the second module are less likely to be answered correctly and more likely to be left blank or skipped, as shown in Table 4.2.

Table 4.2. Correct and missing answers, by module

	Proportion of correct answers			Proportion of missing answers		
	First module	Second module	Difference	First module	Second module	Difference
Austria	0.6158	0.5971	-0.0187	0.039	0.066	0.027
Denmark	0.6270	0.5924	-0.0346	0.055	0.101	0.046
England/Northern Ireland (UK)	0.5920	0.5502	-0.0418	0.055	0.098	0.044
Finland	0.6864	0.6525	-0.0339	0.029	0.056	0.027
France	0.5593	0.5197	-0.0396	0.079	0.126	0.047
Germany	0.6140	0.5807	-0.0334	0.050	0.079	0.029
Ireland	0.5784	0.5328	-0.0455	0.048	0.097	0.049
Italy	0.5147	0.4835	-0.0312	0.097	0.149	0.052
Netherlands	0.6510	0.6343	-0.0168	0.034	0.059	0.024
Norway	0.6397	0.6124	-0.0272	0.039	0.063	0.024
Average	0.5962	0.5663	-0.0299	0.054	0.090	0.036
Poland	0.5831	0.5449	-0.0383	0.056	0.105	0.049
Slovak Republic	0.6047	0.5998	-0.0048	0.043	0.064	0.022
Spain	0.5166	0.4913	-0.0253	0.087	0.127	0.040
United States	0.5642	0.5367	-0.0275	0.039	0.068	0.029

Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933960194

In the literature on large-scale assessments, decline in performance in the course of the assessment is now a well-established result (Borgonovi and Biecek, 2016[3]; Brunello, Crema and Rocco, 2018[4]; Borghans and Schils, 2012[5]).

Timing information extracted from log files is important to better understand the mechanisms behind this established result. Past literature has attributed decline in performance during the test as evidence of lack of endurance or lack of motivation. But the decline in time allocated could also (at least partly) be due to a learning effect and to increased efficiency in answering the questions. The next section attempts to disentangle the two channels by examining whether the relationship between time on task and probability of success changes in the course of the assessment (with item position).

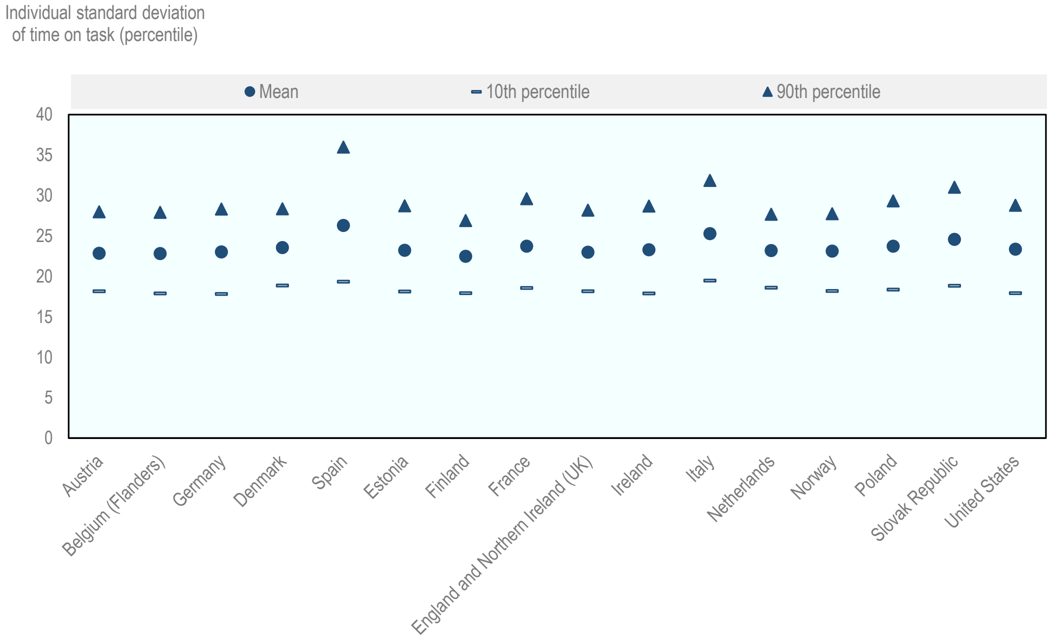
Time-allocation strategies

While the previous section took a predominantly item-level approach, this section focuses on individual respondents, looking at how they allocated time to the different items. Chapter 2 presented information on how the time allocated to the assessment varied across respondents. This deepens that analysis by looking at how time allocation interacts with item characteristics and how it varies during the course of the assessment.

A first question to address is whether respondents differ in the strategy they choose to allocate time between items. One way to answer this question is to compute, for each respondent, the percentile rank of the respondent in the distribution of time on task for each item presented to him/her. It is then possible to analyse the features of the individual-specific distributions of these percentile ranks. A compressed distribution indicates that the respondent adopted a consistent strategy, always devoting the same amount of time (relative to all other respondents who were assigned the same items) to all items in the assessment. On the other hand, a more dispersed distribution would characterise a respondent who spent an unusually large amount of time on some items and an unusually small amount of time on other items. The standard deviation of the individual distributions of percentile rank is used as a summary measure of the degree of dispersion.

Individual standard deviations can be aggregated by countries or by socio-demographic characteristics of respondents. The results are presented in Figure 4.8 and Figure 4.9. The average standard deviation is around 20 percentile points. This indicates a relatively large degree of individual heterogeneity: different respondents interact in different ways with the same item, with the result that the same respondent can be relatively fast on one item and relatively slow on other items. On the other hand, there is very little cross-country variation in this indicator, as shown in Figure 4.8. Similarly, Figure 4.9 shows very little variation across socio-demographic groups (note that the scale of Figure 4.9 ranges from -5 to +5 percentile ranks, while the scale of Figure 4.8 ranges from 0 to 40).

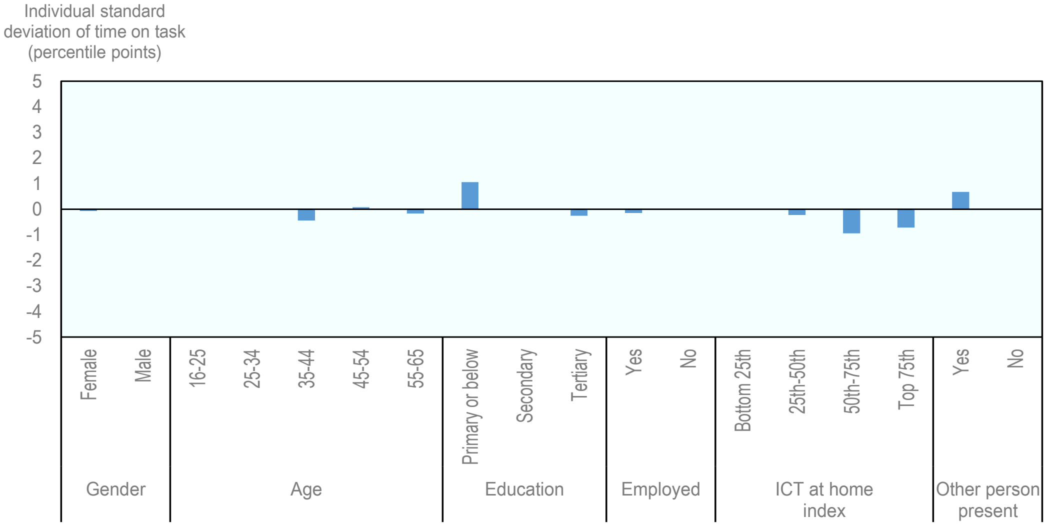
Figure 4.8. Individual distribution of time on task



Note: The figure shows moments of the within-country distribution of individual standard deviations in percentile ranks. The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959833

Figure 4.9. Standard deviations of percentile rank, by socio-demographic characteristics



Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules. Coefficients are jointly estimated in a participants-level regression model with individual standard deviation of time on task as a dependent variable.

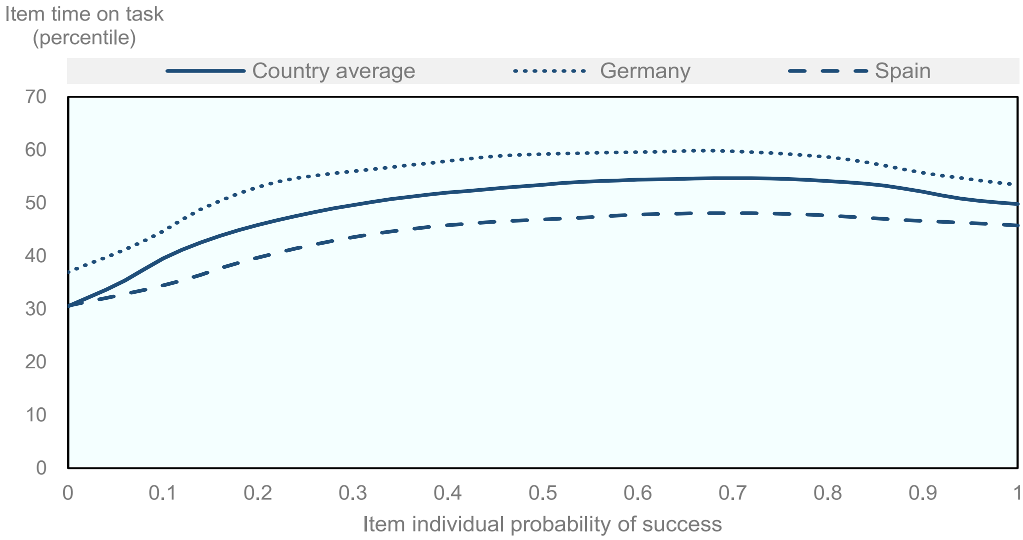
Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959852

The previous section showed a strong relationship between time on task and item difficulty at the item level. It is possible to undertake this analysis at the respondent level, by asking how individuals allocate time to items based on the individual-specific probability of success. While difficulty of a specific item is a fixed characteristic of the item, the ex ante probability of success is an indicator that simultaneously takes into account the difficulty of the item and the respondent’s ability. Indeed, the reported PIAAC proficiency levels are constructed on the basis of the models used to estimate items’ parameters and respondents’ final scores and are defined in terms of a probabilistic relationship between respondents’ skills and item difficulty. It is then possible to make statements such as “A respondent at Level 3 of the PIAAC proficiency scale is able to correctly answer an item of Level 3 difficulty with a probability of 67%.”

A rational individual who values his or her time should not devote too much time to questions that are too difficult, and which he/she is therefore very unlikely to be able to answer correctly. The adaptive nature of the PIAAC assessment makes these extreme situations less frequent. This is because, on average, items are targeted by construction to the expected ability of individual respondents. However, appropriate scaling also requires that some skilled respondents are assigned very easy items and some less skilled respondents are assigned very difficult items. Furthermore, given that the measure of ability used here to compute probability of success is only known at the end of the assessment, there is a good range of variation among individuals.

Figure 4.10. Time on task and relative probability of success



Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

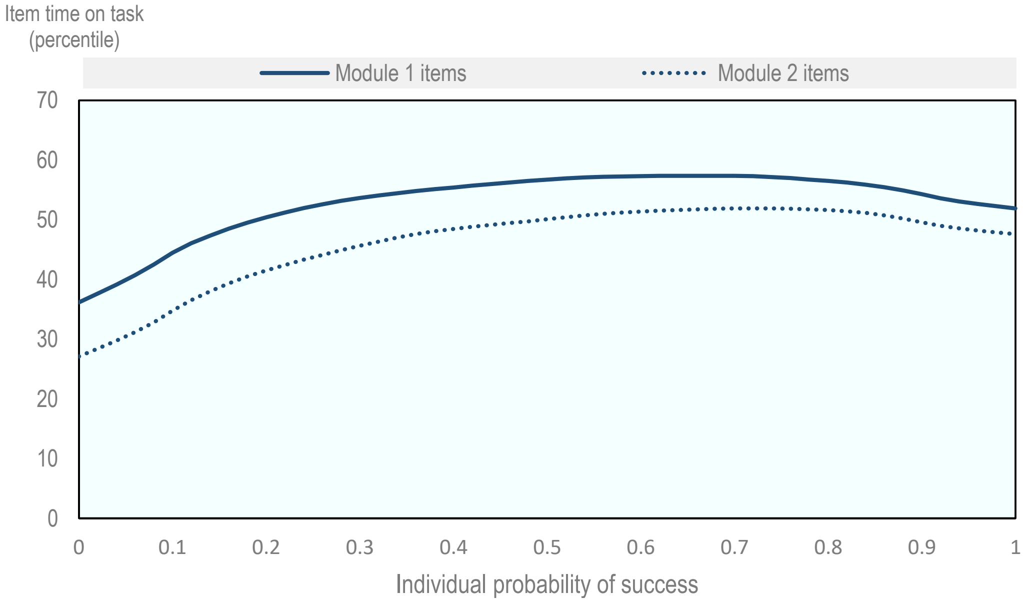
StatLink http://dx.doi.org/10.1787/888933959871

The pattern in Figure 4.10 is consistent with a priori expectations. As the item becomes excessively difficult, respondents devote less time to it (relative to other respondents who faced the same item). Time-on-task percentile also tends to decrease, although to a lesser extent, when the item is very easy. The decline in time on task is lower at the top than at the bottom end of the probability of success distribution, because respondents are more likely to skip difficult items (therefore devoting very little time to them). Easy items, on the other hand, necessarily take some time to answer correctly.

The shape of the relationship between time on task and probability of success does not seem to be affected by item position, as illustrated in Figure 4.11. The curve for the second module is simply shifted downwards, consistent with the fact that respondents spend less time on the second module than on the first module.

Figure 4.12 plots time on task on the individual probability of success depending on whether the final answer was correct, incorrect or missing. In the case of correct answers, there is no decline at the bottom end of the probability of success distribution, while there is a decline at the top end, as is the case for the overall sample. The opposite happens in the case of incorrect answers. Time on task declines as items get harder, although less than in the overall sample, because respondents did attempt to give an answer. At the top end, there is no decline in time on task, which is what one would expect when respondents fail to give a correct answer to an easy item. Items for which respondents did not provide an answer follow a pattern similar to the overall sample, but the distribution of time on task is shifted downwards, indicating that at some point the respondents decided to give up.

Figure 4.11. Time on task and probability of success, by module

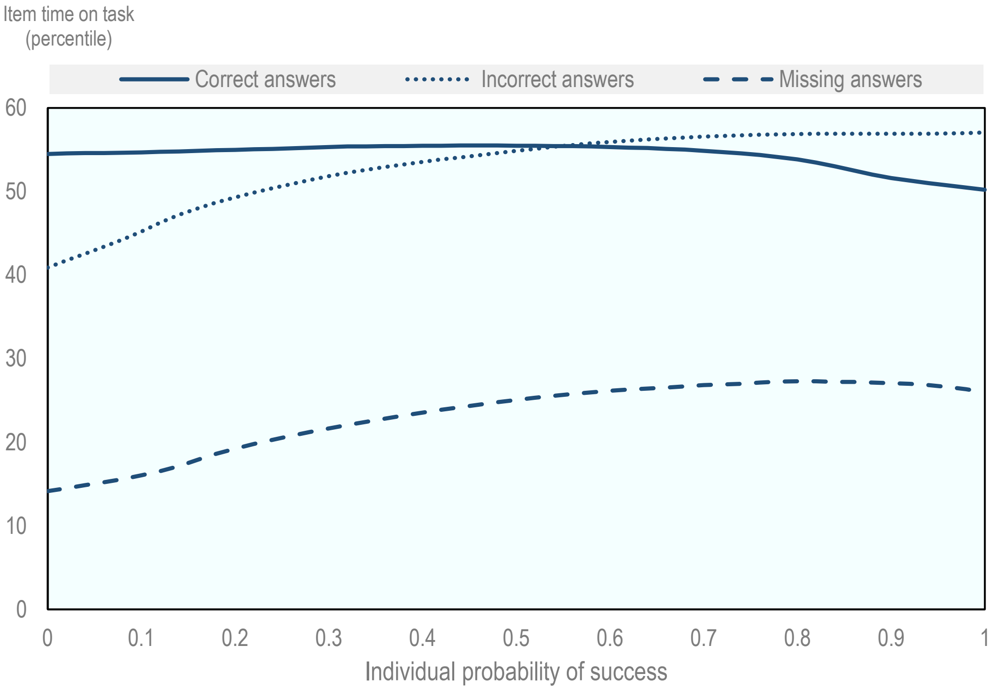


Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.

Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959890

Figure 4.12. Time on task and probability of success, by type of answer



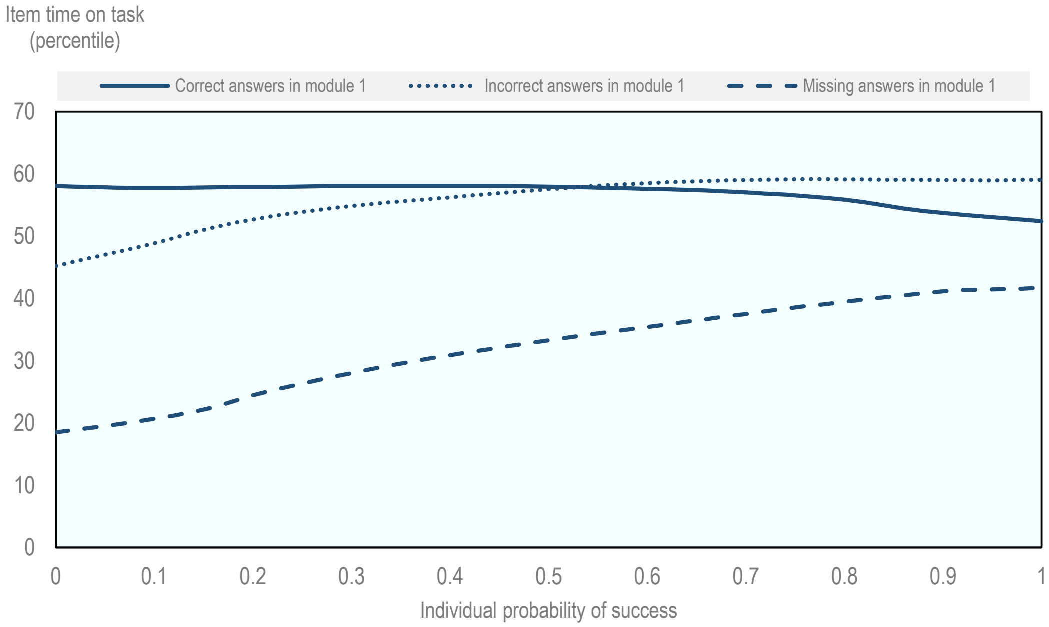
Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.

Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959909

Interestingly, the relationship between time on task and probability of success differs by module, but only for missing answers, as illustrated in Figure 4.13 and Figure 4.14. Moving from Module 1 to Module 2, the curves for correct and incorrect answers are simply shifted downwards, as in the case of Figure 4.11. For missing answers, the curve changes shape and becomes flatter. This means that, in Module 1, respondents spent a relatively larger amount of time before deciding to skip an easy item (i.e. an item with a large ex ante probability of success). In Module 2, decisions to skip easy items are taken much faster. On the other hand, there is no evidence that the increase in missing answers from Module 1 to Module 2 is concentrated in relatively easy or relatively difficult items.

Figure 4.13. Time on task and probability of success in Module 1

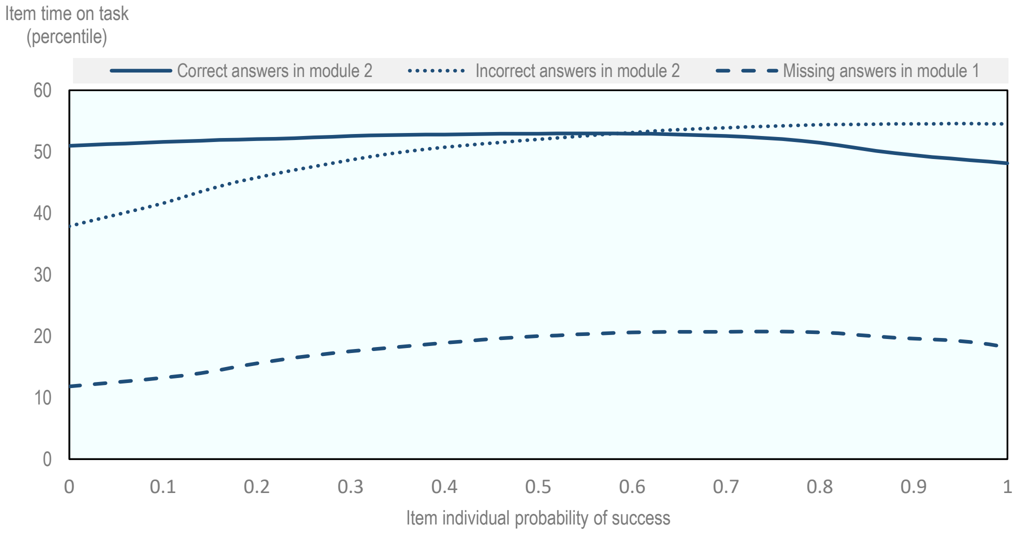


Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959928

More interestingly, and unexpectedly, the relationship between time on task and probability of success is unrelated to (self-reported) perseverance, which can be proxied by the answer to an item of the background questionnaire asking the respondent whether he/she “gets to the bottom of difficult things”.

Figure 4.14. Time on task and probability of success in Module 2

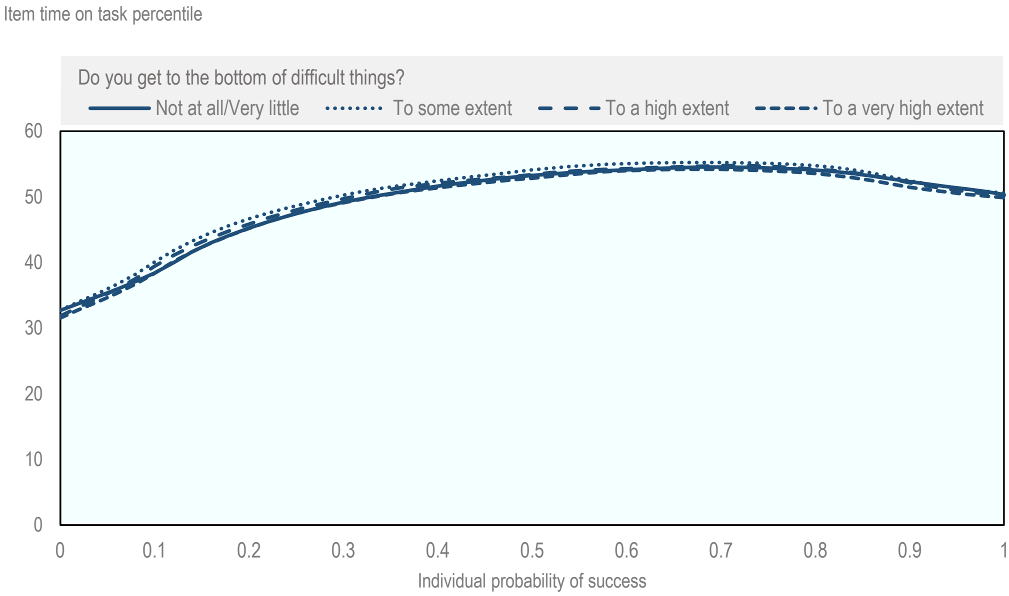


Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.

Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959947

Figure 4.15. Time on task and probability of success, by individual dispositions



Note: The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.

Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959966

The final section of this chapter looks at the relationship between time on task and actual performance on the assessment, measured by the probability of giving a correct answer to an item (Goldhammer et al., 2014[6]). This is not the same as the ex ante individual probability of success that was used in previous parts of the chapter. The ex ante individual probability of success is a measure of how difficult an item is for a respondent with a given ability level. The probability of answering an item correctly is the ex post realisation (i.e. a dummy variable taking a value of 1 if the respondent correctly answered an item and a value of 0 otherwise). No distinction is made between the absence of a response due to skipping and an incorrect answer.

The most straightforward way to investigate whether spending more time on an item actually increases the probability of giving a correct answer is through the following regression:

$$y_{ij} = f\left(ToT_{ij}\right) + \gamma_i + \delta_j + \varepsilon_{ij}$$

where y_{ij} is a dummy taking value 1 if individual i correctly answered item j, $f\left(ToT_{ij}\right)$ is a (quadratic) polynomial in the time on task spent by individual i in item j, γ_i is a respondent fixed effect (that controls for any fixed individual characteristic, like ability and average motivation) and δ_j is an item fixed effect (that controls for any characteristic of item j). ε_{ij} is a random error term.

The regression exploits the fact that the data contain information on a variety of respondents answering the same set of items. As a result, the regression compares the outcome of different individuals who allocated a different amount of time to the same item, controlling at the same time for any fixed characteristic of the respondent (thanks to the fact that the data contain information on the same respondent answering different items).

An alternative specification would replace the individual and item fixed effect by the ex ante individual probability of success, a variable at the individual-item level that is supposed to contain all the relevant information in terms of the interaction between the respondent and the item (i.e. how difficult a given item is for a given respondent) (Table 4.3).

In both specifications, time on task has a positive but declining effect on the probability of giving a correct answer. In other words, spending more time on an item increases the probability of giving a correct answer, but only up to a certain point. Spending an excessive amount of time, in fact, indicates that the respondent has not well understood the requests of the item and is therefore less likely to give a correct answer.

For Models 4 and 5, the time on task indicators are interacted with a dummy for whether the item was taken as part of Module 2. The regression also includes the main effect of Module 2. The coefficient on the main effect of Module 2 is negative, which is consistent with what was shown before: performance significantly declines in Module 2 compared to Module 1 (Table 4.2 showed that the proportion of missing answers increased from 5.4% to 9% in Module 2 and the proportion of correct answers decreased from 60% to 56%). This capture the average effect coming from fatigue or decrease in motivation that occurs at later stages of the assessment.

More interesting is the fact that the coefficient on the interaction term is positive and statistically significant. This means that, compared to Module 1, the returns to time on task are higher in Module 2: spending more time on a given item leads to a higher increase in the probability of giving a correct answer when that item is administered in Module 2.

This result can be interpreted in two ways. On the one hand, respondents could achieve better performance if they spent a bit more time on the items. It is possible that, by the time they get to Module 2, respondents are tired or less motivated, and as a consequence the value they attach to their free time has increased relative to the value they attach to performing well on the assessment. On the other hand, for a given amount of time spent on an item, respondents are more likely to give a correct answer if the item is administered in Module 2 rather than in Module 1. This would suggest that respondents become more efficient in answering items, although it is not possible to determine for what reason.

Table 4.3. Time on task and item performance

	Model 1	Model 2	Model 3	Model 4	Model 5
Time on task	1. 118	0.662	0.478	0.607	0.437
	(0.006)	(0.005)	(0.004)	(0.007)	(0.007)
(Time on task)^2	-0.010	-0.006	-0.004	-0.005	-0.004
	(0.000)	(0.004)	(0.000)	(0.000)	(0.000)
Module 2	-	-	-	-3.412	-1.350
				(0.217)	(0.203)
Time on task*Module 2	-		-	0.088	0.078
				(0.010)	(0.009)
(Time on task)^2*Module 2	-		-	-0.000	-0.000
				(0.000)	(0.000)
Item fixed effects	NO	YES	NO	YES	
Respondent fixed effects	NO	YES	NO	YES	
Ex ante probability of success	-	-	98.436	-	98.431
			(0.105)		(0.105)
R2	0.03	0.34	0.38	0.34	0.38
N. Observations	1 538 752	1 538 752	1 538 752	1 538 752	1 538 752

Note: The table reports results from different regression models. In all of them, the dependent variable is a dummy variable which equals 1 if the respondent has given a correct answer to the item and 0 otherwise. Standard errors are reported in parenthesis. Estimated coefficients and standard errors have been multiplied by 100. The sample includes only participants to the computer-based assessment who were assigned to the literacy and numeracy modules.
Source: OECD (2017[2]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933960213

Conclusions

This chapter investigated the relationship between time on task and item characteristics to shed light on the strategies and criteria respondents use to allocate time across different items in the course of the assessment.

In the first part of the chapter, the analysis was carried out at the item level. A first result is the large degree of between-item variation in time on task. In this respect, the differences between countries are not very pronounced. Time on task is strongly related to intrinsic item characteristics, such as item difficulty. Further evidence in this respect comes from the fact that respondents who correctly answered an item spent about the same amount of time as respondents who provided an incorrect answer. Respondents devoted a considerably smaller amount of time to items administered in the second half of the assessment. This was accompanied by a decrease in performance (measured by the fraction of items answered correctly) and by an increase in the proportion of missing answer. This seems to suggest that the decrease in time on task is due to an increase in fatigue or disengagement.

The second part of the chapter shifted the analysis to the level of the individual respondents. An important result is that respondents seem to allocate time to tasks in a rational way, devoting less time to items that are very difficult and very easy and more time to challenging items for which the probability of success is close to 50%. This pattern is observed in different countries, as well as in the two modules of the assessment. However, in Module 2, respondents seem to be much faster in deciding to skip items. The fact that the relationship between time on task and individual probability of success is the same across the modules provides some evidence of a learning effect. The decrease in time on task during the course of the assessment is, then, not entirely due to fatigue or disengagement, but also to some degree to the fact that respondents become more efficient in the way they interact with the assessment.

Finally, the analysis estimates the impact of time on task on performance, measured by the probability of giving a correct answer to an item. The structure of the dataset and the partially random allocation of items to respondents make it possible to control for item and respondent fixed effects, as well as for the position of the item. Indeed, the analysis shows that spending more time on an item does increase the probability of giving a correct answer, although at declining rates.

References

Borghans, L. and T. Schils (2012), *The Leaning Tower of Pisa: Decomposing Achievement Test Scores into Cognitive and Noncognitive Components*, http://www.sole-jole.org/13260.pdf. [5]

Borgonovi, F. and P. Biecek (2016), “An international comparison of students’ ability to endure fatigue and maintain motivation during a low-stakes test”, *Learning and Individual Differences*, Vol. 49, pp. 128-137, http://dx.doi.org/10.1016/j.lindif.2016.06.001. [3]

Brunello, G., A. Crema and L. Rocco (2018), “Testing at length if it is cognitive or non-cognitive”*, Discussion Paper Series*, No. 11603, IZA, http://ftp.iza.org/dp11603.pdf. [4]

Goldhammer, F. et al. (2014), “The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment.”, *Journal of Educational Psychology*, Vol. 106/3, pp. 608-626, http://dx.doi.org/10.1037/a0034716. [6]

Maddox, B. et al. (2018), “Observing response processes with eye tracking in international large-scale assessments: Evidence from the OECD PIAAC assessment”, *European Journal of Psychology of Education*, Vol. 33/3, pp. 543-558, http://dx.doi.org/10.1007/s10212-018-0380-2. [1]

OECD (2017), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, GESIS Data Archive, Cologne, http://dx.doi.org/10.4232/1.12955. [2]

Chapter 5. Measuring disengagement in the Survey of Adult Skills

This chapter uses timing indicators to estimate and analyse disengagement with the Survey of Adult Skills assessment. The analysis shows that the incidence of disengagement varies substantially across countries. Respondents with low levels of education and low familiarity with information and communications technology (ICT) are more likely to be disengaged, and respondents are more likely to be disengaged with items that appear in the second module of the assessment. Disengagement strongly reduces the probability of giving a correct answer, which results in disengaged individuals performing worse in the assessment. This relationship holds at both individual and country levels.

Introduction

Using log-file data to explore disengagement with the Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC) (hereafter referred to as "PIAAC") assessment, this chapter analyses the distribution of a measure of disengagement based on time on task across test items, countries and respondents – similar to the approach of Goldhammer et al. (2016[1]). It also explores the correlation of this measure with indicators that capture other aspects of disengagement.

What is disengagement

For the purposes of this chapter, participants in an assessment are considered disengaged when they do not devote sufficient effort (or take sufficient care) in responding to test questions to ensure that test results fairly represent their proficiency. Some variation is expected in respondents' efforts to answer test items. However, while comparisons of proficiency across individuals are inevitably influenced by differences in the amount of effort exerted, they are unambiguously biased once participants are disengaged. It is difficult to provide a rigorous definition of "sufficient or reasonable effort". In practice, the choice of definition is necessarily driven by considerations related to what can be reliably measured in order to make valid comparisons.

In this chapter, disengagement is analysed in binary terms: respondents are regarded as either engaged or disengaged, with no attempt made to measure the degree or intensity of disengagement.[1] One consequence of this choice is that no account is taken of variation in the intensity of disengagement, from the extreme case of a respondent who refuses to carry on and skips all remaining items to that of a respondent who skips a specific item because answering is seen as taking too much time, even though he/she has a good chance of providing a correct answer.

It is difficult to operationalise the notion of disengagement. First of all, effort has a subjective dimension, and individuals' perceptions of how much effort a task takes will vary. Inferring effort levels from the observation of certain actions undertaken by respondents requires taking these actions as indicators of effort, without taking into account how they are individually perceived. For instance, playful respondents could view the PIAAC assessment as a kind of game and give their best, while actually enjoying the whole process. Nonetheless, in comparing effort across individuals, it is necessary to equate effort with some instrumental actions and processes.

Second, and more importantly, instrumental actions in the course of a cognitive assessment are difficult to observe because they are mental in nature. From the point of view of the respondent, dealing with an item is a succession of choices and a sequence of actions, generally starting with reading the question. Each sequence of action is associated with a duration cost (time spent), an effort cost and a change in the probability of providing a correct answer to the item. At any point in a respondent's deliberations prior to providing a response, he/she will take into account the costs (in time and effort) and benefits (demonstrating his/her "true" ability) of pursuing any further action, compared to the costs associated with moving to the next item or withdrawing from the assessment.

Personal commitment and cost of effort vary across individuals. Personal commitment will depend strongly on cultural factors (how interviewers and respondents interact and the respondent's own desire to perform well), fatigue and cost of effort on environmental factors (such as distractions), and levels of effort on item characteristics. One of the virtues of an adaptive assessment, as in PIAAC, is its positive impact on engagement. Adaptive allocation of items alleviates the detrimental effects of fatigue, by limiting the frequency of situations in which the participant has to struggle with difficult items for which he/she is unlikely to get the correct answer.

An alternative to action-based measurement would be to consider self-reported measures of effort. This type of measure is not available in PIAAC, and it is difficult to compare these types of measures across countries and individuals, due to their subjective nature. In particular, if it is true that differences in disengagement across countries are driven by differences in perceptions of what constitutes sufficient or reasonable effort, we might expect self-assessed effort scales to be biased accordingly.

Building on the analysis undertaken in Chapter 2 on time on task, this chapter examines the question of disengagement. Time on task does not provide information on how respondents are using their time and thus cannot distinguish between time spent on task-related actions and time spent on activities or actions unrelated to answering a test item. Nonetheless, it can be safely assumed that a respondent who decides to spend more time on an item is, at the very least, not decreasing the effort exerted to answer an item and may even be increasing the effort.

Why disengagement matters

In PIAAC, disengagement may arise simply because of the low-stakes nature of the assessment. Unlike exams or competitions, performance in PIAAC has no consequences for individual respondents and is not related to any kind of incentive (reward or punishment) to exert high levels of effort. In addition, participants do not receive any feedback about their performance, either during the test or on completion. Nonetheless, by agreeing to participate in PIAAC, respondents can be regarded as having entered into some kind of implicit contract to make a minimum effort during the assessment. As participation in PIAAC is not obligatory, respondents must be sufficiently motivated to agree to devote a fair amount of time to the assessment and hence to make a reasonable effort to respond seriously to the various questions.

Interviewers play a major role in gaining the agreement of respondents and ensuring that participants take it seriously. From this point of view, participants in PIAAC start the assessment with a reasonably high level of personal commitment, and disengagement occurs once the cost of participation in time and effort starts to be deemed too high.

Respondents' disengagement matters mostly because it is a source of undesirable variation in estimates of proficiency. Disengagement may mean that respondents do not demonstrate their true level of proficiency, which will affect the validity of inferences that can be made from the assessment. In addition, different levels of disengagement between subgroups within countries and between countries may reduce the validity of comparisons.

However, the relationship between disengagement and performance is a complex question that remains beyond the scope of this chapter for a number of reasons. First, disengagement in PIAAC can only be measured with indicators that partially capture the spectrum of disengagement. As a result, any causal impact of a disengagement indicator on performance would only deliver a partial answer. Second, disengagement and low performance are linked in a complex relationship that cannot be easily disentangled. Third, PIAAC proficiency scores already partially account for disengagement by ignoring (in the underlying model) items on which respondents spent less than five seconds without giving an answer. In particular, following the literature on response latencies (Wise and Kong, 2005[2]; Wise and DeMars, 2005[3]), it was decided that instances in which the interaction between the respondents and the item was very brief are not informative, so they are coded as non-reached items rather than missing items. Nonetheless, respondents in such situations were also strongly disengaged, making almost no effort to give a correct answer to the item. As a result, PIAAC proficiency scores are computed on a sample that already excludes the most extreme cases of item disengagement.

The degree to which external factors, such as motivation, influence the results of low-stakes assessments is an active and growing area of research. One approach consists of comparing the performance of similar respondents in high- and low-stakes testing situations. Using an assessment similar to the Programme for International Student Assessment (PISA), Gneezy et al. (2017[4]) conducted experiments in schools in Shanghai and the United States. They showed that a significant proportion of the gap observed between the two countries in official PISA rankings disappears when students are offered monetary incentives.

Another approach consists of decomposing test scores into two components, one capturing initial performance and the other capturing decline in performance during the test (Borghans and Schils, 2012[5]). Initial performance is often interpreted as the true ability of the individual, as it is assumed to not be contaminated by fatigue effects or by decrease in motivation. Decline in performance during the test is often interpreted as a non-cognitive skill, such as the ability of respondent to remain motivated, or to endure fatigue (Borgonovi and Biecek, 2016[6]; Zamarro, Hitt and Mendez, 2016[7]; Anghel and Balart, 2017[8]; Balart, Oosterveen and Webbink, 2015[9]; Brunello, Crema and Rocco, 2018[10]).

Measuring disengagement at the item level

Rapid item skipping

The simplest indicator of disengagement is rapid skipping of an item. Respondents who spend less than a very short amount of time on an item (i.e. do not give themselves enough time to even read and take full note of the item) can be considered to be disengaged. The analysis in this chapter is based on a threshold of five seconds, below which respondents are considered disengaged. This ensures consistency with the PIAAC rule about rapid omission. However, no account is taken of whether or not the respondent provided an answer.

Table 5.1 shows the proportion of item interactions for each country in which respondents spent less than five seconds and, among them, the proportion with an answer and the proportion with a correct answer. It thus gives a first account of disengagement across countries. The proportion of items that are rapidly skipped varies from 0.7% in Norway to 4% in Spain and Italy. For most countries, respondents who spent less than five seconds on an item did so without giving an answer. The proportion of these items that receive an answer is generally below 5%. This confirms that, in the overwhelming variety of cases, item interactions that last less than five seconds are not productive.

Table 5.1. Rapid item skipping across countries

	Proportion of all item with time on task below 5 seconds	Among items with time on task below 5 seconds:	
		Proportion with an answer	Proportion with a correct answer
Austria	0.8%	1.5%	0.2%
Germany	0.8%	3.8%	0.3%
Denmark	2.1%	0.1%	0.0%
Belgium (Flanders)	1.5%	2.4%	1.2%
Estonia	0.1%		
Spain	4.0%	34.2%	21.0%
Finland	0.8%	0.9%	0.1%
France	2.2%	1.5%	0.3%
England / Northern Ireland (United Kingdom)	1.7%	3.1%	0.5%
Ireland	2.2%	1.5%	0.3%
Italy	4.0%	5.7%	2.7%
Netherlands	0.9%	4.1%	1.7%
Norway	0.7%	3.4%	0.0%
Poland	2.5%	6.2%	2.3%
Slovak Republic	1.5%	16.0%	10.7%
United States	1.5%	8.8%	3.3%

Note: In Estonia, the number of items answered in less than 5 seconds is too small to perform an analysis.
Source: OECD (2017[11]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933960232

However, two countries stand out as exceptions. In the Slovak Republic, 16% of these items were answered. The corresponding proportion was even higher in Spain, where it reaches 34%. In both countries, the proportion of correct answers is similarly high, with two-thirds of these answers being correct. Even though some items feature a multiple-choice format that allows for random guessing, this rate is too high to be plausible.

This phenomenon of rapid correct answers, which is restricted to these two countries, is hard to explain. In particular, data from Spain feature both a high rate of rapid skipping and a high rate of correct answers. This combination would be problematic in the analysis that follows. As a result, data from Spain are excluded from the analysis conducted in the rest of this chapter. Spain also displays a rate of rapid skipping without answers, suggesting that disengagement in Spain is among the highest in the sample of countries.

Rapid item skipping is informative, but it fails to take into account less acute forms of item disengagement. It intends to measure the quasi-absence of interaction (and consequently the quasi-absence of effort) between respondent and item. However, disengagement occurs not once the effort is deemed non-existent, but once it is deemed insufficient. Rapid item skipping, thus, does not capture the range of disengaged items that falls in between these extremes.

T-disengagement

A more refined but less strict concept of disengagement is to see it as a situation in which the respondent has not spent enough time on an item to provide a correct answer.

Operationalisation of this definition requires defining the minimum time necessary to solve an item without resorting to random guessing.

Goldhammer et al. (2016[1]) use the relationship between the likelihood of giving a correct answer and time on task to compute an item-specific threshold below which respondents can be reasonably assumed to not have seriously attempted to solve the item (in which case they are classified as disengaged). This relationship generally starts from a zero probability of success and remains flat up to some threshold at which the probability of success starts to rise.

This chapter adopts a similar approach, adopting the term T-disengagement to represent situations where a respondent spends less time than an item-specific threshold. For each item, it uses this empirical relationship observed in all countries of the sample together (excluding Spain, as explained above). These thresholds will be the same across all countries. For each item, this minimal time is constrained to be at least five seconds. T-disengagement is intended here to extend rapid item skipping. After excluding Spain, the occurrence of correct answers in less than five seconds is limited to four countries (Italy, the Netherlands, Poland and the Slovak Republic) and remains a very rare event. In all the other countries, figures remain anecdotic and justify the statement that respondents who spent less than five seconds are disengaged.

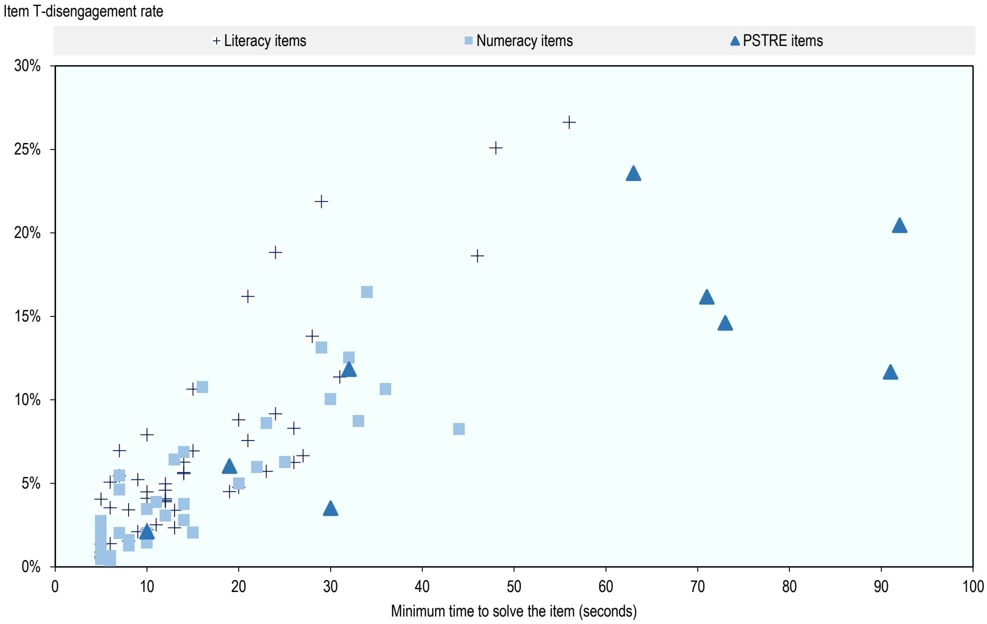
Even though the sample sizes for each item are reasonably large (around 15 000 on average), data at the bottom tail of the distribution of time on task (where minimal time to solve will be found) can be sparse for some items. In order to smooth the relationship between time on task and success, the following procedure is applied: 1) to compute the probability of success at time x, observations with a time on task between x and x+10 are used; 2) if this subsample contains more than 200 observations, success on the item is modelled as a linear function of time; 3) if the subsample contains fewer than 200 observations, the probability of success is not estimated. The minimal time to solve an item will eventually be the smallest x (larger than five seconds) for which the estimated probability of success is higher than 10%.

Respondents who spent less than this minimal time may still have extracted enough information from the item to realise that they will not be able to find a solution. This could be the case, for instance, if they do not understand how the question and stimulus are related or if they do not understand the task they are required to do. This situation is referred to as rational skipping, because respondents have no reason to spend time on items that they know they will not be able to solve. T-disengagement may thus also capture situations in which effort is useless rather than insufficient.

Figure 5.1 plots the distributions of minimum time needed to solve an item and T-disengagement rates for all numeracy, literacy and Problem Solving in Technology-Rich Environments (PSTRE) items. The time required varies between 5 seconds and 3 minutes.[2] Based on these thresholds, items can be classified as either "short" or "long". The shorter the item is, the closer disengaging with this item is to rapid item skipping. PSTRE items are much more time consuming than literacy and numeracy items, with a typical required time for solution of 1 minute. Most literacy and numeracy items can be solved in less than 30 seconds and a good proportion in less than 10 seconds. Only PSTRE items feature minimum times greater than 1 minute. T-disengagement rates vary between 0% and 30% for the most part but reach 60% for one literacy item.

The T-disengagement rate increases in close parallel with the minimum time needed to solve an item. Since the definition of this disengagement indicator is based on the minimum time, this is not all surprising. A respondent could spend 10 seconds on a short item without being considered T-disengaged and 10 seconds on a longer item and be classified as disengaged. Hence, for most items that can be solved in less than 20 seconds, the T-disengagement rate stays below 10%. Items that need more than 20 seconds to solve show higher and more variable rates of T-disengagement. For instance, items that need about 40 seconds to solve have disengagement rates varying between 10% and 25%, meaning that the T-disengagement depends on characteristics other than the time required to solve it, such as type of display, content or difficulty. Among long items, disengagement is more common in the literacy domain. These differences may be driven by booklet selections, as subsamples of respondents to which various items are allocated are not strictly comparable.

Figure 5.1. Item T-disengagement rates



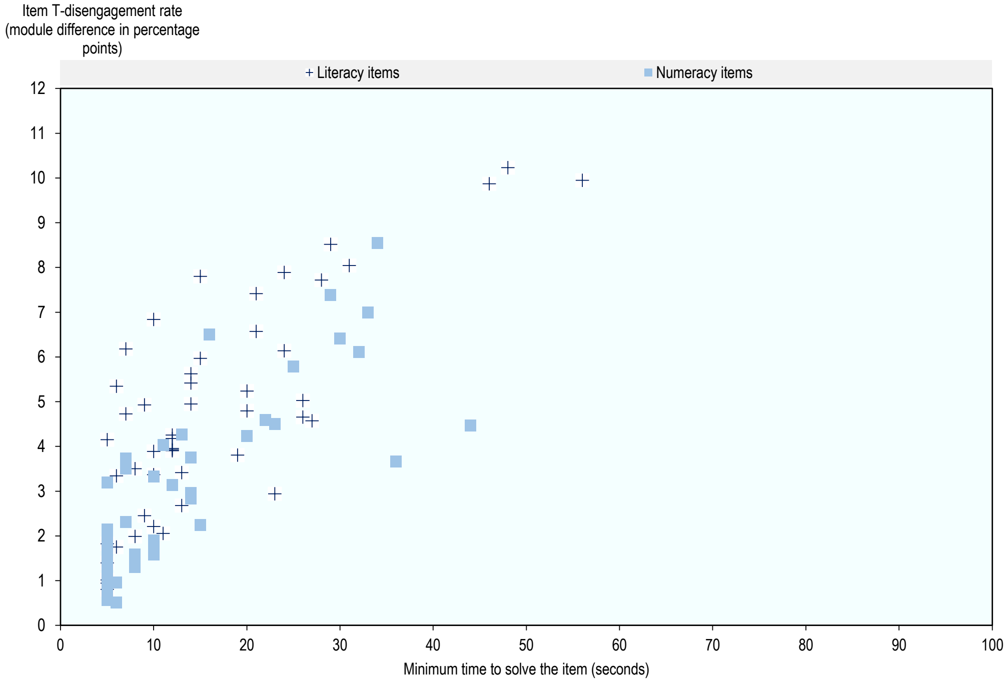
Note: Two PSTRE items are not shown because they are outliers, with minimum time to solve them exceeding 180 seconds.
Source: OECD (2017[11]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933959985

Figure 5.2 highlights the importance of module order for T-disengagement. As mentioned earlier, respondents can be assigned an item in the first or the second module, and this allocation is random. This figure plots the difference in disengagement for each item when the item is answered in the second module compared to the first one. When an item is in the second module, the probability that the respondent is disengaged with that item is between 1 and 10 percentage points higher than if the same item was in the first module. For short items (below 20 seconds), the difference is such that disengagement occurs twice as frequently in the second module. For longer items, this difference does not increase as fast and remains below 10 seconds, but it is still equivalent to a 50% increase between modules

This relationship offers support for the conclusion that T-disengagement represents a good indicator of lack of effort, since the increase in disengagement remains associated with very low success rates in disengaged items. Random allocation of respondents to items guarantees the absence of any selection effects. As a result, the difference between T-disengagement in the first and second module can be fully attributed to the fact of having undertaken the items in different modules (i.e., the difference captures a true "module effect"). This could be related to either fatigue or to a learning effect – more rapid identification of items that are likely to be too difficult for the respondent (rational skipping).

Figure 5.2. Item T-disengagement rates and module order



Source: OECD (2017[11]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933960004

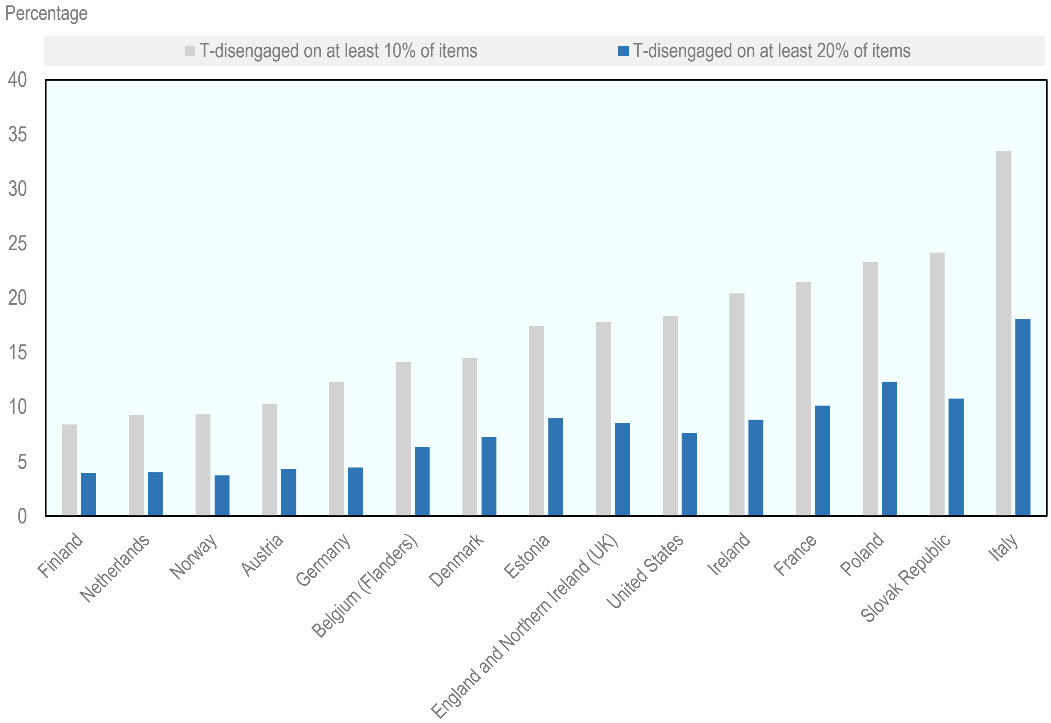
T-disengagement across countries

Figure 5.3 presents T-disengagement rates by country. Consistent with the rest of the report, the focus is on literacy and numeracy items. This is because PSTRE items have features that differentiate them from items in other two domains: there are fewer of them, they are longer and they have high disengagement rates. All respondents answer 20 literacy items and 20 numeracy items. Instead of plotting the average proportion of disengaged items, the choice was made to plot the proportion of respondents who disengaged on at least a given proportion of items. This choice is made to simplify the analysis and maintain the useful dichotomy between disengaged and non-disengaged respondents.

Figure 5.3 shows the proportions of the population who T-disengaged on at least 10% and at least 20% of items. Disengagement concerns respondents in all countries, but to a varying extent. Disengagement is much less frequent in northern European countries, such as Finland, Norway or the Netherlands. In these countries, about 8% of the sample disengage on at least 4 items out of 40. This proportion approaches 35% in Italy. The same differences between countries emerge when looking at more severe cases of disengagement, in which respondents disengage on at least 20% of items. The proportion drops below 5% in Finland, Norway and Netherlands, but it remains above 15% in Italy.

These rates give some indication of how PIAAC country scores might be affected by disengagement. They suggest that comparison of countries with low T-disengagement rates, such as Austria, Finland, Germany, the Netherlands and Norway, are probably more reliable than comparisons with countries characterised by higher rates of item-specific disengagement, as the much higher disengagement rates in Italy suggest that proficiency in these countries might be underestimated compared to others. This does not imply that proficiency estimates for disengaged respondents do not convey important and valuable information about them. But they are likely to contaminate the measurement of latent ability, as defined in the conceptual framework of the PIAAC assessment. In the end, a joint analysis of test scores and disengagement rates provide a more accurate and complete picture of the proficiency of respondents in participating countries. The remainder of the chapter further explores T-disengagement to assess the validity of this indicator.

Figure 5.3. T-disengagement across countries



Source: OECD (2017[11]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933960023

T-disengagement and background characteristics

These important differences across countries could be driven by several factors, reflecting the manner in which respondents interact with items and determine their effort levels. This section describes the association of T-disengagement with individual background characteristics.

Figure 5.4 explores the relationship between disengaging on more than 10% of items and several background characteristics. The figures reported are averages of estimated coefficients across all available countries. These coefficients are estimated in a single ordinary least squares (OLS) regression model, meaning that each estimate takes into account the effect of all other covariates.

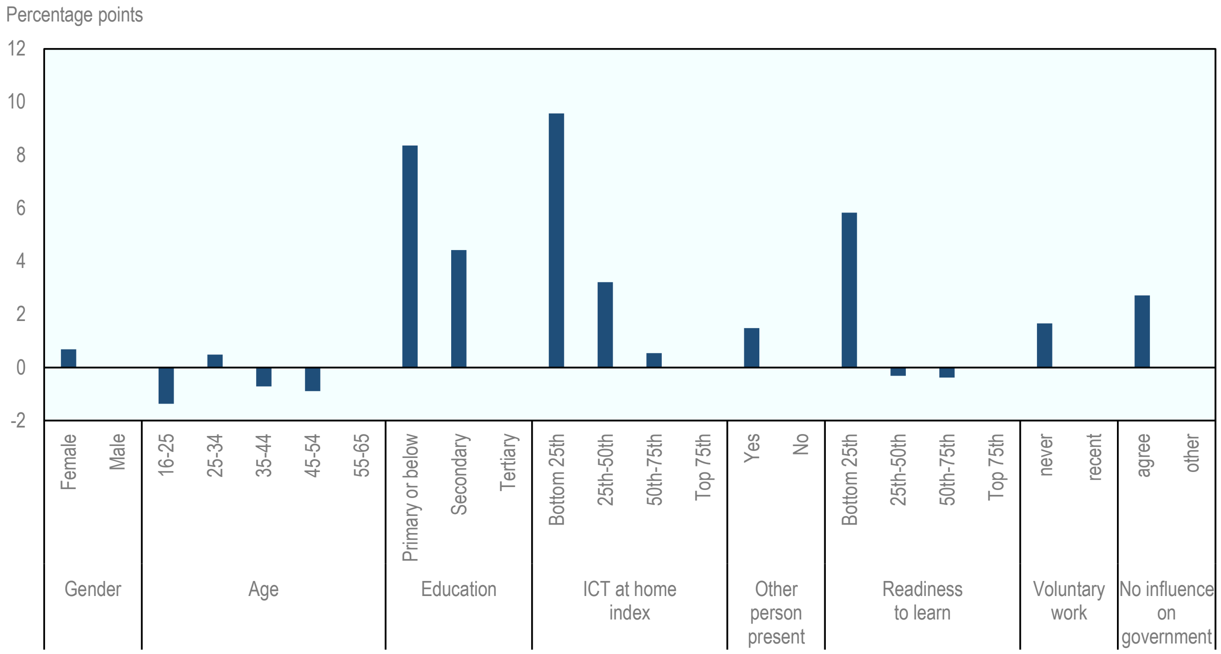
A first result is the absence of gender differences, with female and male respondents being equally likely to be T-disengaged in all countries.

The relationship between age and T-disengagement is also relatively weak on average. Young respondents are less likely to be disengaged, than respondents over age 25, while middle-aged respondents are slightly less likely to be disengaged, based on item-specific thresholds. The relationship between T-disengagement and age is highly country-specific. This suggests that biological factors, such as ability to concentrate or fatigue, while possibly explaining the relationship between age and disengagement, are not dominant. Most importantly, in England / Northern Ireland (United Kingdom), there is a steep decrease in T-disengagement with age, with the youngest respondents being 17 percentage points more likely to T-disengage.

Education levels are negatively associated with T-disengagement. This impact is sizeable with rates being, on average, 8.5 percentage points higher for respondents with less than secondary education than for respondents with tertiary attainment. In addition, the relationship is stable across countries, although its magnitude varies. The association between T-disengagement and education might be related to several factors. One reason may be that since respondents with higher education are, on average more proficient, they need to answer fewer items that are (from their point of view) relatively difficult, although the adaptive nature of the assessment partly corrects for this. Another reason could be that they are more accustomed to or have acquired more experience with testing and assessment environments. As a result, they may experience less fatigue than other respondents, even though they spend more time on the assessment (see Chapter 3). This would also suggest that fatigue is related to cognitive demand rather than test length. In addition, more highly educated respondents could also have a stronger sense of commitment to completing the assessment to the best of their ability. Nonetheless, as mentioned earlier, these differences could be driven by rational skipping as well. Less educated respondents may be more likely to not understand some questions or to be aware that they are unable to solve them.

Given that the assessment was taken on a computer, familiarity with information and communications technology (ICT) can plausibly affect respondents' motivation, fatigue and engagement. The frequency of use of ICT at home is indeed strongly associated with T-disengagement. Respondents in the bottom quartile of the ICT-use index are on average 9 percentage points more likely to be T-disengaged than those in the top quartile. This effect seems to be concentrated in the lowest quartile and has a straightforward interpretation. Respondents who are not familiar enough with computers, but successfully complete the ICT core and pass the computer-based assessment will have more difficulty undertaking the assessment on a computer than other respondents, due to their lack of familiarity with computers.

Figure 5.4. T-disengagement and background characteristics



Note: Country averages of regression coefficients with 'disengaged at least 10% items' as a dependent variable.
Source: OECD (2017[11]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933960042

The presence of another person during the assessment is associated with an increase of 1.5 percentage points in the probability of disengagement. The presence of another person is an environmental factor that might increase the cost of effort, because respondents' attention and focus on the assessment could potentially be distracted by communication with the other person. However, the estimated effect is quite small, suggesting that this potential source of disturbance did not play a significant role.

Disengagement is strongly associated with readiness to learn. The readiness-to-learn index is constructed on the basis of a set of questions about the respondent's perception of himself/herself as a curious and perseverant individual (respondents are asked questions such as 'Do you get to the bottom of difficult things?'). Respondents in the lowest quartile of this index are more likely to be disengaged than those in higher quartiles, by a margin of 8 percentage points. In so far as this index is associated with how respondents attach value to the search for a correct answer, this relationship exhibits another mechanism through which a respondent decides on the level of effort to exert. The less respondents value success, the lower level of effort they would accept. This association suggests that respondents choose effort levels rationally, by comparing the benefits of actions to the costs.

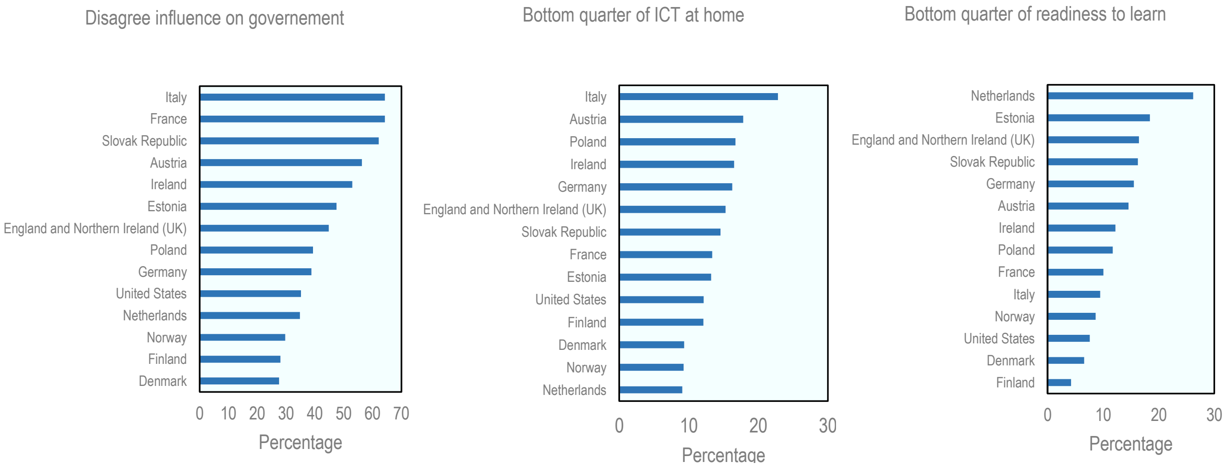
Respondents who do not engage in voluntary work, as those who agree that they do not have influence on the government are more likely to be T-disengaged. This association is not as high as the association with readiness to learn but it is not negligible. One important source of disengagement is insufficient commitment to the effort required for the assessment. This commitment is eventually the reason why respondents agree to participate in the survey. Although it is difficult to know what respondents are willing to accept, it is logical to assume that respondents with stronger ties to civic life would accept more.

Overall, the relationship between T-disengagement and respondent background variables seems to be in line with a simple model of how respondents choose their effort levels. Moreover, it suggests how disengagement might affect some important socio-demographic gaps. In particular, education proficiency gaps might be smaller than what is featured in PIAAC, and in some countries (such as England and Northern Ireland [United Kingdom]), age differences might be affected by varying levels of disengagement.

Further analysis of T-disengagement across countries

The question then arises of the importance of these background variables in shaping variations across countries and, more generally, of the sources of the differences in T-disengagement rates.

Figure 5.5. Variations across countries of some important factors



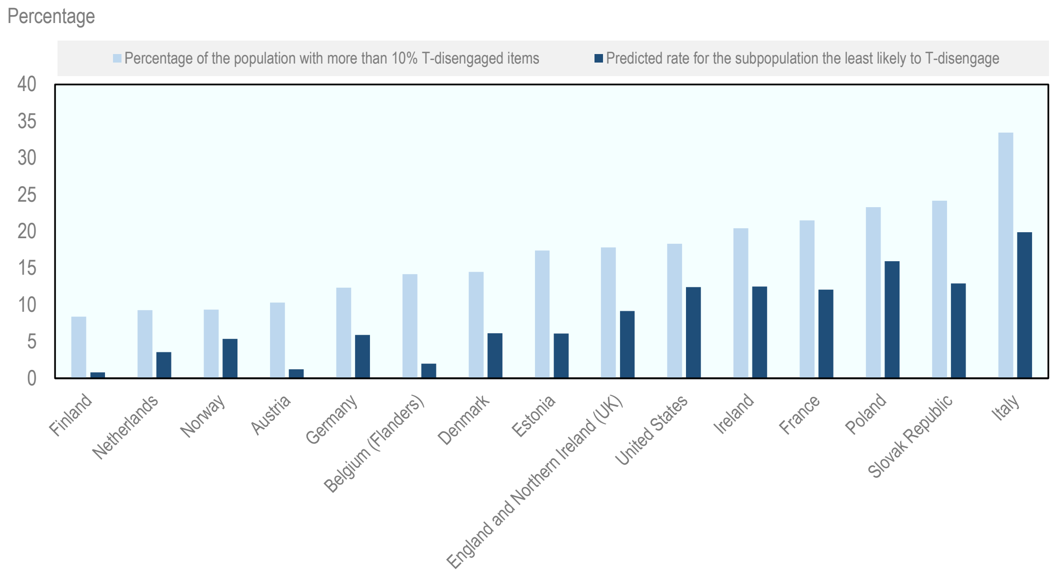
Source: OECD (2015[12]), *OECD Survey of Adult Skills (PIAAC) (Database 2012, 2015)*, http://www.oecd.org/skills/piaac/publicdataandanalysis/.

StatLink http://dx.doi.org/10.1787/888933960061

Figure 5.5 shows country averages for three of the background factors that have the strongest association with T-disengagement (other than educational attainment): 1) the proportion of respondents who disagree that they have an influence on the government; 2) the proportion who belong to the bottom quartile of use of ICT at home; and 3) the proportion of respondents who belong to the bottom quartile of readiness to learn. For two of these factors, variations across countries are sizeable. The share of respondents who disagree that they have an influence on the government varies from 25% of the population (in Denmark) to 65% (in Italy). The proportion of respondents who fall in the bottom quarter of the readiness-to-learn index is lowest in Finland (less than 5%) and highest in the Netherlands (24%). The proportion of respondents who are in the bottom quartile of the ICT-use-at-home index features smaller variations. Country rankings on these variables do not seem to mirror T-disengagement rankings, with the notable exception of the proportion of respondents who disagree that they have an influence on government. This similarity suggests (but does not prove) that disengagement and this factor may be related.

Figure 5.6 plots raw T-disengagement rates along with the rate adjusted for all the factors positively associated with T-disengagement in Figure 5.4, with the exception of age.[3] This adjusted rate is thus the predicted rate of T-disengagement for the subpopulation the least likely to be disengaged: a male with tertiary education who agrees that he has influence on government, participates in voluntary activity and belongs to the top quartile of the readiness-to-learn and ICT-use-at-home indices.

Figure 5.6. Raw and predicted rates of T-disengagement for the subpopulation least likely to disengage



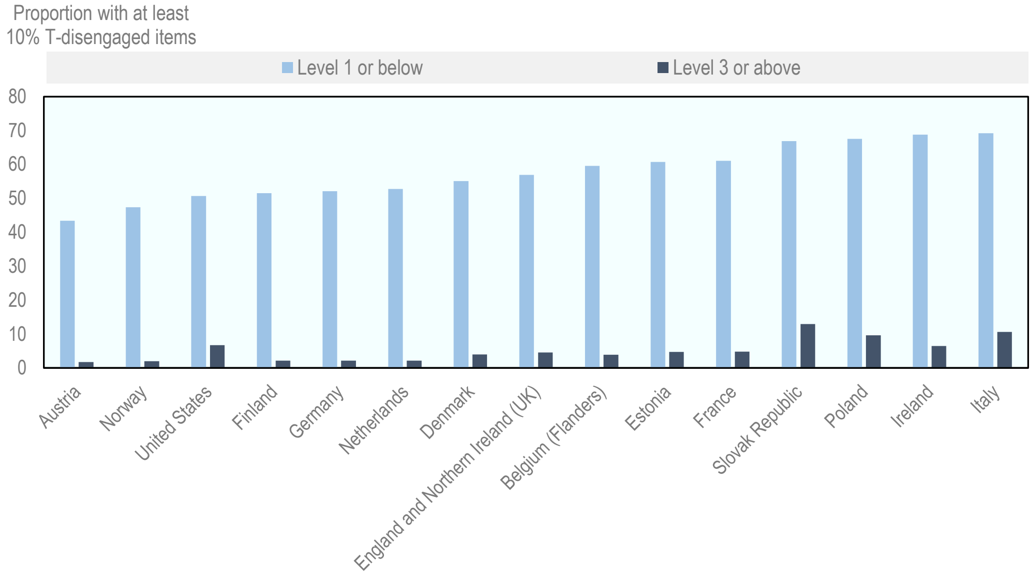
Source: OECD (2017[11]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933960080

This adjustment typically decreases rates by 5 to 15 percentage points. Most surprisingly, there is a large decrease even in countries in which the raw rate is low. As a result, in Austria, Finland or Netherlands, the rate falls close to zero, while in Italy it remains in the 15% to 20% range. Consequently, while T-disengagement seems to be a matter of personal characteristics in the first group of countries, with a fringe of the population not likely to disengage, disengagement in the second group has an endemic component: even the subpopulation with characteristics associated with lower disengagement is likely to disengage.

Figure 5.7 shows T-disengagement across countries for respondents with low and high literacy levels. In all countries, a majority of respondents who score at Level 1 or below are disengaged. This proportion varies from almost 40% in Austria to up to 70% in Italy. There are two strong reasons for these high shares in all countries. Less able respondents are more often required to answer items that are difficult for them than respondents of high ability. Their propensity to T-disengage will thus increase because of accumulated fatigue and because they rationally skip more items than respondents of high ability. Moreover, disengaged items are items that were not successfully completed; as a result the estimated proficiency of disengaged respondents will be mechanically lower. Across the whole proficiency distribution, as measured in PIAAC, the lower end of the proficiency distribution is the most subject to disengagement bias.

Figure 5.7. T-disengagement and literacy proficiency



Source: OECD (2017[11]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

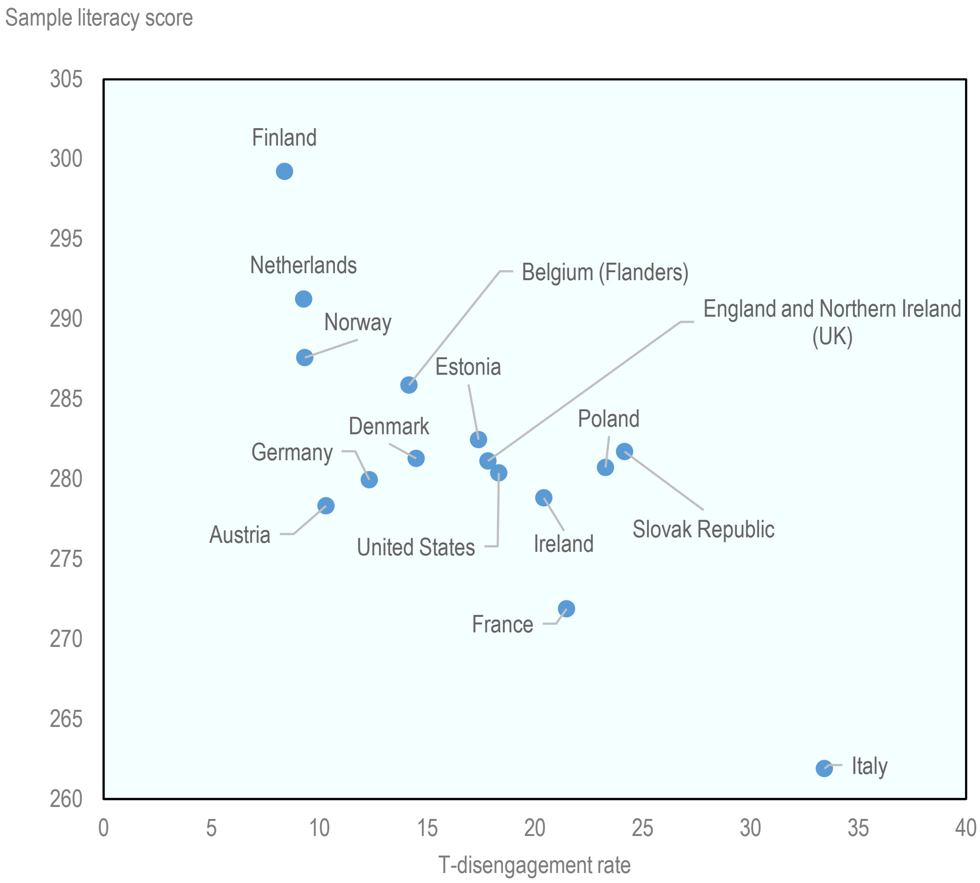
StatLink http://dx.doi.org/10.1787/888933960099

The picture for high-proficiency respondents is strikingly different, featuring a pattern very similar to the one found in Figure 5.6. Among respondents with high proficiency in Austria, Finland and Norway, less than 2% are T-disengaged, while more than 10% are T-disengaged in the Slovak Republic. These differences cannot be explained by rational skipping. Rational skipping is mostly related to relative difficulty, and these rates are computed on a population with high proficiency.

Figure 5.8 plots T-disengagement rates against the literacy performance of the subsample on which the T-disengagement rate was computed. Once again, Finland on the one hand and Spain and Italy on the other stand apart. Finland features both high average literacy scores and low disengagement, while the high T-disengagement rates observed in Spain and Italy are associated with much lower literacy performance.

It is not possible to provide causal estimates of the impact of disengagement on literacy performance. Figure 5.8, however, suggests that it is possible to identify a cluster of countries that differ in terms of level of engagement (as measured by this particular indicator). This might serve as a first step in furthering understanding of the role that engagement plays in contributing to cross-country differences in proficiency in low-stakes assessments such as PIAAC.

Figure 5.8. T-disengagement and sample literacy score



Source: OECD (2017[11]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933960118

Comparisons of T-disengagement and other indicators

The indicator described above is only one aspect of disengagement. PIAAC offers other possibilities that might help build a more detailed picture of disengagement variations across countries.

Table 5.2 presents a comparison across countries of T-disengagement with three other indicators. As mentioned earlier, fast item skipping is a more restrictive version of item disengagement based on time on task. The second column shows an indicator that summarises fast item skipping with the proportion of respondents who spent less than five seconds on at least 10% of items. The third indicator considers disengagement during the background questionnaire, rather than during the assessment. The last indicator is based on a question about respondents' perception of the length of the assessment. This question comes from the observation module, which is completed by the interviewer right after the interview. In describing whether the respondent felt that the length of the assessment was reasonable or not, this question does not indicate disengagement as such but describes one of its potential sources. These three indicators are strong predictors of T-disengagement at the individual level. On average across countries, 97% of respondents who rapidly skip items are also T-disengaged, (compared to 23% for other respondents), 34% of those are among the fastest on Section I of the background questionnaire (compared to 25% for other respondents) and 36% of those thought that the assessment was too long (compared to 22% for other respondents).

Table 5.2. Comparisons of disengagement indicators across countries

	Proportion who T-disengaged on more than 10% of items,	Proportion who spent less than 5 seconds on more than 10% of items	Proportion among the 25% fastest on Section I of the background questionnaire	Proportion who thought the assessment was too long
Italy	33.4%	13.2%	41.1%	47.8%
Slovak Republic	24.2%	5.1%	46.4%	34.3%
Poland	23.3%	8.0%	28.0%	46.5%
France	21.5%	7.4%	8.0%	45.3%
Ireland	20.4%	6.3%	33.3%	26.5%
United States	18.3%	6.3%	33.3%	26.5%
England / Northern Ireland (UK)	17.8%	6.3%	15.9%	23.0%
Estonia	17.0%	6.0%	2.0%	26.4%
Denmark	14.5%	6.0%	4.1%	14.0%
Belgium (Flanders)	14.0%	4.0%	25.0%	15.6%
Germany	12.3%	3.1%	15.0%	10.2%
Austria	10.3%	2.2%	9.4%	21.3%
Norway	9.3%	2.8%	4.9%	13.6%
Netherlands	9.3%	2.8%	12.7%	15.8%
Finland	8.4%	3.0%	6.1%	13.2%

Note: For each column, the three highest values are highlighted in dark blue and the three lowest ones in light blue.

Source: OECD (2017[11]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933960251

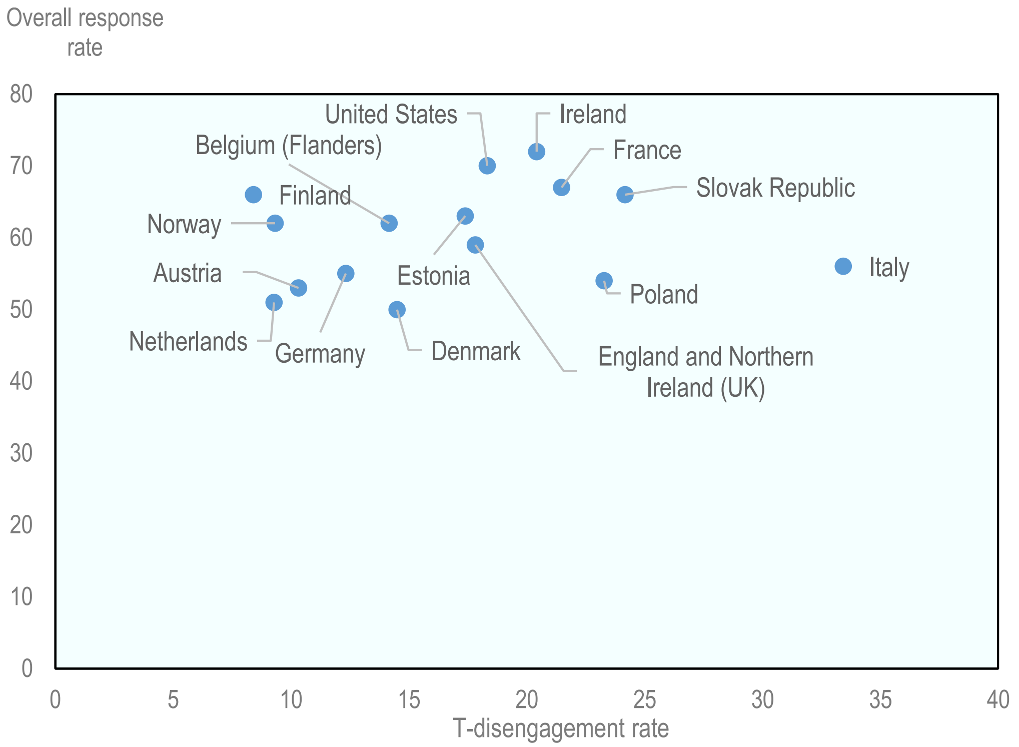
All these indicators exhibit large variations across countries and, most importantly, these variations are closely related to those found for T-disengagement. In Austria, Finland, the Netherlands and Norway, less than 3% of the sample belongs to the category of rapid skippers, while this proportion exceeds 13% in Italy and Spain. In Denmark and Norway, less than 5% are among the fastest on Section I of the background questionnaire, compared with more than 40% in Italy and the Slovak Republic. In all countries, at least a small minority of respondents found the assessment too long. The proportion remains among the lowest (below 14%) in Finland, Germany and Norway, but it is close to a majority in Italy and Poland. For all indicators, Italy and the Slovak Republic rank among the highest, while Norway and Finland rank at the bottom.

The similarity between the country rankings of these various indicators of disengagement shows how disengagement varies across countries. In particular, it confirms that disengagement seems to matter the most in Italy, Poland and the Slovak Republic.

While this chapter has focused so far on disengagement during the interview, potential participants who refuse to be surveyed can also be considered as disengaged, or more precisely, as refusing to engage. The relationship at the country level between the prevalence of disengagement during the interview and overall response rates might thus stem from a trade-off. Respondents who are at the margin of refusing to participate are also among those most likely to disengage during the survey. As a result, improving the response rate might also have the side-effect of increasing disengagement.

Figure 5.9 plots the relationship between T-disengagement and response rates at the country level. The lack of a clearly positive empirical relationship between the two rates proves that country-specific forces that determine both rates dominate the potential trade-off between response rates and the prevalence of disengagement among those who agree to participate. Nonetheless, this figure highlights valuable contrasts. For instance, among countries with low T-disengagement, only Finland and Norway have a satisfying response rate, while Austria, Denmark, Germany and the Netherlands are among the countries with the lowest response rates. France, Ireland and the United States have all high response rates and average T-disengagement rates. And, while the above discussion highlights the Slovak Republic and Italy among countries where disengagement is prevalent, only the Slovak Republic has a high response rate.

Figure 5.9. T-disengagement and response rates, by country



Source: OECD (2017[11]), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, http://dx.doi.org/10.4232/1.12955.

StatLink http://dx.doi.org/10.1787/888933960137

Conclusions

Comparisons of performance in a cognitive assessment can produce misleading conclusions if not all participants exert a sufficient amount of effort. Without sufficient effort, performance on the assessment will not accurately represent the underlying ability of the respondent. This problem is particularly relevant in the case of low-stakes assessments, where participants do not have external incentives to perform at their best.

The information contained in log files makes it possible to more precisely observe the behaviour of respondents in the course of such assessments and to construct indicators that can be used to proxy the amount of effort exerted.

This chapter presented and analysed various indicators that can be used to classify respondents as either engaged or disengaged with assessment items. The incidence of disengagement varies substantially across countries. In Norway, Finland, and the Netherlands, less than 10% of respondents are disengaged in at least 10% of items, compared to more than 20% in France, Ireland, Poland and Slovak Republic, and more than 30% in Italy.

Low levels of education and low familiarity with ICT (proxied by the frequency of performance of ICT-related tasks in everyday life) are positively associated with the probability of being disengaged in the course of the assessment. Similarly, respondents who report that they are generally less perseverant are also more likely to be disengaged.

Respondents are also more likely to be disengaged with items that appear in the second module of the assessment rather than in the first. This is consistent with the findings discussed in Chapter 3 that respondents tend to spend less time on items positioned in the second module.

Not surprisingly, disengagement strongly reduces the probability of giving a correct answer, which results in disengaged individuals performing worse in the assessment. This relationship holds at both the individual and the country level.

Indicators of disengagement are, therefore, very useful in two respects. On the one hand, disengagement provides important information on the respondent and can be used to proxy a variety of individual traits (such as conscientiousness or the ability to endure fatigue) that are likely to be important determinants of real-life economic and non-economic outcomes. On the other hand, these traits are not part of the skills cognitive assessments typically try to measure. As a result, the presence of disengagement (or any kind of difference in the effort respondents exert during an assessment) biases the results of assessments and can make comparison of results across countries problematic. In this sense, information on the extent of disengagement is a useful complement to actual estimates of proficiency that can be used to make more accurate comparisons across countries.

Notes

[1] In a sense, and with all the caveats discussed in previous chapters, time on task could be interpreted as a continuous measure of the effort respondents exert in solving the items and, therefore, as a measure of the degree of engagement.

[2] Three minutes were required for two PSTRE items that are not shown in Figure 5.1.

[3] Age is excluded here, because the effect of age is not homogeneous and the choice of a reference is therefore not natural. For instance, while old respondents are the least disengaged in England / Northern Ireland (United Kingdom), this is not true in all countries.

References

Anghel, B. and P. Balart (2017), "Non-cognitive skills and individual earnings: New evidence from PIAAC", *SERIEs*, Vol. 8/4, pp. 417-473, http://dx.doi.org/10.1007/s13209-017-0165-x. [8]

Balart, P., M. Oosterveen and D. Webbink (2015), "Test scores, noncognitive skills and economic growth", *IZA Discussion Paper*, No. 9559, The Institute for the Study of Labor, Bonn, http://ftp.iza.org/dp9559.pdf. [9]

Borghans, L. and T. Schils (2012), *The Leaning Tower of Pisa: Decomposing Achievement Test Scores into Cognitive and Noncognitive Components*, http://www.sole-jole.org/13260.pdf. [5]

Borgonovi, F. and P. Biecek (2016), "An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test", *Learning and Individual Differences*, Vol. 49, pp. 128-137, http://dx.doi.org/10.1016/j.lindif.2016.06.001. [6]

Brunello, G., A. Crema and L. Rocco (2018), "Testing at length if it is cognitive or non-cognitive", *Discussion Paper Series*, No. 11603, IZA, Bonn, http://ftp.iza.org/dp11603.pdf. [10]

Gneezy, U. et al. (2017), "Measuring success in education: The role of effort on the test itself", *NBER Working Paper*, No. 24004, National Bureau of Economic Research, Cambridge, MA, http://dx.doi.org/10.3386/w24004. [4]

Goldhammer, F. et al. (2016), "Test-taking engagement in PIAAC", *OECD Education Working Papers*, No. 133, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jlzfl6fhxs2-en. [1]

OECD (2017), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, GESIS Data Archive, Cologne, http://dx.doi.org/10.4232/1.12955. [11]

OECD (2015), *OECD Survey of Adult Skills (PIAAC) (Database 2012, 2015)*, http://www.oecd.org/skills/piaac/publicdataandanalysis/. [12]

Wise, S. and C. DeMars (2005), "Low examinee effort in low-stakes assessment: Problems and potential solutions", *Educational Assessment*, Vol. 10/1, pp. 1-17, http://dx.doi.org/10.1207/s15326977ea1001_1. [3]

Wise, S. and X. Kong (2005), "Response time effort: A new measure of examinee motivation in computer-based tests", *Applied Measurement in Education*, Vol. 18/2, pp. 163-183, http://dx.doi.org/10.1207/s15324818ame1802_2. [2]

Zamarro, G., C. Hitt and I. Mendez (2016), "When students don't care: Reexamining international differences in achievement and non-cognitive skills", *EDRE Working Paper*, No. 2016-18, SSRN, Rochester, NY, http://dx.doi.org/10.2139/ssrn.2857243. [7]

Annex A. The PIAAC LogDataAnalyzer

Log files for 18 countries that participated in the first round of the Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC) (hereafter referred to as "PIAAC") in 2011/12 have recently been released and can be downloaded from the German Social Science Infrastructure Services (GESIS) Data Catalogue (OECD, 2017[1]).[1] The files have been fully anonymised to prevent identification of individual respondents. The records can nevertheless be matched with information already available in the PIAAC Public Use File, which contains the individual answers to the background questionnaire, as well as the performance of test-takers in the PIAAC assessment.

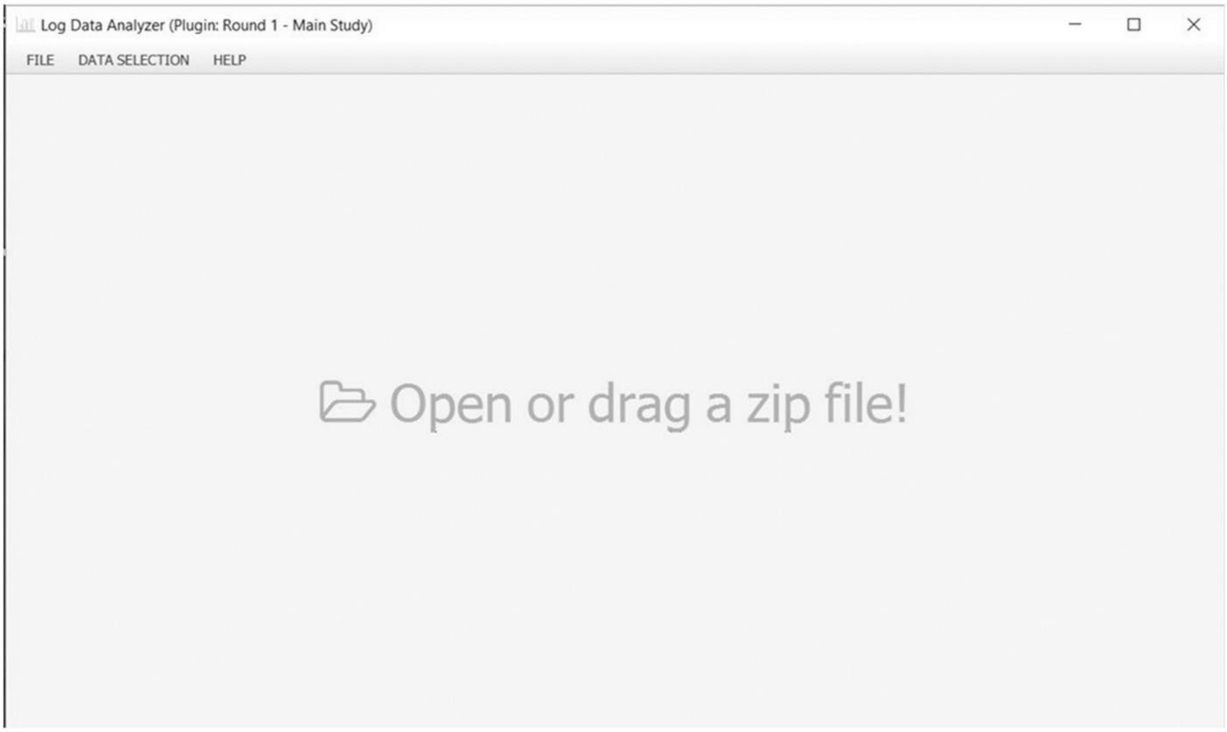
To facilitate the analysis of log files, the OECD has made available the LogDataAnalyzer (LDA), a software programme developed by GESIS. The LDA can be used to: 1) extract a number of predefined variables from the log files (which are in xml format); 2) export these variables in an external file, in txt format, which can be easily imported in the majority of software programmes used for statistical analysis; 3) compute and export descriptive statistics of the predefined variables; and 4) graphically visualise the predefined variables.

In particular, users can generate and extract the following variables:

- number of using cancel button
- number of using help menu
- time on task
- time till the first interaction
- final response
- number of switching environment
- sequence of switching environment
- number of highlight events
- time since last answer interaction
- number of created e-mails
- sequence of viewed e-mails
- number of different e-mail views
- number of revisited e-mails
- number of e-mail views
- sequence of visited web pages

- time-sequence of time spent on web pages
- number of different page visits
- number of page visits
- number of page revisits.

Figure A A.1. Initial screen of the LDA

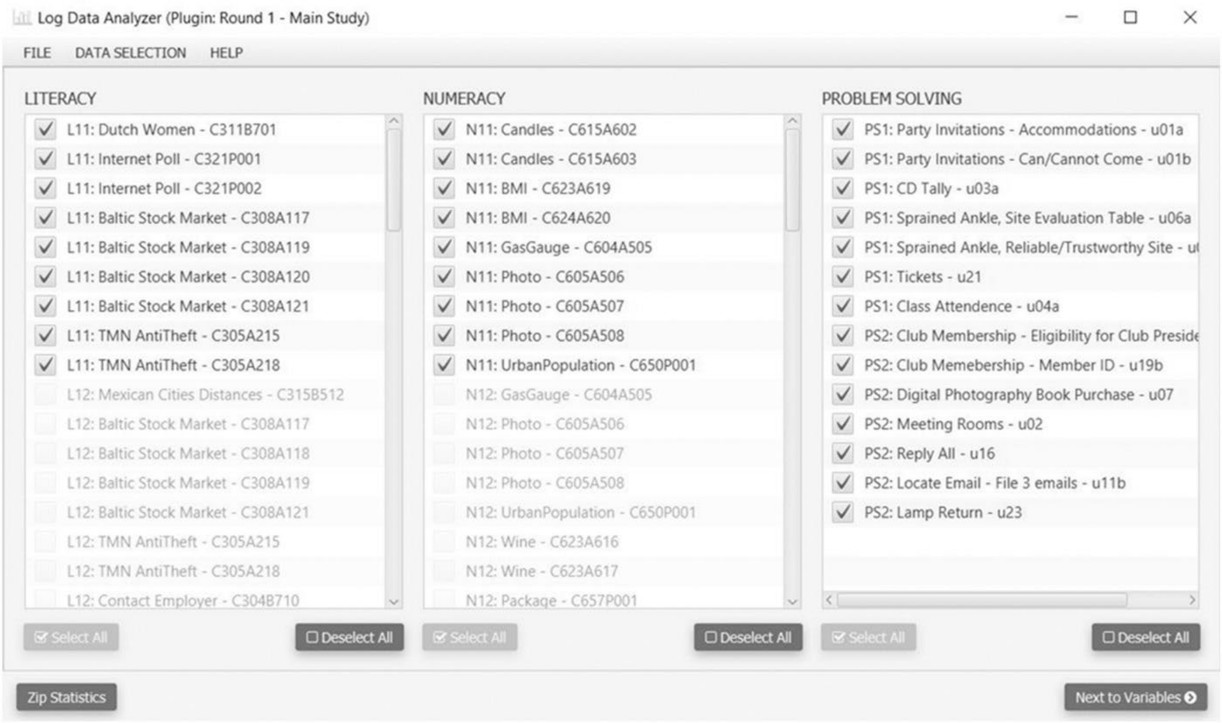


Log Data Analyzer (Plugin: Round 1 - Main Study)

Open or drag a zip file!

Once the files are loaded, it is possible to select both the items (Figure A A.2) and the variables (Figure A A.3) to be analysed. At this stage, it will become apparent that not all variables are available for all items, so available variables will depend on the items selected in the first step. The selected variables, once extracted, can be exported in the form of a text file that can then be imported easily into other software.

Figure A A.2. Selection of items in the LDA



LITERACY

- ☑ L11: Dutch Women - C311B701
- ☑ L11: Internet Poll - C321P001
- ☑ L11: Internet Poll - C321P002
- ☑ L11: Baltic Stock Market - C308A117
- ☑ L11: Baltic Stock Market - C308A119
- ☑ L11: Baltic Stock Market - C308A120
- ☑ L11: Baltic Stock Market - C308A121
- ☑ L11: TMN AntiTheft - C305A215
- ☑ L11: TMN AntiTheft - C305A218
- ☐ L12: Mexican Cities Distances - C315B512
- ☐ L12: Baltic Stock Market - C308A117
- ☐ L12: Baltic Stock Market - C308A118
- ☐ L12: Baltic Stock Market - C308A119
- ☐ L12: Baltic Stock Market - C308A121
- ☐ L12: TMN AntiTheft - C305A215
- ☐ L12: TMN AntiTheft - C305A218
- ☐ L12: Contact Employer - C304B710

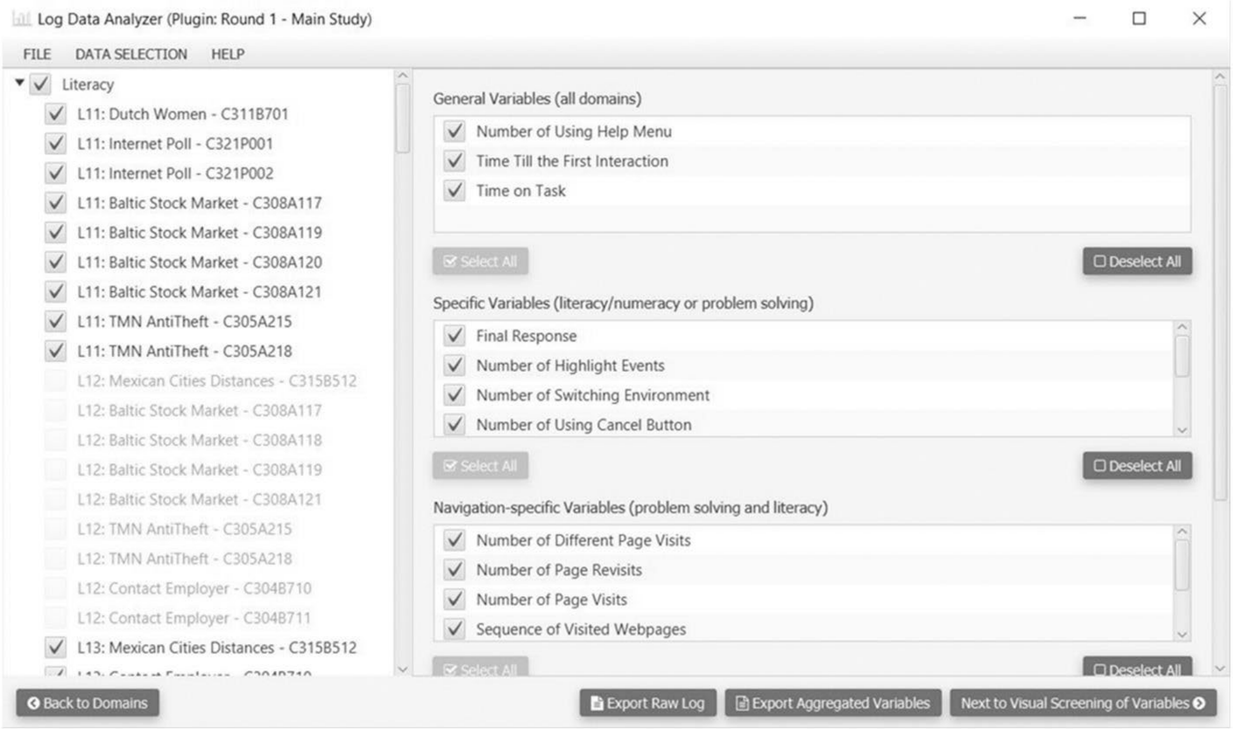
NUMERACY

- ☑ N11: Candles - C615A602
- ☑ N11: Candles - C615A603
- ☑ N11: BMI - C623A619
- ☑ N11: BMI - C624A620
- ☑ N11: GasGauge - C604A505
- ☑ N11: Photo - C605A506
- ☑ N11: Photo - C605A507
- ☑ N11: Photo - C605A508
- ☑ N11: UrbanPopulation - C650P001
- ☐ N12: GasGauge - C604A505
- ☐ N12: Photo - C605A506
- ☐ N12: Photo - C605A507
- ☐ N12: Photo - C605A508
- ☐ N12: UrbanPopulation - C650P001
- ☐ N12: Wine - C623A616
- ☐ N12: Wine - C623A617
- ☐ N12: Package - C657P001

PROBLEM SOLVING

- ☑ PS1: Party Invitations - Accommodations - u01a
- ☑ PS1: Party Invitations - Can/Cannot Come - u01b
- ☑ PS1: CD Tally - u03a
- ☑ PS1: Sprained Ankle, Site Evaluation Table - u06a
- ☑ PS1: Sprained Ankle, Reliable/Trustworthy Site - u
- ☑ PS1: Tickets - u21
- ☑ PS1: Class Attendence - u04a
- ☑ PS2: Club Membership - Eligibility for Club Preside
- ☑ PS2: Club Memebership - Member ID - u19b
- ☑ PS2: Digital Photography Book Purchase - u07
- ☑ PS2: Meeting Rooms - u02
- ☑ PS2: Reply All - u16
- ☑ PS2: Locate Email - File 3 emails - u11b
- ☑ PS2: Lamp Return - u23

Figure A A.3. Selection of variables in the LDA


Log Data Analyzer (Plugin: Round 1 - Main Study)
FILE
DATA SELECTION
HELP
Select All
Deselect All
Back to Domains
Export Raw Log
Export Aggregated Variables
Next to Visual Screening of Variables

- ☑ Literacy
  - ☑ L11: Dutch Women - C311B701
  - ☑ L11: Internet Poll - C321P001
  - ☑ L11: Internet Poll - C321P002
  - ☑ L11: Baltic Stock Market - C308A117
  - ☑ L11: Baltic Stock Market - C308A119
  - ☑ L11: Baltic Stock Market - C308A120
  - ☑ L11: Baltic Stock Market - C308A121
  - ☑ L11: TMN AntiTheft - C305A215
  - ☑ L11: TMN AntiTheft - C305A218
  - ☐ L12: Mexican Cities Distances - C315B512
  - ☐ L12: Baltic Stock Market - C308A117
  - ☐ L12: Baltic Stock Market - C308A118
  - ☐ L12: Baltic Stock Market - C308A119
  - ☐ L12: Baltic Stock Market - C308A121
  - ☐ L12: TMN AntiTheft - C305A215
  - ☐ L12: TMN AntiTheft - C305A218
  - ☐ L12: Contact Employer - C304B710
  - ☐ L12: Contact Employer - C304B711
  - ☑ L13: Mexican Cities Distances - C315B512

General Variables (all domains)

- ☑ Number of Using Help Menu
- ☑ Time Till the First Interaction
- ☑ Time on Task

Specific Variables (literacy/numeracy or problem solving)

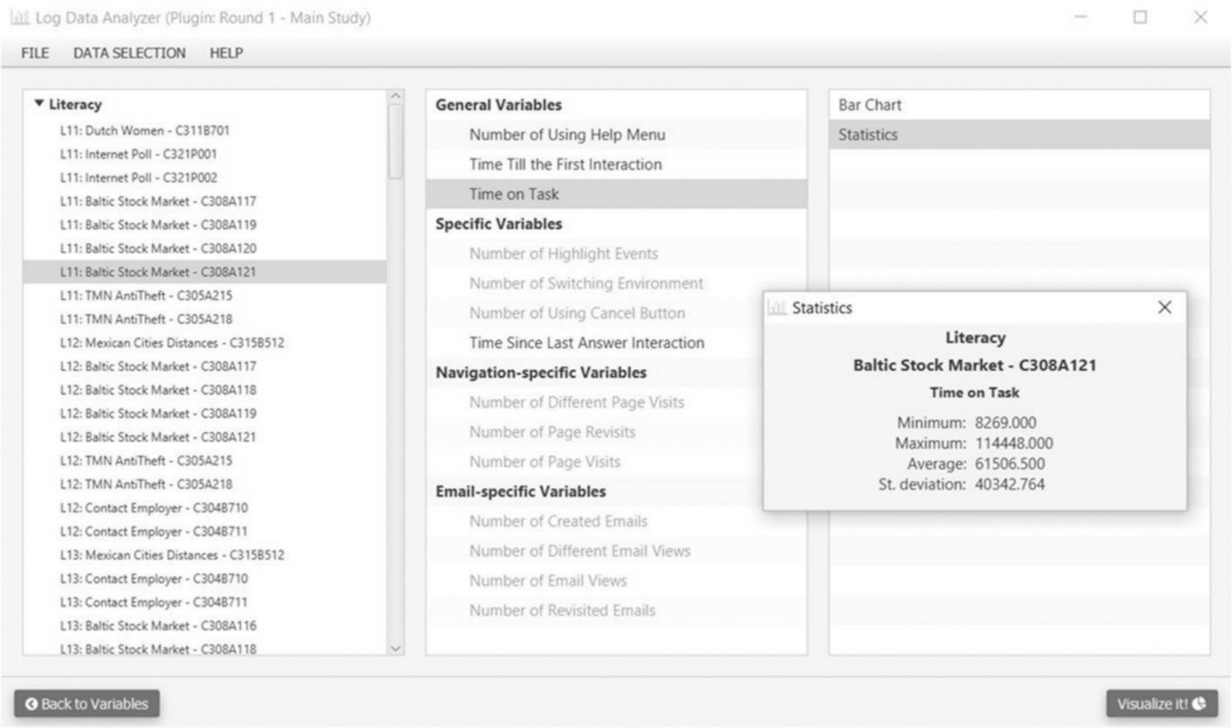
- ☑ Final Response
- ☑ Number of Highlight Events
- ☑ Number of Switching Environment
- ☑ Number of Using Cancel Button

Navigation-specific Variables (problem solving and literacy)

- ☑ Number of Different Page Visits
- ☑ Number of Page Revisits
- ☑ Number of Page Visits
- ☑ Sequence of Visited Webpages

Finally, it is possible to compute simple descriptive statistics of the variables of interest, and display them either numerically or graphically (Figure A A.4).

Figure A A.4. Analysis of variables in the LDA



▼ Literacy
- L11: Dutch Women - C311B701
- L11: Internet Poll - C321P001
- L11: Internet Poll - C321P002
- L11: Baltic Stock Market - C308A117
- L11: Baltic Stock Market - C308A119
- L11: Baltic Stock Market - C308A120
- L11: Baltic Stock Market - C308A121
- L11: TMN AntiTheft - C305A215
- L11: TMN AntiTheft - C305A218
- L12: Mexican Cities Distances - C315B512
- L12: Baltic Stock Market - C308A117
- L12: Baltic Stock Market - C308A118
- L12: Baltic Stock Market - C308A119
- L12: Baltic Stock Market - C308A121
- L12: TMN AntiTheft - C305A215
- L12: TMN AntiTheft - C305A218
- L12: Contact Employer - C304B710
- L12: Contact Employer - C304B711
- L13: Mexican Cities Distances - C315B512
- L13: Contact Employer - C304B710
- L13: Contact Employer - C304B711
- L13: Baltic Stock Market - C308A116
- L13: Baltic Stock Market - C308A118

General Variables
- Number of Using Help Menu
- Time Till the First Interaction
- Time on Task

Specific Variables
- Number of Highlight Events
- Number of Switching Environment
- Number of Using Cancel Button
- Time Since Last Answer Interaction

Navigation-specific Variables
- Number of Different Page Visits
- Number of Page Revisits
- Number of Page Visits

Email-specific Variables
- Number of Created Emails
- Number of Different Email Views
- Number of Email Views
- Number of Revisited Emails

Bar Chart
Statistics

Statistics

Literacy
Baltic Stock Market - C308A121
Time on Task

Minimum: 8269.000
Maximum: 114448.000
Average: 61506.500
St. deviation: 40342.764

Full documentation of the content of the log files is available, including information on the content of the items, which is essential to properly interpret the variables extracted. Full access is restricted, as the documentation concerns items that have not yet been used in the assessment and should therefore be treated confidentially. Interested researchers should submit an application form, including a signed confidentiality agreement, to the OECD contact officer at edu.piaac@oecd.org.

Note

[1] More information and documentation is available at www.oecd.org/skills/piaac/log-file/.

Reference

OECD (2017), *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, GESIS Data Archive, Cologne, http://dx.doi.org/10.4232/1.12955. [1]

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

The OECD is a unique forum where governments work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to co-ordinate domestic and international policies.

The OECD member countries are: Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Latvia, Lithuania, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The European Union takes part in the work of the OECD.

OECD Publishing disseminates widely the results of the Organisation's statistics gathering and research on economic, social and environmental issues, as well as the conventions, guidelines and standards agreed by its members.

OECD PUBLISHING, 2, rue André-Pascal, 75775 PARIS CEDEX 16
ISBN 978-92-64-59082-3 – 2019

OECD Skills Studies

Beyond Proficiency

USING LOG FILES TO UNDERSTAND RESPONDENT BEHAVIOUR IN THE SURVEY OF ADULT SKILLS

Computer-based administration of large-scale assessments makes it possible to collect a rich set of information on test takers, through analysis of the log files recording interactions between the computer interface and the server. This report examines timing and engagement indicators from the Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC), both of which indicate large differences across countries and socio-demographic groups, in the amount of time spent by respondents and their levels of disengagement, which reduce the probability of giving a correct answer and consequently reduces measured performance. Such insights can help policy makers, researchers and educators to better understand respondents' cognitive strategies and the underlying causes of low and high performance. This, in turn, can help improve the design of assessments and lead to more effective training and learning programmes.

Consult this publication on line at ***https://doi.org/10.1787/0b1414ed-en***.

This work is published on the OECD iLibrary, which gathers all OECD books, periodicals and statistical databases.
Visit ***www.oecd-ilibrary.org*** for more information.

ISBN 978-92-64-59082-3

9 789264 590823

OECD*publishing*
www.oecd.org/publishing