Organisation for Economic Co-operation and Development

EDU/WKP(2018)27

Unclassified

English text only

30 November 2018

DIRECTORATE FOR EDUCATION AND SKILLS

INTERVIEWERS, TEST-TAKING CONDITIONS AND THE QUALITY OF THE PIAAC ASSESSMENT

OECD Education Working Paper No. 191

By François Keslair, OECD

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Mr. François Keslair (Francois.Keslair@oecd.org)

JT03440723

OECD Education working papers series

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

---

www.oecd.org/edu/workingpapers

Abstract

This paper explores the impact of test-taking conditions on the quality of the Programme for the International Assessment of Adult Competencies (PIAAC) assessment. Interviewers record information about the room of assessment and interruptions that occurred during each interview. These observations, along with information on interviewer assignment size and a careful look at interviewer effects, provide insights into the quality of the assessment. This working paper first describes the variations in test-taking conditions among participating countries. Second, it examines interviewer assignment sizes and the frequency of interruptions, finding that both vary markedly among countries (contrary to the room of assessment). The paper then looks at the relationship between these variations and response rates and engagement measures. While neither the room of assessment nor the recorded interruptions impact quality differences among countries, interviewer assignment size and interviewer effects may have a mild impact on results.

Résumé

Ce papier explore l'impact des conditions de passation sur la qualité de l'évaluation des compétences dans le Programme pour l'évaluation internationale des compétences des adultes (PIAAC). Pour chaque interview, les enquêteurs ont noté dans quelle pièce l'interview a eu lieu et l'occurrence d'éventuelles interruptions. Ce document de travail décrit en premier lieu les variations des conditions de passation entre pays participants, puisse penche sur la charge de travail des enquêteurs et la fréquence des interruptions, qui toutes deux varient notablement suivant les pays (contrairement à la pièce où a lieu l'interview). Le papier s'achève sur une étude de la relation entre ces variations le taux de réponse et diverses mesures d'engagement. Alors que ni la pièce où a lieu l'interview, ni les interruptions enregistrées ne semblent avoir de conséquences notables sur les différences de qualité entre pays, la charge de travail des enquêteurs et les effets interviewers pourraient avoir un impact modéré.

Table of contents

Tables

Figures

1. Introduction

The Programme for the International Assessment of Adult Competencies (PIAAC) developed and conducts the Survey of Adult Skills. Managed by the OECD (OECD, 2013; OECD, 2016), this international survey aims to assess how adult information-processing skills differ both among and within countries. The comparability of the background questionnaire and cognitive test, the stringent technical standards, and wide country coverage make it a unique source of information for researchers and policy makers interested in the interplay between skills, education, and work and life outcomes. So far, 33 countries and economies have participated in the study. The study has included two waves of data collection, the first in 2011-12 and the second in 2014-15. The study involves two main components: an assessment of literacy, numeracy and problem-solving in technology-rich environments; and a background questionnaire that collects information on the correlates of the assessed skills. These correlates include the socio-demographic characteristics of respondents, like participation in education and the labour market. The study is administered by trained interviewers in the respondents' homes or at some other agreed-upon location. The background questionnaire is delivered in computer-assisted personal interviewing (CAPI) format by the interviewer. The assessment is completed directly by the respondent either on a laptop computer or using paper-based forms. Respondents can use paper-based forms if they do not wish to use a computer, or in the event that they do not have sufficient familiarity with computers to take the assessment using the computer-based format.

The quality of data from the PIAAC assessment is potentially affected both by the skills and behaviour of interviewers, and the conditions within which the PIAAC background questionnaire and cognitive assessments are administered. Interviewers are the cornerstone of data collection in PIAAC. They are responsible for establishing contact with potential respondents, and for the administration of the survey instruments. They represent the primary contact that potential respondents and survey participants have with the survey organisation that is responsible for data collection. Interviewers are also men and women whose work is centred on human contact. Thus, their work cannot be perfectly standardised. Unlike school-based assessments, where testing takes place in relatively standardised conditions, PIAAC respondents complete the assessment in their own homes or some other agreed-upon location. As a result, the conditions under which respondents take the assessment are likely to be highly variable. The variability of conditions depends on factors such as family situation and living conditions.

The purpose of this paper is to explore the role interviewers and test-taking conditions play in shaping the quality of the data gleaned from the PIAAC assessment. The paper presents descriptive statistics regarding variation among countries in terms of the structure of interviewer recruitment, the room of assessment and a set of distractions occurring during the interview, as recorded by the interviewer. It also assesses how these variations might contribute to data quality as measured by engagement levels, and, in the case of interviewers only, response rates.

Interviewers and test-taking conditions can have an important impact on data quality in a study such as PIAAC. Tools like the Total Survey Error framework (TSE) help researchers identify potential errors and design high-quality surveys. The TSE framework suggests that interviewers can contribute to survey errors in several ways (West and Blom, 2016). TSE represents the difference between the true value of a construct – skills concepts in the case of PIAAC – for the full target population, and the value estimated by the survey. TSE gives a special care to describe the various sources of errors. The TSE represents the sum of all the myriad ways in which survey measurement can deviate from the true underlying factors that the survey attempts to measure (Smith, 2011). As interviewers are in charge of the conversion of potential respondents who appear on a survey sampling plan into participating respondents, they can contribute to coverage and unit non-response errors. During the interview itself, interviewers are responsible for asking survey questions, recording the answers of respondents, and, importantly in the case of PIAAC, maintaining the respondent's motivation during the cognitive assessment. Because of this, they can contribute to measurement, item non-response and processing errors. The background questionnaire is a complex instrument. Respondents may have difficulties understanding questions, and may ask interviewers for an explanation. In turn, interviewer explanations can differ in quality. Interviewers can also code respondent answers incorrectly. Interviewers may also willingly or mistakenly skip questions. Moreover, the sheer length of the PIAAC interviews (which are close to two hours) and the potential difficulty of the skills assessments can exacerbate respondent fatigue or disengagement. Interviewers have to make sure these situations do not result in respondents failing to answer the survey instruments seriously.

In addition, the conditions in which respondents take the assessment may affect measurement and item non-response errors, largely via their effect on respondents' engagement and motivation. In the case of international surveys such as PIAAC, variation in interviewers' qualifications, training and skills across countries is one of the major contributors to TSE. This is particularly true in terms of interviewer effects on country comparisons (Hibben, Pennell and Scott, 2018).

This paper is organised in three sections. The first section presents the research framework and data. This research framework focusses on the cognitive measures in PIAAC. The paper then presents results on the interview space and interviewers. Both sections are structured in the same way. They start with a short literature review and further information on guidelines. They then present a set of descriptive statistics and an analysis of the impact of the variables in question on the rate of complaints about the length of the assessment and respondent disengagement. Both sections then conclude with an assessment of the potential for bias. The section on interviewers also contains an analysis of the impact of interviewers on response rates, and a discussion of the empirical strategy used to address this question. Item non-response and measurement error in the background questionnaire are not addressed in this paper. The paper concludes with a set of recommendations for the second cycle of PIAAC, as well as suggestions regarding the data needed to improve these types of analyses.

2. Framework and data

The conceptual framework of the analysis

Figure 2.1. Model framework

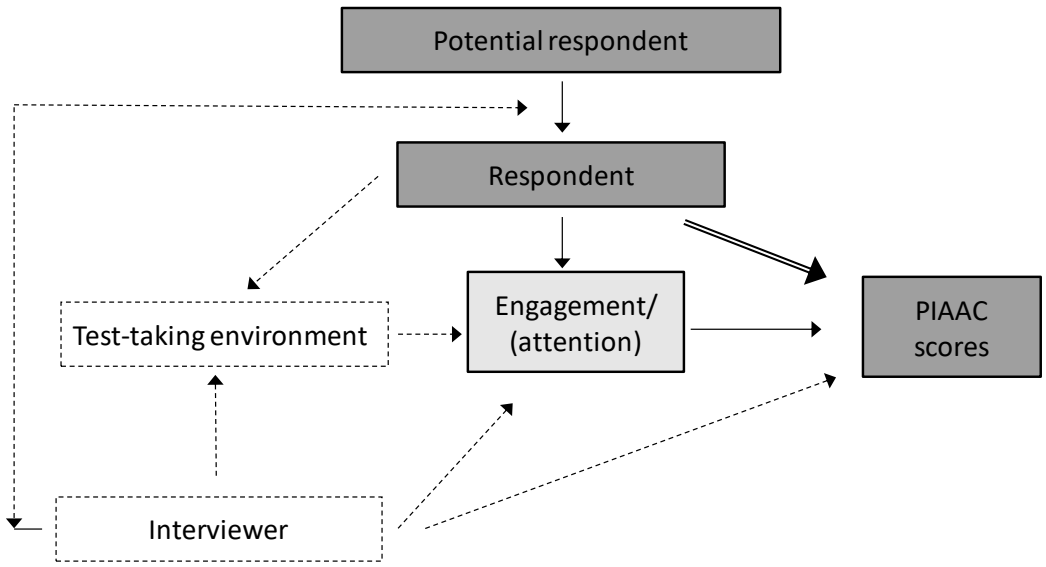


Figure 2.1 summarises the framework that underpins the analysis presented in this paper. It represents the channels through which interviewers and test-taking conditions can influence the performance of PIAAC respondents. The model focuses on the relationship between the level of engagement of the respondent and the test (assumed to be unaffected by the assessment mode). The test score is understood as a combination of the outcome of the respondent's underlying cognitive ability and his or her engagement with the assessment. The engagement of respondents is thought to be influenced by: 1) his or her disposition and motivation; and 2) the external conditions that influence the attention that the respondent can give to the test. These two factors are different, but are not distinguished empirically in the analyses that follow. A respondent could place great importance on performing at his or her best on the assessment, but could still be unable to do so because of a noisy environment. Alternatively, a respondent could take the assessment undisturbed without necessarily being motivated to perform well. In both cases, the respondent's engagement with the test is affected.

Variations in the behaviour and performance of interviewers, as well as variations in test-taking conditions, are indirect features of the PIAAC mode. Test-taking conditions are understood as influencing only the engagement and attention of respondents. The interviewer plays several roles in this framework. First, interviewers are responsible for the conversion of potential respondents (i.e. sampled individuals) into actual respondents.

In this way, they have a decisive impact on sampling and coverage error. Second, interviewers interact with the respondent during the interview. In addition to carrying out the background interview, they provide information about the purpose of PIAAC, deal with reluctant respondents, help establish the conditions under which the test is taken and explain the workings of the questionnaire and assessment. Thus, they have a direct impact on engagement of respondents and on the test-taking environment. Interviewers can also have a direct impact on PIAAC scores. This is the case when interviewers help the respondent during the assessment, for instance. Although this is notable, this paper will not consider direct interviewer impact on PIAAC scores. Interviewer behaviour and test-taking conditions are important for test scores only in so far as they have an influence on the attention, engagement and effort of the respondent.

The data

Engagement and attention

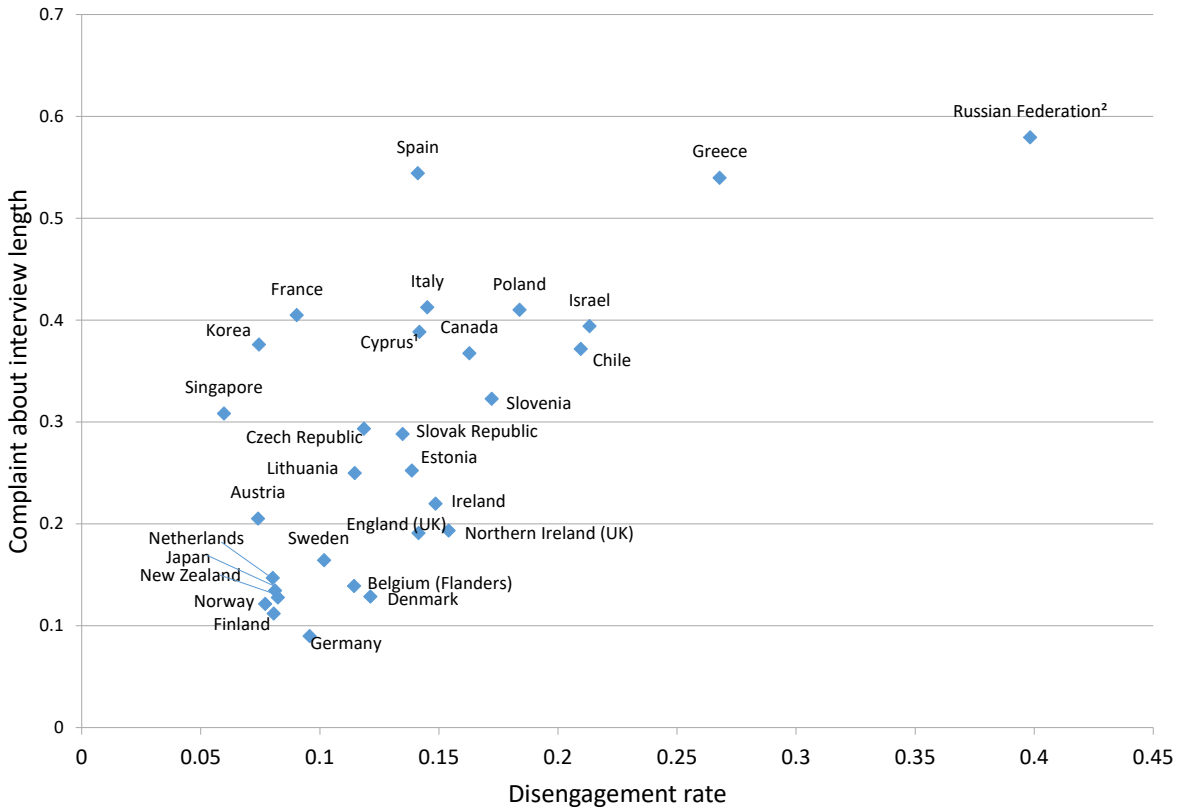
The analysis presented below uses two measures of attention and engagement. The first is whether or not the respondent complained about the length of the interview. This information is derived from a question in the observation module that is completed by the interviewer after the PIAAC interview session ends. It is blind to (not affected by) respondent responses in the background questionnaire and cognitive test. This variable is used as a proxy for the overall commitment of respondents to participation over the course of the PIAAC survey (the background questionnaire and the assessment). The hypothesis is that disaffected respondents will be most likely to complain about the length of the assessment. The variable used is not equivalent to a direct measure of engagement or attention, however. Depending on the circumstances (having to undertake family duties for instance), a respondent could complain about the survey length while still answering questions to the best of his or her ability. This measure will be called length complaint in the remainder of this paper.

The second measure of engagement or attention is based on information from the cognitive assessment. It draws on the work of Frank Goldhammer et al.'s study of test-taking engagement in PIAAC (Goldhammer et al., 2016). This measure is based on the principle that disengaged respondents are unlikely to attempt to answer certain items and are unlikely to spend enough time on these items to provide a correct answer. The measure is constructed in several steps. First, a threshold value representing the minimum time needed to provide a correct answer is computed for every item in the cognitive assessment. If a participant does not spend at least this threshold time on an item, he or she is considered disengaged and considered as not having tried to attempt the item. Second, for each respondent, this paper calculates the proportion of items not attempted in relation to the items that were assigned. Finally, a respondent is defined as disengaged if he or she did not attempt 10% or more of the items to which he or she was assigned. The 10% threshold is somewhat arbitrary and was chosen mainly in order to obtain a group of disengaged respondents large enough to be properly studied. As a result, the levels of disengagement and length complaint rates cannot be compared with each other because the former depends on this arbitrary threshold. Unlike the length complaint measure, the disengagement indicator is built on behavioural information available from the cognitive assessment. Therefore, it does not consider the level of engagement of the respondents during the background questionnaire session.

Some limitations of the disengagement measure need to be noted. First, it can only be computed for respondents who took the computer version of the assessment, as detailed

timing is not available for those who took the paper version. However, as 72% of respondents took the computer-based assessment, this measure still applies to a large majority of respondents. Second, some individuals do not attempt items that they anticipate to be too difficult. Such behaviour should not be seen as disengagement, however. Third, some respondents may not attempt to answer an item even if the time taken before moving to the next item is above the minimum time necessary to provide a correct answer.

Figure 2.2. Engagement measures (national averages)



1. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".
Note by all the European Union Member States of the OECD and the European Union:
The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.
2. The sample for the Russian Federation does not include the population of the Moscow municipal area.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

Figure 2.2 plots the national averages of the disengagement and length complaint measures. The proportion of respondents who complained about the length of the survey is 28% on average among countries. The proportion of disengaged respondents among countries is 14% on average. The proportion of participants complaining about the survey

length varies between 10% and 60%. The rate of disengagement falls between 5% and 25% in the majority of countries – with the exception of Greece and the Russian Federation. In Greece, this rate reaches 27%, and in the Russian Federation it reaches 40%. The correlation between the two measures is 0.68. At one end of the spectrum, there is a group of countries that combines very low rates of disengagement and low rates of length complaints. This group is comprised of Finland, Japan, the Netherlands, New Zealand, and Norway and to a lesser extent, Belgium (Flanders), Denmark, Germany and Sweden. Respondents in these countries appear positively disposed towards the survey. At the other end of the spectrum, Greece and the Russian Federation have high rates of both length complaints and disengagement. To a lesser extent, participants in Chile, Israel and Poland also report somewhat high rates of length complaints and disengagement.

All countries in the positively disposed group have PIAAC scores above the OECD average. Performances are generally below average in the negatively disposed group, with only the Russian Federation having above average scores. At the country level, low rates in the two measures of engagement are associated with high average PIAAC scores. To explore the relationship between these measures and performance, and to assess the potential for bias arising from test-taking conditions, simple OLS (ordinary least squares) regressions have been used at the individual level. At the individual level, engagement is associated with background characteristics that also influence literacy scores. Therefore, the effect of background characteristics on literacy has to be taken into account. Controlling for a range of covariates (gender, parental education, education levels and country of origin), disengagement is associated with a 40-point score decrease in literacy at the individual level on average among OECD countries. This figure does not vary much among countries, and an effect of similar magnitude is found for numeracy. Complaining about survey length is associated with a decrease of around eight score points. As a rule of thumb, a 10-percentage point difference in disengagement or length complaints between two countries is associated with a range of potential bias and score effects between one point and four points.

Many factors can impact engagement. For instance, in all countries, older participants and less-educated participants are more likely to be disengaged or to complain about the interview length. General attitudes towards surveys are also important, especially with regards to explaining differences among countries. This paper will focus on how environmental factors shape engagement.

Interviewers

The data available concerning interviewers is limited to an individual identifier that allows the linking of interviewers to potential and actual respondents. The existence of this link means that interviewer performance can be compared among and within countries. Unfortunately, no data on the personal characteristics of interviewers are available for use.

Potential respondents

Data on potential respondents (i.e. sampled individuals) is limited to their age and gender in most cases. These are the two the basic pieces of information necessary to determine whether a person was eligible for the survey.

Respondents

A considerable amount of information is available about respondents. This includes scores on the cognitive assessments, as well as socio-demographic and other information collected in the background questionnaire.

3. Test-taking conditions

Room of assessment

Interviewers are asked to indicate whether the interview took place in one of eight settings. These settings include different rooms or spaces within the respondent's home, as well as a location (which is not specified further) outside the respondent's home. A study of the distribution of different test-taking spaces both within and between countries is a first step to assessing variations in test-taking conditions. Unfortunately, PIAAC does not have access to information on some other relevant and related characteristics that might also affect the progress of the interview. These characteristics include the position of the interviewer relative to the respondent, for example.

Table 3.1 presents the distribution of interviews by location and country. In all countries, the majority of interviews took place within the respondent's home (91% on average). More than 80% of interviews took place in one of two rooms in the respondent's house – the kitchen (24%) or the living room (57%). These rooms are those in which guests are traditionally received in most countries. However, the balance between them varies among countries. For instance, 55% of interviews in Canada took place in the kitchen, compared to around 3% in Japan and Singapore.

Japan, Finland, Sweden and, to a lesser extent, Singapore differ from other countries in that a much lower than average proportion (around 50%) of PIAAC interviews took place in the kitchen or in the living room. In these countries, around one-third of interviews took place outside of respondents' residences. According to written comments by interviewers, these locations included respondents' workplaces, public places and also respondents' gardens. The proportion of interviews taking place outside of the respondent's residence is lower than 5% in 18 of the 32 countries for which information on the room of assessment is available. Such a wide variation could be the result of cultural differences. It could also be due to variation in the freedom granted to interviewers to negotiate a location for the interview with respondents in different countries.

The average proportion of interviews that took place in spaces different from those mentioned above is comparatively low. This was expected. That said, some countries stand out. On average among countries, 1.4% of interviews took place in the bedroom, 0.8% in the entrance and 0.5% in the hallway or corridor. 3% of interviews took place in the respondent's office, and 3.5% in a different space inside the home. 15% of interviews in Korea took place in the bedroom. This figure is notable, as it is larger than in other countries. Similarly, 7% of interviews in the Russian Federation took place in the bedroom, as well as 6% in Poland. The proportion of interviews that took place in the

office is higher than 5% only in Austria, Italy, Korea and Sweden. Very few interviews took place in spaces like the entrance, hallway or corridor of a home. This may be because these spaces are unsuited to sitting and engaging in long interviews and assessments. Still, 4.9% of interviews in Israel and 6.3% of interviews in Japan took place in the entrances to homes. These countries had the highest proportion of interviews taking place in this space. In Israel, the Russian Federation and the Slovak Republic a small proportion (3-4%) of interviews took place in the hallway or in the corridor. Finally, the proportion of interviews that took place in other kinds of spaces varies greatly between countries. Interviewer comments suggest that, in many cases, this other space is actually the dining room. It seems likely that these cases should have been included in the living room category, and thus should be regarded as misclassified. The proportion of interviews taking place in other spaces in the home is the highest in Slovenia (17%), Austria and Jakarta (Indonesia) (12%), Japan (9%), and Canada and Israel (6%).

Table 3.2 presents the results of simple OLS regressions that aim to uncover how socio-demographic determinants relate to the type of room or the location in which the assessment takes place. These variables cover age, gender, country of origin, family structure, education and parental education. Each category of space is associated with an independent regression. The coefficients must be scaled up to reflect overall likelihood of an interview occurring in the particular space in order to get a sense of their magnitude. Overall, socio-demographic characteristics appear to have little influence on the kind of space in which the interview took place. This is particularly true with regards to interviews conducted in kitchens and living rooms. The probability of an interview taking place in the living room is very similar across all demographic groups. Interviews in the kitchen are spread almost equally across these demographic groups, even though female, lower educated and older participants tend to be interviewed slightly more often here than the average. A few other findings are noteworthy. Young participants are interviewed in their bedrooms slightly more often than older respondents. An interview in the office is more likely if the participant is a man, is aged 35 or older, or has a tertiary education degree. Interviews outside of the house more often involve participants who are single, are without children, have a tertiary education degree, are men or are 25-44 years-old. These findings have intuitive explanations. Young participants may be interviewed in their bedrooms in order to leave the living room free for their parents. More educated participants are more likely to have better incomes and larger dwellings with a home office. They also often have jobs, and thus may find it easier to set up an interview at their place of work. In such cases, PIAAC considers the interview to have occurred outside the house. However, these variations across demographic groups are never large enough to create large deviations from the overall average distribution. Whatever the socio-demographic category concerned, the two most likely interview spaces are still the living room and kitchen.

Table 3.1. Room of assessment by country (percentages)

	Living or dining room	Kitchen	Bedroom	Entrance	Hallway or corridor	Office	Other space in the house	Other space outside of the house
	%	%	%	%	%	%	%	%
Austria	46.8	33.6	0.2	0.5	0.1	6.0	11.9	0.9
Belgium (Flanders)	75.8	17.0	0.1	0.0	0.0	4.0	1.8	1.3
Canada	30.6	55.5	0.2	0.6	0.3	1.7	6.4	4.7
Chile	86.1	5.2	1.6	0.8	0.7	1.3	2.2	2.0
Cyprus*	66.9	21.7	0.3	1.7	0.2	3.0	2.2	4.0
Czech Republic	46.5	37.5	0.9	0.0	1.0	3.2	2.3	8.7
Germany	66.4	20.6	0.3	0.3	0.5	4.3	3.4	4.2
Denmark	67.9	24.0	0.4	0.1	0.1	2.7	1.2	3.6
England (UK)	77.4	18.2	0.9	0.0	0.2	1.0	1.3	0.9
Spain	74.3	10.6	1.5	0.9	0.1	2.5	2.6	7.5
Estonia	44.6	24.5	1.7	0.1	0.3	4.5	1.6	22.7
Finland	21.8	35.8	0.9	0.0	0.0	2.7	1.0	37.8
France	73.1	16.6	0.6	0.0	0.2	2.3	2.1	5.2
Greece	71.2	16.9	1.2	1.1	0.5	1.5	3.8	3.9
Ireland	56.4	40.1	0.3	0.0	0.2	1.2	1.1	0.7
Israel	57.7	13.5	1.2	4.9	3.9	1.9	5.9	11.0
Italy	46.2	35.0	0.4	0.5	0.1	5.2	2.7	9.9
Japan	42.8	3.2	0.4	6.3	0.1	2.3	9.1	35.8
Korea	60.5	9.6	14.8	0.2	0.0	7.7	0.7	6.5
Lithuania	48.8	36.7	0.4	1.1	1.1	2.1	5.1	4.7
Northern Ireland (UK)	64.5	33.2	0.1	0.0	0.1	1.1	0.7	0.3
Netherlands	80.9	14.9	0.3	0.0	0.0	2.0	1.2	0.7
Norway	68.8	26.9	0.5	0.0	0.0	1.6	2.1	0.0
New Zealand	79.9	11.1	1.2	0.2	0.0	2.2	1.7	3.7
Poland	59.8	23.7	5.7	0.3	0.6	1.8	3.8	4.4
Russian Federation	44.2	33.6	7.5	2.4	2.9	2.8	2.7	4.0
Singapore	62.5	3.2	1.5	0.1	0.5	2.4	1.3	28.5
Slovak Republic	47.3	31.1	0.4	0.7	4.0	2.9	2.8	10.8
Slovenia	33.1	33.2	0.6	1.4	0.8	2.6	17.4	10.9
Sweden	17.4	36.2	0.4	0.0	0.0	6.2	0.8	39.1
United States**								
Average	57.3	24.1	1.5	0.8	0.6	2.9	3.4	9.3

Note: * See note 1 under Figure 2.2. ** Room of assessment data is not available for the United States. The sample for the Russian Federation does not include the population of the Moscow municipal area.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

Table 3.2. Room of assessment and background characteristics (OLS country average)

		Living or dining room	Kitchen	Bedroom	Entrance	Hallway or corridor	Office	Other space in the house	Other space outside of the house
Age	16-24	-1.4	-0.5	0.9	0.0	0.2	0.5	1.0	-0.7
	25-34	ref	ref	ref	ref	ref	ref	ref	ref
	35-44	-2.2	0.6	-0.2	-0.1	0.1	1.5	0.1	0.3
	45-54	-2.8	1.1	-0.3	-0.1	0.2	2.0	0.4	-0.6
	55-65	-1.7	2.1	-0.5	0.0	0.3	1.8	0.5	-2.5
Parental education	Neither parent has attained upper secondary	-0.8	2.0	-0.1	0.3	0.0	-0.5	-0.2	-0.7
	At least one parent has attained secondary education	ref	ref	ref	ref	ref	ref	ref	ref
	At least one parent has attained tertiary education	-0.3	-1.6	0.0	-0.2	-0.1	0.5	0.4	1.3
Education	Below upper secondary	-1.2	2.9	0.4	0.0	-0.1	-0.9	-0.1	-1.1
	Upper secondary	ref	ref	ref	ref	ref	ref	ref	ref
	Tertiary education	-0.7	-3.8	-0.1	-0.1	-0.1	2.4	0.0	2.5
Men		-0.9	-3.2	0.0	0.1	0.1	1.5	0.6	1.7
Foreign born		3.3	-5.0	1.4	-0.1	0.1	-0.3	-0.2	0.8
Living with a partner		1.3	1.3	-0.9	-0.2	-0.1	0.5	-0.1	-1.9
Has a child younger than 16		-0.8	3.0	0.0	0.1	0.1	-0.3	-0.1	-1.9
Average		57.3	24.1	1.5	0.8	0.6	2.9	3.4	9.3

Note: Every column stands for a regression model with the column head as a dependent variable and the row names as control variables. 'ref' is the omitted category. The very last row is not part of the regressions but is a reminder of the average proportion of interview taking place in each type of rooms. OLS: ordinary least squares.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

Distractions

In addition to collecting information on the room of the house in which the interview takes place, the interviewers also collect other information in the observation module. This information includes events that occurred during the interview, and the activities in which the respondent was engaged. This information helps interviewers document test-taking conditions that may have distracted respondents and could have potentially affected their performances on the cognitive assessment. Interviewers are asked to indicate the following events or activities: the respondent held a conversation with someone else; the respondent answered a phone call, text message or e-mail; the respondent was looking after children; the respondent was undertaking domestic tasks; a television, radio, game console or music player was in use in the vicinity of respondent; the respondent was interrupted by some other activity, task or event.

This list merits a few comments. First, these events and activities represent different types of distractions. All but one represent events or activities that cause an interruption to the assessment. The use of a television set (or other media device) in the vicinity of the interview is somewhat different. A television is more likely to create an environment in which it is difficult to concentrate, rather than an interruption. Second, depending on the

circumstances, most of these events could be considered either as a constraint or a choice. For instance, the events fitting in the "a conversation with someone else" category go from a simple chat that could be postponed to more a practical and necessary exchange relating to household life. Third, no information is available regarding the timing and duration of the interruptions. A phone call could last ten seconds or ten minutes. This type of information is important. It helps to discriminate between behaviours that suggest a lack of engagement, and those that do not. This type of information could also give a sense of the importance of the interruption. Because more information is not available, however, each of these events and activities could be significant or trivial in their effects on test performance, depending on the circumstances. Finally, the responses to these questions are unavoidably influenced by the perceptions of interviewers. Nonetheless, this list re-groups a set of events which is diverse, informative and potentially relevant – despite the potential noise in the measurement. These events are valuable sources of information that help researchers compare the frequency of interruptions among countries. Finally, the observation module asks interviewers to report the presence of another person during the interview. This information is important, since most interruptions imply the presence of another person, including possible interruptions about which information is not captured. As a result, this variable is also included alongside the set of interruptions.

Table 3.3 presents the probability of each type of potentially distracting event or activity by country. On average, 32.8% of interviews took place with another person present in addition to the interviewer. Variation across countries is large, with the lowest proportion of interviews taking place with another person present in Japan (11%) and the highest proportion in Israel (73%). Even though the interviewer may play a role, the presence of another person during the interview is likely to be related to the size and design of dwellings, as well as social norms and habits. Unfortunately, it is difficult to determine how much these country-level differences stem from cultural differences or from variations in interviewer behaviour.

On average among countries, a potentially distracting event or activity was noted in 36% of interviews. When considered together, interruptions are as frequent as the presence of another person. The most widespread interruption appears to be a conversation with someone. This interruption occurs in 15% of interviews. In order of decreasing frequency, occurrences of other interruptions during interviews include: an electronic communication or phone call (12.3%), the use of an entertainment device such as a TV set or video game console (8.9%), other kinds of interruptions (7.3%) and, with a very low frequency, domestic tasks (1.2%). Countries where the interview frequently takes place in the presence of another person tend to have higher overall rates of interruptions as well. Chile, Cyprus[1] and Israel rank high on both lists. The rate of uncategorised interruptions is above 10% in six countries only. Evidence from interviewers' comments suggest these non-categorized interruptions could be coffee breaks, visits to the bathrooms, or interruptions that are related to the presence of a pet.

Rates for individual interruptions vary at the country level. They are generally all high (as in Israel and Cyprus[2]) or all low (as in Germany and Denmark). This pattern can be confirmed by a simple principal component analysis applied to the set of country rates of

[1] See note 1 under Figure 2.2.

[2] See note 1 under Figure 2.2.

interruption. A single underlying factor explains 74% of their variance. Moreover, at the individual level within each country, the correlations between different interruptions are very low. None of the average pairwise correlations are above 0.15 in absolute terms.

Table 3.3. Interruptions by country (percentages)

	Interruptions							
	Other person present	Conversation with someone	Phone call, e-mail or SMS	Looking after child	Domestic tasks	Television	Others	At least one distraction
	%	%	%	%	%	%	%	%
Austria	24.6	11.8	8.7	2.7	0.3	6.1	3.9	26.0
Belgium (Flanders)	27.0	5.2	4.9	2.9	0.3	5.4	3.3	18.0
Canada	45.0	24.5	16.3	8.0	2.8	17.6	12.8	50.0
Chile	62.5	41.9	16.3	13.4	3.3	24.8	14.2	62.7
Cyprus*	50.7	27.9	20.1	6.5	2.1	15.2	3.8	52.3
Czech Republic	30.1	15.6	9.6	7.0	0.8	10.2	8.4	37.5
Germany	20.8	5.4	7.9	2.3	0.2	2.9	4.2	18.9
Denmark	18.2	8.2	7.4	3.1	0.5	3.6	4.9	22.5
England (UK)	27.1	14.1	11.2	6.5	1.0	7.4	8.4	34.9
Spain	34.2	19.0	12.3	5.3	1.0	10.5	8.0	41.9
Estonia	22.7	12.3	17.3	5.7	1.4	13.9	7.8	41.7
Finland	16.0	12.4	12.5	5.2	1.0	4.4	9.6	32.5
France	27.9	9.3	9.7	4.3	0.5	6.0	4.2	26.6
Greece	34.5	9.5	11.7	3.9	1.0	7.2	1.8	27.4
Ireland	30.5	14.9	7.9	8.6	1.3	7.6	6.8	33.6
Israel	73.4	32	20.4	10.2	2.0	11.9	8.6	53.3
Italy	37.0	7.2	8.3	1.8	0.1	1.8	2.0	19.5
Japan	11.3	6.3	7.9	3.8	0.2	4.5	7.6	24.6
Korea	45.7	17.4	16.3	4.5	1.2	12.5	4.3	42.8
Lithuania	45.4	19.9	16.8	5.2	0.9	13.7	3.0	47.6
Northern Ireland (UK)	24.3	13.2	12.8	6.4	0.8	8	6.8	35.5
Netherlands	24.4	5.5	7.6	2.7	0.3	10.6	5.5	26.2
Norway	19.8	8.2	9.5	3.1	0.3	3.3	7.1	25.0
New Zealand	34.7	22.1	11.3	7.5	1.5	8.8	10.1	40.2
Poland	47.4	14.2	11.4	5.0	1.4	7.5	7.1	35.2
Russian Federation	36.9	26.6	22.2	8.9	6.7	13.3	20.3	66.9
Singapore	37.0	14.8	16.0	3.1	0.5	6.5	9.6	36.8
Slovak Republic	21.7	9.4	9.8	4.1	0.5	7.3	4.4	30.5
Slovenia	27.3	15.7	10.2	2.9	0.7	7.1	6.9	34.0
Sweden	17.5	9.3	12.8	4.2	0.9	2.5	10.0	29.7
United States	41.6	18.7	14.7	8.3	1.4	12.8	12.5	45.9
Average	32.8	15.2	12.3	5.4	1.2	8.9	7.3	36.1

Note: * See note 1 under Figure 2.2.
The sample for the Russian Federation does not include the population of the Moscow municipal area.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

Table 3.4. Interruptions and background characteristics (OLS country average)

		Interruptions							
		Other person present	Conversation with someone	Phone call	Looking after child	Domestic tasks	Television	Other	At least one of them
Other person present			29.5	0.0	10.5	0.9	9.5	4.3	30.6
Age	16-24	6.2	-1.1	-1.5	-3.8	-0.5	-2.5	-1.6	-6.4
	25-34	ref	ref	ref	ref	ref	ref	ref	ref
	35-44	-5.5	1.2	0.8	-4.8	0.2	-0.5	0.4	-1.1
	45-54	-8.4	2.0	1.9	-7.2	0.3	-1.0	0.8	-0.4
	55-65	-6.5	2.1	0.7	-4.6	0.3	-2.1	0.9	0.0
Parental education	Neither parent has attained upper secondary	3.0	0.2	-1.1	0.4	0.0	0.1	-0.3	-0.6
	At least one parent has attained secondary education	ref	ref	ref	ref	ref	ref	ref	ref
	At least one parent has attained tertiary education	-2.4	-0.2	-0.2	-0.1	0.0	-0.6	0.1	-0.7
Education	Below upper secondary	8.6	0.9	-1.8	-0.4	-0.1	0.3	-0.2	-0.2
	Upper secondary	ref	ref	ref	ref	Ref	ref	ref	ref
	Tertiary education	-4.9	-0.8	0.4	0.0	0.0	-1.8	0.1	-1.3
Men		0.9	-0.9	0.4	-5.0	-1.2	-0.1	-0.1	-3.2
Foreign born		7.4	2.1	1.3	0.1	0.3	0.3	0.6	2.5
Living with a partner		11.0	1.6	-1.0	0.9	0.0	-0.9	-0.6	0.1
Has a child younger than 16		9.7	1.2	1.5	10.5	0.3	0.5	1.2	7.2
Average		32.8	15.2	12.3	5.4	1.2	8.9	7.3	36.1

Note: Every column stands for a regression model with the column head as a dependent variable and the row names as control variables. 'ref' is the omitted category. The very last row is not part of the regressions but is a reminder of the average proportion of interview taking place in each type of rooms. OLS: ordinary least squares.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

Table 3.4 completes this description of interruptions and presents their relationships to individual background characteristics. The pattern is rather simple. There is a strong link between individual background characteristics and the presence of another person (first column). There is also a strong link between the presence of another person and some interruptions, but once this latter is accounted for, almost no direct link between individual characteristics and interruptions is estimated. As a result, the link between individual background characteristics and interruptions is mostly driven by the presence of another person. The presence of another person relates in a very predictable manner to the household composition. Interruptions are much more likely if the participant is living with a partner, or if he or she has a child under the age of 16. Socio-demographic characteristics matter as well. Educated participants tend to be interviewed alone more often. The opposite is true with respect to foreign-born participants. These patterns might result from income-related differences in the size of dwellings. The age of a respondent is also strongly related to the probability of being interviewed alone. Older participants are more often alone with the interviewer than younger ones. With all of these factors accounted for, parental education has only a second order effect on the likelihood that another person will be present for the interview. In addition, with participants originating from a lower social background, this is more often the case. Table 3.4 shows a set of similar regressions documenting interruptions, from column two to column eight. Both

tables clearly demonstrate that the presence of another person is a strong source of interruptions, with dramatic effects on all interruptions other than electronic communications. Most of these interruptions actually imply the presence of another person, as there is no conversation possible without the presence of another person, for instance. Nonetheless, socio-demographic characteristics have little impact on interruptions. Once the presence of another person is accounted for, age, education, parental education, gender and country of origin have little effect on the level of interruptions. The only noteworthy effect is actually the link between household structure and childcare.

Impact on engagement and attention

The differences observed among participants regarding the location in which the assessment took place and the incidence of distracting events or activities matters to the extent that they have an influence on attention and engagement, as well as on test scores. As we have no way of addressing the question of causality, we show instead the results of separate OLS regressions for each of our two measures of engagement and attention as dependent variables with: 1) the room of assessment; and 2) the set of distracting events or activities as covariates. As shown above, they are associated with socio-demographic variables. In order to avoid confounding effects, we also add this set of variables as controls. Unfortunately, as sample sizes are too small to reliably study these relationships at the country level, we only show country averages here. Importantly, the room of assessment and the occurrence of interruption are not exogenous. It is possible, for example, that disengaged respondents are less concerned than others with choosing an environment that facilitates maximum performance in the assessment.

Table 3.5. Impact of the room of assessment on disengagement and complaints about interview length (OLS country average)

	Living or dining room	Kitchen	Bedroom	Entrance	Hallway or corridor	Office	Other space in the house	Other space outside of the house
Disengagement	ref	**1.3**	**4.9**	**13.2**	4.8	0.9	0.8	-0.3
Complaints regarding interview length	ref	0.8	2.9	2.3	**8.3**	**2.5**	**2.2**	1.2

Note: Each row stands for one regression model with the row head as a dependent variable, a set of dummies for the rooms of assessment and a set of additional controls matching those in Table 3.4. Only coefficients for the different rooms of assessment are shown. 'Ref' is the omitted category. OLS: ordinary least squares.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

The association between the room of assessment and engagement or attention is significant only in a few situations. The room of the interview is related to disengagement only in the case of the bedroom and the entrance. The probability of being disengaged is higher for respondents who undergo their interviews in these rooms than for respondents who undergo the interview in the living or dining room. When we look at complaints about interview length, there are significant differences between the living or dining room and the hallway or corridor (8.3) and to a lesser extent, the office and other space in the house (2.5 and 2.2 respectively).

Table 3.6. Impact of interruptions on disengagement and complaints about interview length (OLS country average)

	Interruptions						
	Other person present	With someone	Phone call	Looking after child	Domestic tasks	Television	Other
Disengagements	0.7	1.4	0.2	2.2	4.3	0.5	0.4
Complaints regarding interview length	0.2	3.3	0.8	1.8	16.7	0.3	1.4

Note: Each row stands for one regression model with the row head as a dependent variable, a set of dummies for the interruptions and a set of additional controls matching those in Table 3.4. Only coefficients for the interruptions are shown. OLS: ordinary least squares.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

Interruptions are only weakly associated with disengagement or complaints about interview length. In either case, the presence of another person does not seem to matter. Their presence is associated with a difference of only 0.7 percentage points with regards to the likelihood of disengagement. Furthermore, the presence of another person is only associated with a difference of 0.2 percentage points in terms of the probability of lodging a complaint about the length of the interview. With regards to distracting events and activities, around half are associated with either disengagement or interview-length complaints. In addition, domestic tasks have a strong link to disengagement (4.3 percentage-point difference) and to complaints about interview length (16.7 percentage-point difference). To a lesser extent, having a conversation with someone is associated with higher rates of disengagement (1.4 percentage-point difference) and interview-length complaints (3.3 percentage-point difference), as is looking after a child.

The variations observed among countries are unlikely to create significant biases in the estimation of proficiency scores at the country level. For a given country, the potential for bias is the product of three figures. The product of: 1) the effect of the outcome on the numeracy and literacy score; and 2) the effect of interruptions and rooms of assessment on the outcome gives the potential bias for an individual. In order to calculate the potential bias for the country population average, this figure needs to be multiplied by: 3) the proportion of the population affected by a distracting event or activity, or affected by being assessed in a given room. Several assumptions underlie this formula. The two associations (between outcomes and proficiency scores, and between outcomes and events) are not country dependent. They are also considered as upper bounds for causal effects. All in all, the threat for bias appears to be negligible because the situations associated with a sizeable increase in disengagement or complaints are all rare. Thus, they have a tiny aggregated impact. For instance, interviews occurring in the entrance of a home are associated with a 13.2 percentage-point increase in disengagement, which yields a decrease of around five points in literacy and numeracy scores. However, the country where this interview setting is the most frequent is Japan, where 6% of interviews happen in the entrance to a home. This results in a potential downward bias of 0.3 points on average. Among all combinations of an engagement measure, a country and a room of assessment, this one is associated with the largest bias.

4. Interviewers

Interviewers have long been recognised by survey researchers as having a decisive impact on survey quality. Loosvelt (2008) gives an overview of the interviewer's role and position in survey data collection, with a brief theoretical account of interviewer effects. In their research synthesis, Brady West and Annelies Blom (2016) summarise this literature with a focus on interviewer characteristics that explain interviewer effects. These characteristics include experience, gender, education, non-cognitive skills and specific training. Only a small share of this literature deals with large-scale international surveys and compares interviewers working on the same survey, but in different countries. Thus, the European Social Survey (ESS) is a rare and valuable source, as it collects detailed information about interviewers and their work. This information is all the more valuable given the considerable overlap between the countries participating in the ESS and those participating in PIAAC (all European countries present in PIAAC are also present in the ESS). Data from the ESS shows systematic differences in interviewer characteristics exist between countries in terms of gender, experience and age. For instance, the mean for years of experience of ESS interviewers is 12.7 years in Finland, compared to 1.8 years in Switzerland (Blom, de Leeuw and Hox, 2011).

Ideally in these types of surveys, interviewers should not influence the data collected, and interviewer effects should be kept at a minimum. However, systematic differences in the magnitude of interviewer effects have also been documented by various researchers (Blom, de Leeuw and Hox, 2011; Beullens and Loosveldt, 2016). Blom, de Leeuw and Hox. (2011) study variation in non-response rates among countries participating in the ESS. Interestingly, they distinguish between establishing contact with a respondent and converting this contact into a successful interview subject. They also show that interviewers explain a much larger share of the dispersion of contacts (27%) than co-operation (8%). Beullens and Loosveldt (2016) analyse interviewer effects on ESS survey instruments. They find that interviewers barely affect measurement in some countries, such as Finland and Denmark. In other countries, like the Russian Federation and Greece, interviewers have considerably more of an effect on measurement.

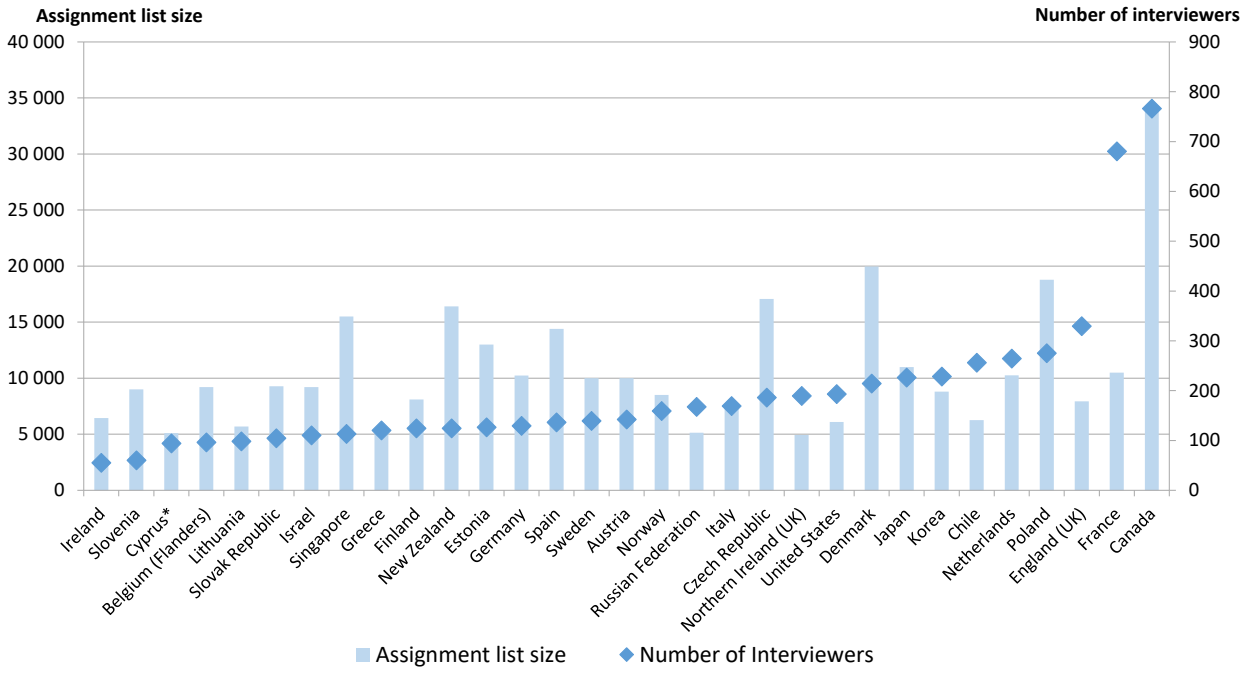
Similar international comparisons of differences in interviewer characteristics and the effect these differences may have on data quality in PIAAC are lacking; thus, this paper attempts to fill the gap. A few country-specific studies have looked at the work of PIAAC interviewers without estimating any direct interviewer effects. One study uses audio recordings of German interviewers and shows that they do not systematically follow instructions (Ackermann-Piek and Massing, 2014). For instance, interviewers incorrectly skip at least two questions in one out of four recorded interviews. For a forthcoming study (Maddox, Javrh and Keslair, forthcoming), a researcher accompanied interviewers during data collection in Slovenia in order to observe the interaction between respondent, interviewer and the CAPI software in the field. Importantly, the study showed that interviewers can play a crucial role in maintaining motivation to finish the cognitive assessment among less skilled respondents. These important tasks draw, to a large extent,

on the social skills and non-verbal communication interviewers. As such, they cannot be standardised.

In order to minimise interviewer effects, the PIAAC Technical Standards and Guidelines provide guidance to participating countries regarding the recruitment and training of interviewers, interviewer workloads and remuneration, and the monitoring of interviewer performance. In particular, given the complexity of PIAAC instruments, countries receive a training package that covers general interviewing techniques, the details of PIAAC and computer-assisted interviewing.

This standardisation has obvious limits, however. The most important are the diversity in the labour market for interviewers, and their conditions of employment. Interviewers are, in some cases, long-term employees of the institute in charge of data collection (as is the case in France). In other cases, interviewers may be recruited directly through short-term contracts, or via a global contract signed with a specialised firm. Thus, the skills, education levels and experience interviewers have varies among countries. Moreover, in most countries, interviewers are not recruited at the national level, but rather at the regional level in order to avoid long travel distances.

Figure 4.1. Number of interviewers and assignment list size, by country

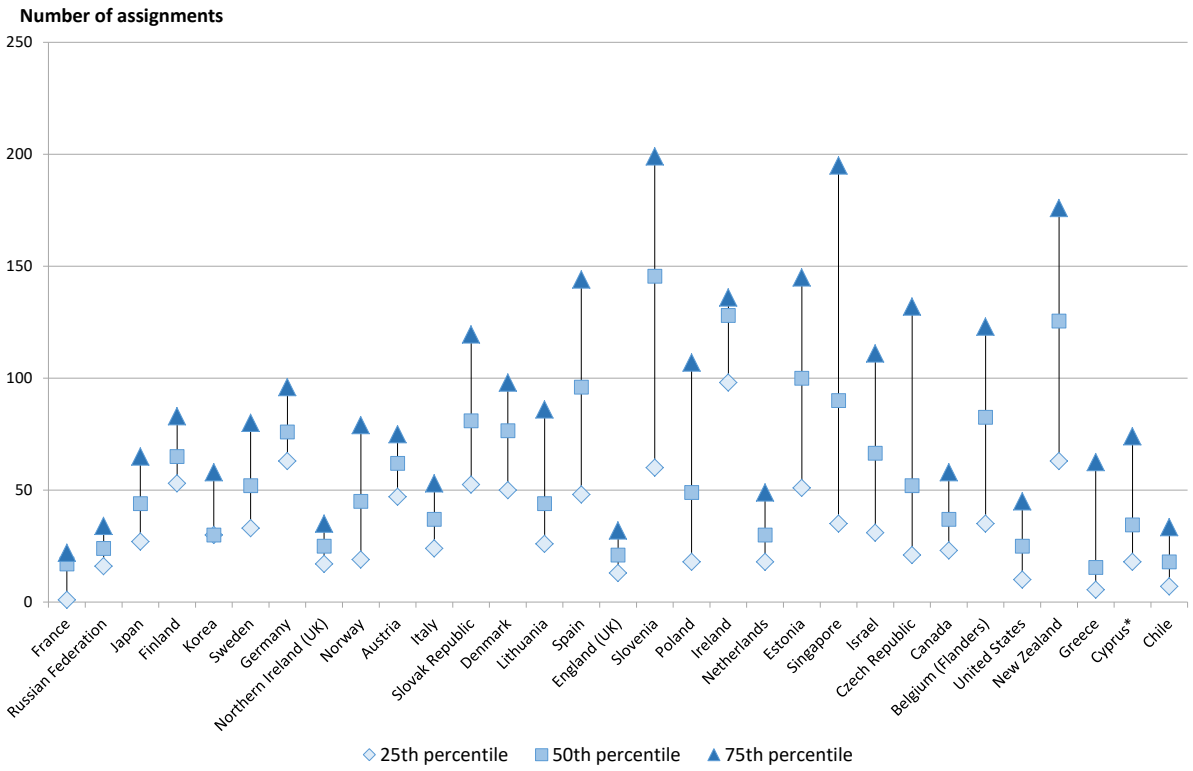


Note: * See note 1 under Figure 2.2.
The sample for the Russian Federation does not include the population of the Moscow municipal area.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

As shown in Figure 4.1, the number of interviewers varies greatly by country and is not that closely related to total assignment size. In most countries, between 100 and 200 interviewers were recruited for PIAAC. However, a sizeable gap exists between countries with small numbers of interviewers and those with large numbers. The number of interviewers who worked on PIAAC in Ireland was 55; in Slovenia, it was 60. In

comparison, 680 interviewers were employed in France and 766 in Canada. As such, France and Canada employed ten times as many interviewers as Ireland and Slovenia. The sheer difference in the number of interviewers employed implies a qualitative difference in the management challenge. Figure 4.1 also shows assignment list sizes. Assignment list sizes are different from sample sizes in so far as they represent the total number of persons to be contacted for the PIAAC assessment (including non-respondents), rather than completed interviews. Canada and France stand out as having very high interviewer numbers. However, they differ considerably in assignment size. Overall, interviewer numbers and assignment size are not very closely related (the spearman correlation is 0.28). As a result, their ratio and the number of assigned interviews per interviewer are country specific. Its distribution is studied in the following paragraph.

Figure 4.2. Distribution of interviewer assignment size, by country



Note: * See note 1 under Figure 2.2.
The sample for the Russian Federation does not include the population of the Moscow municipal area.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

Figure 4.2 presents the country distributions of interviewer assignment size, and shows the 25th, median and 75th percentiles. These figures are closely related to the interviewer design effects, which measures how interviewers affect the representativeness of the sample. An increase in the average number of interviews by interviewer lowers the representativeness of the sample. At one extreme, we have Ireland, New Zealand and Slovenia, where the median interviewer is assigned more than 120 potential respondents. At the other extreme, in France, the median interviewer is in charge of only 17 potential interviews. These extreme cases exemplify two different choices for the management of

data collection: 1) have a few interviewers working on many assignments; or 2) allocate many interviewers to work on a few assignments, at the expense of an increase in sample size. Between these two extremes, a range of possible trade-offs exist. The variation in the median interviewer assignment among countries shows that none of the possible combinations is favoured. There are seven countries that assign 15-25 interviews per interviewer. Nine countries assign 25-50 interviews per interviewer. Five countries assign 50-75 interviews per interviewer, and seven countries assign 75-100 interviews per interviewer. Ireland, Slovenia, and New Zealand all have a median of more than 120 assignments per interviewer.

Figure 4.2 also shows the distribution of interviewer assignment size within each country. Again, we can see much variation among countries. In some countries, the difference between the interviewer ranked at the 25th percentile and the one ranked at the 75th one is quite small. This difference is around 20 interviews in the Russian Federation, England (UK), Northern Ireland (UK) and France. On the contrary, in Slovenia, Singapore, New Zealand and the Czech Republic, this difference exceeds 100 interviews. These large interquartile ranges are driven by exceedingly high 75th percentiles, as in Slovenia and Singapore, where the 75th percentile is close to 200 interviews. This means that about one-quarter of interviewers in these countries are assigned more than 200 interviews. In contrast, in Greece, France and Chile, the 25th percentile is below 10. In these countries, a sizeable proportion of interviewers ended up with a light workload.

These important variations epitomise the variety in the local supply of interviewers. In particular, the breadth of the distribution of interviewer assignments is related to interviewer attrition. Interviewers who stop working have their assignments re-allocated to interviewers who are still working, and this transfer widens the distribution. Interviewer attrition rates – whether they are due to interviewers quitting their jobs or being laid off – are detailed in the PIAAC technical report. Attrition rates appear to be related to greater variation in assignment size. In the 12 countries in which attrition was less than 10%, the interquartile range was 35 on average. In the eight countries in which this rate was above 50%, the average interquartile range was 72. The overall time period during which data is collected is another factor that affects assignment size. Data collection generally ranged from around 180 days in the Slovak Republic to 327 days in Greece, with a typical length of 250 days. France is an outlier with data collection lasting only 79 days. In France, the very narrow distribution of interviewer assignments is, to a large part, a consequence of this very short data collection period. The variations exhibited above show the diversity of approaches to interviewer recruitment, both among and within countries.

The analysis of interviewer effects on survey quality measures involves the framework of an econometric model in order to disentangle the contributions of individuals, interviewers and sampling units. The previous analysis studying interruptions and rooms of assessment covered engagement and attentions measures on which interviewers could have an impact as well. However, on top of appropriately conducting interviews, interviewers are also in charge of contacting potential respondents. As a result, interviews can also influence response rates. We apply the following analysis to both outcomes, (response rates and engagement measures). In a given country, the general shape of the model that we use is the following:

$$Y_{ijs} = \alpha_c X_i + \beta_c X_j + \gamma_c X_s + \varepsilon_i + \varepsilon_j + \varepsilon_s$$

Subscript I stands for the individual (respondent or potential respondent). Subscript j stands for the interviewer. And subscript s stands for the sampling unit. Y is the outcome

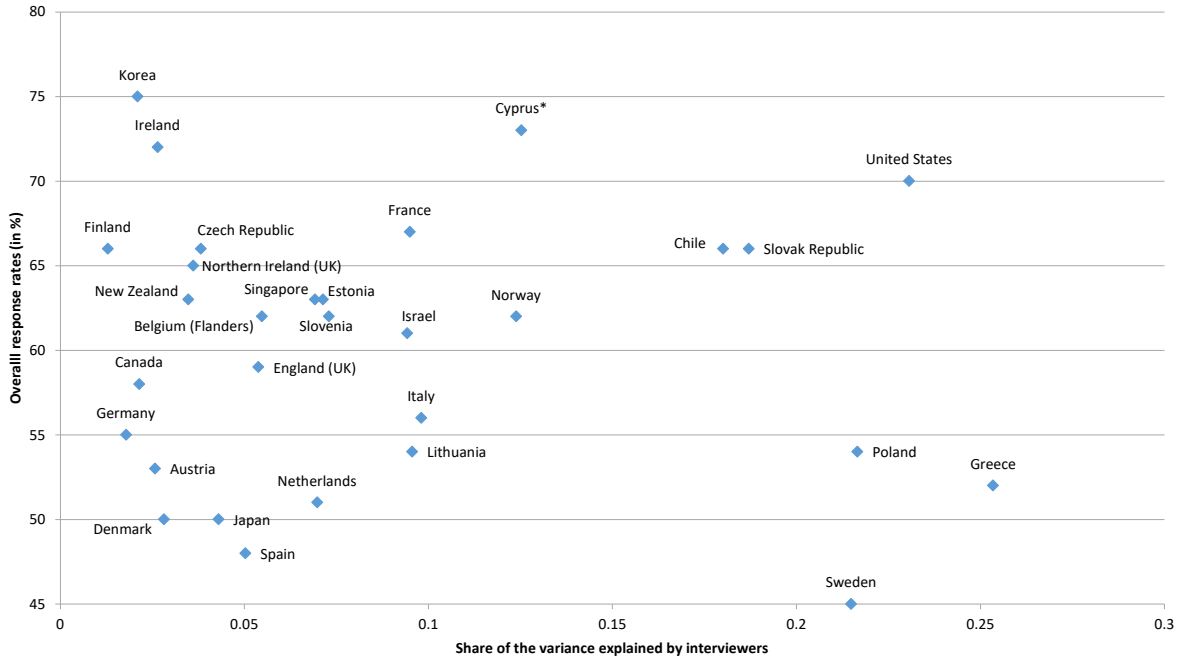
of interest. In this analysis, *Y* will be either the completion of the interview by a potential respondent, or one of the measures of respondent engagement, as previously described. The *X*s are sets of observable characteristics for each type of unit. Finally, the εs are unobservable effects. This equation has to be restrained for two reasons. First, no observable characteristics of interviewers are available, and there are no terms in X_j. Second, sampling units and interviewer assignment overlap too much to not cause irremediable multicollinearity issues. As a consequence, we need to get rid of the ε_s term. The X_i and $X_s$will be the only channels through which we can account for the association between interviewers and sampling units. The resulting equation is thus:

$$\mathrm{Y_{ijs}} = \alpha_c \mathrm{X_i} + \gamma_c \mathrm{X_s} + \varepsilon_i + \varepsilon_j$$

The crucial parameters here are the ε_j, which are called interviewer effects. In this work, we are much more interested in the way these ε_j are distributed than in the attribution to each interviewer of an effect. Usually, such a case calls for estimating random effect models. However, these models assume the independence of the $\varepsilon_j s$ with all other covariates. They are valid when each individual assignment is randomly allocated to interviewers. In most cases, interviewers are actually allocated to geographical sampling units close to where they live. It cannot, therefore, be assumed that the $\varepsilon_j s$ are independent of X_i and X_s. We estimate thus this model using a bootstrap method similar to the method used by Braga, Paccagnella and Pellizzari (2016) and proceed according to the following steps. First, we build two equal samples, with each interviewer's assignment equally spread between the two. Unfortunately, this operation is possible only for interviewers with enough cases assigned (the threshold is 20). Second, we estimate separately in these two samples the previous equations using interviewer fixed effects. Each interviewer effect is estimated without bias, but with error. As a result, the standard deviation of the fixed effect distribution is biased upwards since it also accounts for the standard deviation of these estimation errors. However, for a given interviewer, each sample gives an estimation of its effect. Fittingly, the estimation errors from the two samples are independent of each other. Finally, by exploiting this property, we can obtain an unbiased estimation of the standard deviation of interviewer effects by computing the covariance between the two sets of interviewer effects. As is standard practice, we report the proportion of the total variance explained by interviewers (intra-correlation coefficient), instead of the standard deviation of interviewer effects.

Importantly, this model cannot give an average interviewer effect because it cannot distinguish between an average interviewer effect common to all interviewers and an average individual effect common to all individuals. In this framework, national averages are the indistinguishable sum of a national average of individual effect, which accounts for cultural attitudes and survey promotion, and a national average of interviewer effect. Thus, we can only compare countries according to the variation of interviewer performance, and not according to their average performances. Viewing the share of variance explained by interviewers side by side with national averages is, nonetheless, informative. Generally speaking, the interpretation of interviewer effects will be asymmetric. If their variance is low, we cannot say whether they are uniformly bad or good. That said, if this variance is high, the pool of interviewers is heterogeneous and necessarily contains poorly performing interviewers.

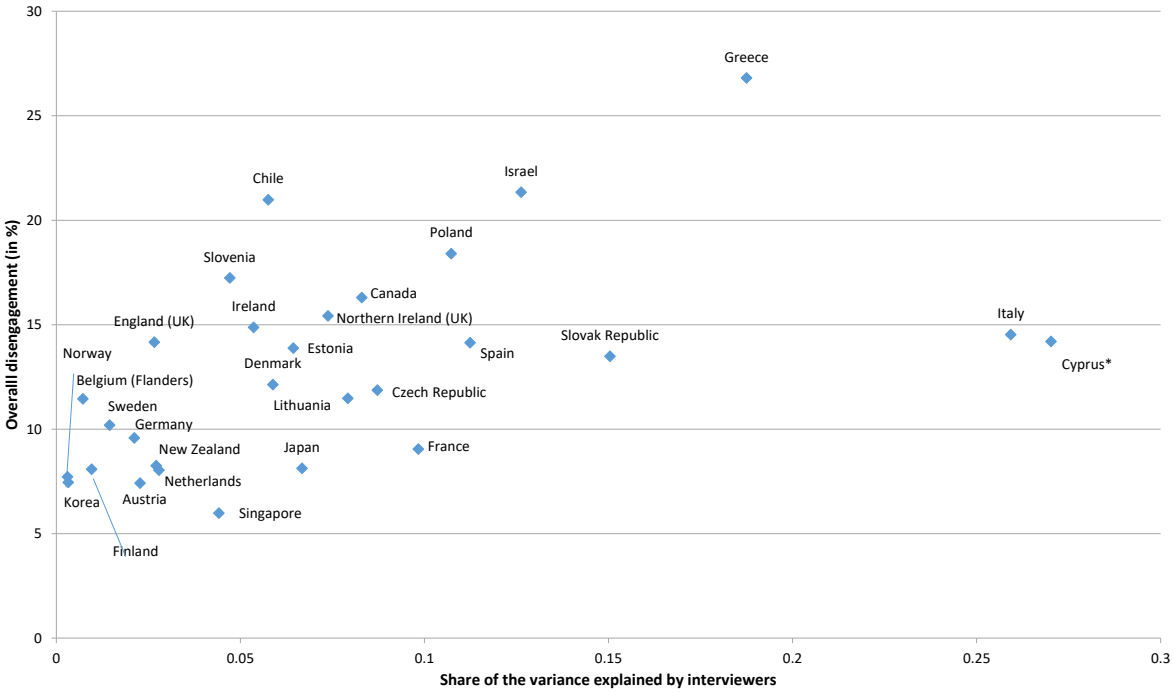
Figure 4.3. Interviewers and response rates



Note: * See note 1 under Figure 2.2.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

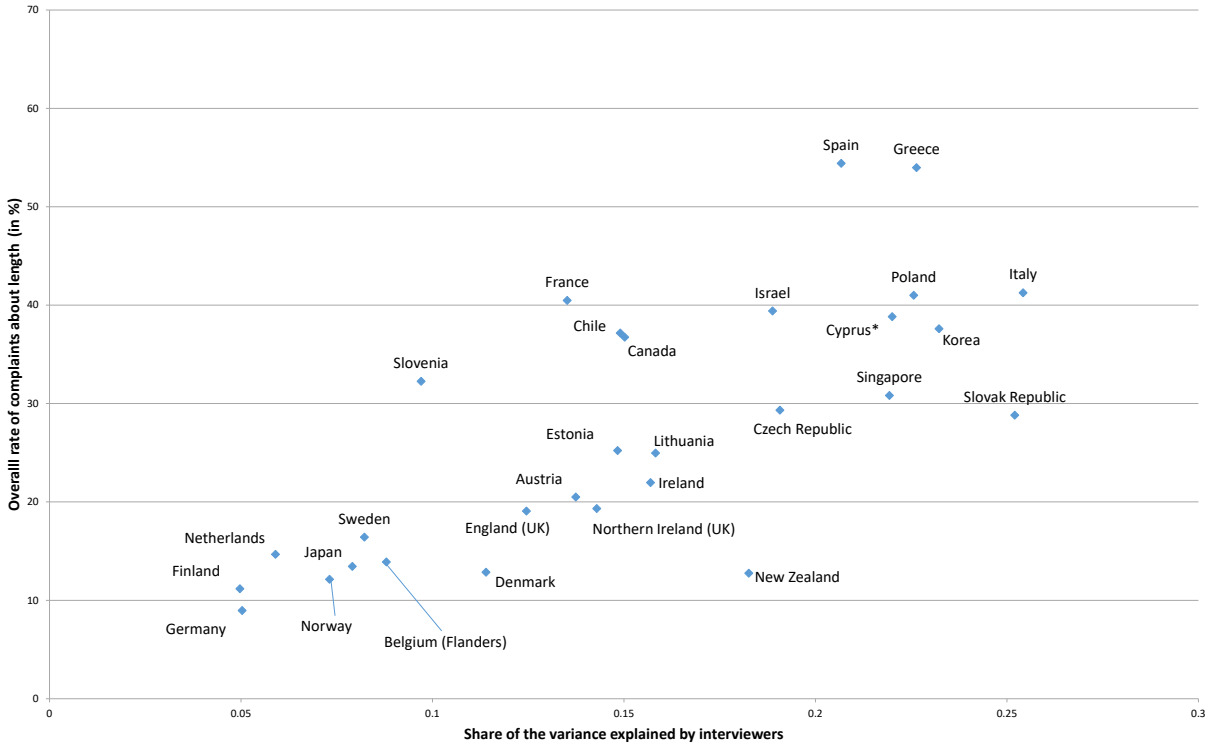
Figure 4.3 shows overall response rates by country, along with the share of the variance of response rates that is explained by interviewers. Interviewer effects were computed using the set of potential respondents. Few controls were available, and as a result, $X_i$and X_S are restricted to age and gender. There are wide variations in the magnitude of interviewer effects. In most countries, interviewer effects are low and explain less than 10% of the variance of non-response rates. In particular, in Korea, Ireland, Finland, Canada and Austria, this proportion is close to zero. On the contrary, in Greece, Sweden, Poland and the United States, a relatively large share (over 20%) of response decisions are explained by interviewer effects. In particular, Poland, Sweden and Greece combine this high share of interviewer effects with low overall response rates. Large variations in interviewer effects indirectly indicate that some interviewers perform poorly in turning potential respondents into actual respondents. This could be because of inappropriate recruitment or management, for instance. In these cases, the overall response rates would be higher without this issue. The lack of a clear association between interviewer effects and overall response rates is in line with the findings of Blom, Leeuw and Hox (2011). These researchers found that co-operation, which is decisive in explaining non-response, is mostly related to country effects and hence to general attitudes towards surveys.

Figure 4.4. Interviewers and disengagement



Note: * See note 1 under Figure 2.2.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

Figure 4.5. Interviewers and complaints about interview length



Note: * See note 1 under Figure 2.2.
Source: Survey of Adult Skills (PIAAC) (2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

Figure 4.4 shows the relationship between interviewers and respondent disengagement. Figure 4.5 shows the relationship between interviewers and respondent complaints about interview length. These models are by nature limited to the pool of actual respondents. As a result, a larger number of individual controls are possible. These include education, social background and employment. They give extra confidence to the reliability of estimates.

The primary motivation behind these figures is to assess how interviewers affect the attitude of respondents towards PIAAC.

Interviewer effects on disengagement rates vary enormously (Figure 4.4). Importantly, the results suggest that interviewer behaviour may be able to reduce disengagement, as the effects are very low in a number of countries. In 10 countries, the share of the variance explained by interviewer effects is below 0.03. It is above 0.1 in only seven countries, and exceeds 0.25 only in Italy and Cyprus.[3] Overall, disengagement rates are generally associated with higher interviewer effects. Inadequate interviewing seems to contribute at least partially to the high level of disengagement in some countries. Nonetheless, even in countries where interviewer effects are almost zero, disengagement rates are above 5%. Interviewer effects on complaints about length offer a similar picture and confirm the previous results. They show large variations among countries and a generally positive association between interviewer effects and the overall rate of complaints. However, contrary to interviewer effects on disengagement rates, they are never close to 0, possibly because length complaints are reported by interviewers themselves – and there is no consensus with regards to what constitutes a complaint.

Importantly, these interviewer effects are correlated across measures. In individual countries, interviewer effects tend to be low across all measures (as in Finland, Germany and Korea) or, alternatively, high across all measures (as in Greece, Poland, the Slovak Republic and Israel). This pattern is similar to the pattern found by Beullens and Loosveldt (2016) in their analysis of the ESS. Their research found that Germany and Finland showed low interviewer effects, while interview effects were higher in the Slovak Republic, Greece and Israel. These regularities across measures as well as across surveys tend to confirm the importance of cultural differences and local labour supply factors as major driving forces of interviewer effects levels.

In Singapore, Korea, Japan, the Netherlands, Austria and New Zealand interviewer effects on disengagement are around 0.07. Not surprisingly, these countries also have very low interviewer effects on response rates. In most countries, the share of variance related to the interviewer lies between 0.1 and 0.2. Interviewers' effects are high and above 0.2 in Chile and Israel. In Greece they are even higher, reaching close to 0.3. Figure 4.5, which shows interviewer effects on complaints, gives a similar picture with a similar ranking. The share of the variance explained by interviewers is slightly higher, with Greece again showing the biggest effects. In these two graphs, interviewer effects are positively associated with high national averages of length complaints, suggesting that countries in which interviewer-effect variance is high might also have high average interviewer effects – which could be detrimental. The impossibility of estimating interviewer-effect country averages makes bias assessment a difficult task. Thus, only a rough order of magnitude can be given. Countries can be compared only according to the standard deviation of interviewer effects – not according to the average of interviewer

[3] See note 1 under Figure 2.2.

effects. The way in which the magnitude can be computed was discussed earlier. The spread of interviewer effects can be regarded as an indicator of the diversity of interviewer attitude and performance. In particular, high interviewer effects or, more precisely, high variance of interviewer effects, implies that the pool of interviewers includes both poorly performing and high-performing interviewers. As a result, a margin exists for improving recruitment processes, training and monitoring. A way of assessing bias in such a context would be to ask the question: "How would our quality outcome change if the average of interviewer effects increased by, say, one standard deviation?" The country average of standard deviations for interviewer effects are: 8.5 percentage points for response rates, 7.5 percentage points for disengagement rates and 15 percentage points for length complaints. Even though the variation of overall response rates among countries is higher, 8.5 percentage points would still be a sizeable improvement for response rates. Regarding the link between interviewers, engagement and attention, and performance, a one standard deviation increase in average interviewer effects would affect literacy and numeracy scores in a country by one point for disengagement rates and three points for length complaints (once converted into literacy score points following the analysis presented in the data section).

5. Conclusion

This paper has explored two sets of factors that have the potential to influence the quality of data from PIAAC and the validity of country comparisons. These factors are variation in testing conditions, and the quality of interviewers.

Because the PIAAC assessment is conducted in respondents' homes or some other agreed-upon location, PIAAC is unable to standardise the conditions under which the assessment is taken to the same extent as other school-based studies (such as the Programme for International Student Assessment, the Trends in International Mathematics and Science Study or the Progress in International Reading Literacy Study). The analysis conducted in this paper suggests that while there are variations in testing conditions between countries (both in terms of the type of space in which the test is taken and in terms of the frequency of certain distractions), this variation seems unlikely to threaten the quality of the measurement in PIAAC. In addition, the distractions that potentially contribute to measurement errors are too rare to be considered of great concern.

At the same time, the information that is collected on testing conditions could be improved. For example, PIAAC could easily enhance the collection of information on distractions, for instance by recording their length or intensity.

Regarding the impact of variation in interviewer performance, some recommendations can be made in the context of preparing for the second cycle of PIAAC. First, a lack of data prevents us from completing a detailed analysis of interviewer effects. This lack of data relates to information about assignments and interviews, as well as information about interviewers. In several countries, the paradata documenting the completion of an interview is not appropriately recorded. Identifying whether the interview was not

completed because the respondent refused to participate, because he or she was not available, or because he or she could not be reached is impossible. Other variables are missing entirely. These include the number of times the interviewer contacted the potential respondent in order to obtain an interview, the start time of the interview and the re-allocation of a potential respondent to another interviewer. Very often, the data have been collected at the country level but have not been centralised into an international database. In addition, no personal information is available about the interviewer, which makes it impossible to examine the relation of interviewer effects and interviewer characteristics such as age, education, gender and experience. This paradata is easy to collect – as long as this collection does not infringe upon any privacy rule. Thus, PIAAC should consider collecting personal information of interviewers for its second cycle.

Accordingly, we suggest that a thorough review of the data collected about assignments, interviews and interviewers be conducted prior to the main study data collection of the second cycle of PIAAC. Additionally, the information required for analysis of interviewer effects should be collected and made available for analysis. Second, the computation of interviewer effects is plagued by the overlapping of sampling areas and interviewers, and it is generally difficult to disentangle the two. As a result, an unknown proportion of the interviewer variance actually captures differences among sampling areas – instead of differences among interviewers. Partially interpenetrated sample assignments, in which sampling areas and groups of interviewers are pooled at a regional level in order to implement a random allocation of cases, would help alleviate this bias (Hibben, Pennell and Scott, 2018). Third, design effects are crucially linked to the number of the number of interviews interviewers take on. Some countries rely heavily on very productive interviewers, who survey more than 100 respondents. The need to ensure that a sufficient number of interviewers are recruited, and the fact that countries should avoid giving large assignments to any single interviewer should be emphasised in the PIAAC Technical Standards and Guidelines for the second cycle. As a final note, it is worth remembering that interviewer effects vary enormously between countries and that the main cause of this variation is ultimately differences in capacity. Some countries such as France, Germany and Finland have a large pool of experienced interviewers whose work can be supervised more easily. This in turn minimises interviewer effects. In some other countries, interviewer capacity has to be built almost from scratch. It is important that the team in charge of the survey at the international level pay close attention to data collection in such countries and provide help and expertise that might be lacking at the national level. The findings of researchers looking at quality assurance in multinational studies show that participation in an international survey can also help some countries improve their survey infrastructures – as long as adequate monitoring is implemented (Pennel et al., 2017).

References

Ackermann-Piek, D. and N. Massing (2014), "Interviewer behavior and interviewer characteristics in PIAAC Germany", *Methods, Data, Analyses*, Vol. 8/2, pp. 199-222, https://mda.gesis.org/index.php/mda/article/view/2014.008.

Beullens K. and G. Loosveldt (2016), "Interviewer effects in the European Social Survey", *Survey Research Methods*, Vol. 10/2, pp. 103-118.

Blom, A., E. de Leeuw and J. Hox (2011), "Interviewer effects on nonresponse in the European Social Survey", *Journal of Official Statistics*, Vol. 27/2, pp. 359-377.

Braga, M., M. Paccagnella and M. Pellizzari (2016), "The impact of college teaching on students' academic and labor market outcomes", *Journal of Labor Economics,* Vol. 34/3, pp. 781-822.

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.

Hibben, C., K. Pennell and L. Scott (2018), "Interviewer effects in multicultural, multinational and multiregional surveys". *Quality Assurance in Education,* Vol. 26/2.

Loosveldt G. (2008), "Face-to-face interviews", in E. de Leeuw, J. Hox and D. Dillman (eds.), *International Handbook of Survey Methodology*, Taylor & Francis Group, New York.

Maddox, B. , P. Javrh and F. Keslair (forthcoming), "Investigating testing situations", in B. Maddox (ed.), *International Large-Scale Assessments in Education: Insider Research Perspectives*, Bloomsbury, London.

OECD (2016), *Skills Matter: Further Results from the Survey of Adult Skills*, OECD Skills Studies, OECD Publishing, Paris, https://doi.org/10.1787/9789264258051-en.

OECD (2015), Survey of Adult Skills (PIAAC) (Database 2012, 2015), www.oecd.org/skills/piaac/publicdataandanalysis.

OECD (2013), *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*, OECD Publishing, Paris, https://doi.org/10.1787/9789264204256-en.

Pennell, B.-E., et al. (2017), "A total survey error perspective on surveys in multinational, multiregional, and multicultural contexts", in P. P. Biemer et al. (eds.), *Total Survey Error in Practice*, pp. 179-202, John Wiley & Sons, Hoboken, New Jersey.

Smith, T. (2011), "Refining the total survey error perspective" *International Journal of Public Opinion Research*, Vol. 23/4, pp. 464-484, https://doi.org/10.1093/ijpor/edq052.

Tourangeau, R. and N. Bradburn (2010), "The psychology of survey response", in P. Marsden and J. Wright (eds.), *The Handbook of Survey Research*, second edition, Emerald Group Publishing Limited, Bingley.

West, B. and A. Blom (2016), "Explaining interviewer effects: A research synthesis", *Journal of Survey Statistics and Methodology*, Vol. 5/2, pp. 175-211, https://doi.org/10.1093/jssam/smw024.