SFI -The Danish National Centre for Social Research
4. December 2014

PISA 2000, 2003, 2006, 2009, and 2012 data; PIAAC 2011-12 data and combined PISA 2000 and PIAAC 2011-12 data.

1. Introduction

This paper is made for researchers, including students and other analysts, interested in using data from PISA, PIAAC, and PISA-PIAAC in Denmark.

PISA means The Programme for International Student Assessment. PIAAC is The Programme for the International Assessment of Adult Competencies (a kind of adult PISA). Both PISA and PIAAC are international OECD surveys (http://www.oecd.org). PISA covers pupils 15-16 years old and it is conducted every three years, for the first time in 2000. PIAAC covers 16-65-year-olds and was conducted in 2011/12. PISA-PIAAC is a special investigation undertaken only in Denmark where participants in PISA 2000 were re-tested and interviewed again in PIAAC 2011-12.

Danish data from PISA and PIAAC have been collected by SFI (the former SFI-Survey). The data are owned by The Ministry of Education, Denmark. The Ministry has decided that Danish PISA and PIAAC data should be placed in open access by SFI – The Danish National Centre for Social Research within the framework of Centre for Survey and Survey/Register data (CSSR), cf. www.sfi.dk and http://www.sfi.dk/cssr-7745.aspx. "Open access" means that anonymous data are made accessible via the internet, so that researchers may download data to their own computer and make analyses in this way. Downloading of data is free of charge for the individual researcher.

Using the Danish person identification number (CPR-number) it is also possible to combine PISA or PIAAC survey data with register data. Such combined data cannot be placed in open access. Researchers interested in analyzing combined register and survey data should send an application to the Ministry of Education, cf. Box 1 next page. Acceptance from the Ministry is a precondition for use of PISA or PIAAC data in this way. If an application is accepted, the Ministry will inform the

applicant about the procedure and conditions for using PISA and/ or PIAAC data together with register data.

Concerning technical issues related to "open access" survey data, contents of data, access to data not or only partly placed in open access (e.g. sensitive data), you may contact the persons in Box 2.

Box 1: Ministry of Education responsible for PISA and PIAAC in Denmark

Claus Jepsen, Special Advisor
The Danish Ministry of Education
Test and Evaluation Division
National Agency for Quality and Supervision
Frederiksholms Kanal 25, 1220 Copenhagen K, Denmark
Direct: +45 3392 5431. Email: claus.jepsen@ktst.dk
www.uvm.dk

Box 2: Contact persons at SFI in relation to use of PISA and PIAAC data

Anders Rosdahl, Senior Research Fellow
SFI – The Danish National Centre for Social Research
Herluf Trolles Gade 11, 1052 Copenhagen K, Denmark
Direct: +45 33 48 09 20. Email: ar@sfi.dk
www.sfi.dk

Leif Jensen, Head of Centre
CSSR – Centre for Survey and Survey/Register data
SFI – The Danish National Centre for Social Research
Herluf Trolles Gade 11, 1052 Copenhagen K, Denmark
Direct: +45 33 48 08 20. Email: leh@sfi.dk.
www.sfi.dk

The remaining parts of this paper include the following. *Section 2* presents links to data, documentation, questionnaires, and reports with results. *Section 3* includes links to special programs which have been developed to analyze PISA and PIAAC data. *Section 4* presents the PIAAC survey in Denmark and issues related to analysis of PIAAC data. *Section 5* describes the special PISA-PIAAC survey in Denmark and some technical issues related to these unique longitudinal data.

2. Links to data, documentation, questionnaires, and results

Data from PISA and PIAAC have been placed in open access within the framework of CSSR, http://www.sfi.dk/cssr-7745.aspx. You may access data via this link or the following more direct link:

http://www.sfi.dk/surveys-alphabetically-10676.aspx

Under this link you will see a number of surveys including:

- PISA (2000, 2003, 2006, 2009, 2012)
- PISA-PIAAC
- PIAAC

You click on "*open access*" to the right of the survey. Then you see a list of surveys including PISA, PISA-PIAAC and PIAAC. By clicking on PISA you get a menu with PISA 2000, PISA 2003, PISA 2006, PISA 2009 and PISA2012. When you click on one of these PISA surveys, PISA-PIAAC or PIAAC you get a menu with two entries: "*Metadata*" and "*Variable description*". Under "*Metadata*" you find two entries: "*Study description*" and "*Other documentation*". The latter entry includes:

- Questionnaires (in English and in Danish)
- Result reports (main reports with results in Danish and English)
- Documentation (codebooks and technical reports from OECD)

The PIAAC and PISA surveys are placed in open access from mid-December 2014. The PISA-PIAAC survey will be placed in open access in 2015.

3. Links to programs

PISA and PIAAC surveys are complex: the samples are not simple random due to explicit design as in PISA, and/or non-response rate, which is high and biased in PIAAC. In PISA and PIAAC databases the complex survey design is indicated with population weights and replicate weights. Both types of surveys measure skills in e.g. literacy/reading. The skills are estimated with plausible values based on multiple imputation technique. For each respondent with given characteristics, a skills-distribution is estimated. From this distribution is randomly drawn a number of so-called plausible values (e.g. 10 in PIAAC), which are the information (variables) about the respondents' skills in the data from PISA and PIAAC. Present standard versions of SPSS, SAS, and STATA cannot, without further programming, calculate unbiased estimates of mean and variance in analyses including plausible values in the context of complex surveys.

Therefore certain special kinds of programs have been developed to handle PISA and PIAAC types of data in the correct way; that is, in a way that ensures correct estimation of variance in particular.

Perhaps the simplest program from a user point of view is The International Database (IDB) Analyzer, which can handle both PISA and PIAAC data. The program is available from International Association for the Evaluation of Educational Achievement (IEA) website. It can be downloaded for free here:

http://www.iea.nl/data.html

A guide to installation can be found here:

http://www.iea.nl/fileadmin/user_upload/IEA_Software/Installing_the_IDB_Analyzer__Version_3_0_.pdf

With installation follows a user manual, which can be printed.

The IDB Analyzer creates SPSS code that can be used with SPSS to conduct statistical analyses of PISA or PIAAC data, taking into account the complex sample structure of the data and the plausible values method. It provides several different procedures for analysis, such as the computation of means or percentages of any background variable of interest for a whole country or subgroup within

a population. Analysis of regression and certain other types of analysis specially aimed at handling plausible values may also be undertaken with the program.

Use of IDB Analyzer presupposes that the user has installed SPSS on his/her computer, but it is not necessary to have much knowledge/skill related to SPSS to be able to use the program, which is easy to learn, tested and maintained through many years.
You may also use SPSS, SAS, and STATA more directly, cf. the following.

OECD has published comprehensive PISA guides for users of SPSS and SAS, respectively, cf. the following two links, which also refer to special programs (macros) developed for PISA data:

PISA Analysis Manual SPSS, second edition, 2009
http://browse.oecdbookshop.org/oecd/pdfs/free/9809031e.pdf

PISA Analysis Manual SAS, second edition, 2009
http://browse.oecdbookshop.org/oecd/pdfs/free/9809021e.pdf

A new module from 2014 using STATA to handle both PISA and PIAAC data is called REPEST, cf. the link:
http://econpapers.repec.org/software/bocbocode/S457918.htm

The module supports analyses with plausible values and replicate weights according to OECD data analysis guidelines, complementing the existing PISA-macros for SAS and SPSS, cf. above. Any questions or feedback on this module should be directed to francesco.avvisati@oecd.org.

Another STATA based module to handle both PISA and PIAAC data has been developed by Kevin Macdonald, cf. the link:
http://econpapers.repec.org/software/bocbocode/s456951.htm

The following SAS and STATA based programs are designed to handle PIAAC data but not PISA data:

PIAAC SAS Macro User Guide, version 2.0
http://www.oecd.org/site/piaac/PIAAC%20SAS%20Macro%20User%20Guide%20-%202.0.pdf

PIAACTOOLS: Stata® programs for statistical computing using PIAAC data.
http://www.oecd.org/site/piaac/PIACTOOLS_16OCT_for_web.pdf

The following STATA based program is designed to handle PISA but not PIAAC data:

Program developed by Maciej Jakubowski and Artur Pokropek:
https://ideas.repec.org/c/boc/bocode/s457754.html

The links also describe how to access the programs, which may be used for free.

4. PIAAC

PIAAC (The Programme for the International Assessment of Adult Competencies), or Survey of Adult Skills, is an OECD study of skills in literacy, numeracy, and problem solving in technology-rich environments among the population aged 16-65 years. The first round comprised 24 countries, including Denmark where PIAAC was initiated and financed by The Ministry of Education, The Ministry of Employment, The Ministry of Science, Innovation and Higher Education, The Ministry of Finance, and The Ministry of Social Affairs, Children and Integration. The data collection was carried out in 2011-2012. The international results were published by OECD in October 2013 and can be found at: www.oecd.org/site/piaac/surveyofadultskills.htm. The main report with results from Denmark was published at the same time:

- Anders Rosdahl, Torben Fridberg, Vibeke Jakobsen, Michael Jørgensen: *Færdigheder i læsning, regning og problemløsning med IT i Danmark*. København: SFI - Det Nationale Forskningscenter for Velfærd. 13:28. 2013.

The report can be downloaded from www.sfi.dk. An English summary of the Danish PIAAC results is downloadable from: http://www.sfi.dk/danskernes_kompetencer-7149.aspx.

The Danish PIAAC sample: The Danish PIAAC sample was drawn from the population aged 16-65 years in July 2011. This segment of the population comprised 3,629,000 persons. The sample was drawn by random. Persons aged 55-65 years and immigrants were oversampled in order to be able to focus in particular on these categories in the population. The total number of respondents was 7,328. The response rate was around 50 per cent.

PIAAC in Denmark also included interviews and tests of 1,881 persons who participated in PISA in year 2000. These respondents are not part of the PIAAC Main Survey. These data are described in section 5 below.

The PIAAC data collection was undertaken in the respondents' homes. An interviewer with a PC first conducted an interview (about 40 minutes) with the respondent, who was subsequently asked to solve the assessment tasks on the interviewer's PC or - if this was not possible - with paper and pencil. The assessment took about 60 minutes, but the respondent was allowed to use as much time as necessary.

In addition to the measured skills, PIAAC includes a wealth of information from the interviews with respondents including detailed data on: education and training, adult education, labour market experience, employment (industry, sector, occupation, contents of the job), wage/ income from employment, unemployment, activities related to cognitive foundation skills (reading, writing, calculating, and ICT use both at work and outside work), leaning motivation and learning strategies, household and family, parents highest level of education, and immigration status.

The skills measured in PIAAC are Cognitive Foundation Skills (CFS) or "key information-processing skills":

- *Literacy:* The ability to understand, evaluate, use, and engage with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential.
- *Numeracy:* The ability to access, use, interpret, and communicate mathematical information and ideas in order to engage in and manage the mathematical demands of a range of situations in adult life.

- *Problem solving in technology-rich environments:* The ability to use digital technology, communication tools and networks to acquire and evaluate information, communicate with others, and perform practical tasks (ICT skills - that is, skills in using information and communications technology)

Proficiency in these domains is measured on a scale from 0 to 500. OECD has divided the literacy and numeracy scales into six levels (below 1, 1, 2, 3, 4, and 5). The ICT skills are divided into five levels (no score, below 1, 1, 2, and 3). The "no score" category includes persons with no computer experience and persons who could not or did not want to do the assessment on the interviewer's computer.

Analysis of PIAAC data: Danish PIAAC Main Survey data with 7,328 respondents include a full sampling weight (variable name: SPFWT0), the inverse of the sampling probability, taking both sampling design and non-response into consideration. The sum of SPFWT0 amounts to the total Danish population aged 16-65 July 2011 from which the sample was drawn, cf. above. SPFWT0 is used for calculating unbiased estimates. The data also includes 80 replicate weights (SPFWT1-80) for each person. The replicate weights in the Danish PIAAC are calculated by a statistical method called jackknife1. Each replicate weight also sums up to the total population. The replicate weights are used for making correct variance estimates, taking the sampling design and non-response into consideration. Both types of weights have been calculated by Statistics Denmark based on e.g. non-response bias analysis including a large number of register variables.

Special standard programs exist that take the nature of the sample into consideration, i.e. population weights and replicate weights. SFI has used the svy commands in STATA, cf. the stata manual on survey data/estimation. You open the PIAAC stata dataset and tell STATA about the weights etc. by issuing the "svyset" command. Hereafter, the svy estimation commands are run - see the STATA manual on survey data.

Standard STATA (and standard SAS in an analogous way) can always be used in PIAAC when the analysis does not include plausible values.

In PIAAC, Cognitive Foundation Skills (CFS) are, as mentioned, measured in three domains: Literacy, Numeracy, and Problem Solving with ICT. Each PIAAC respondent solves a *sample* of

tests in an adaptive testing design. The persons typically get diverse combinations of tests, although with some overlap of tests between categories of persons. For each PIAAC respondent with given traits is estimated a *distribution* of CFS within each domain - based on Item Response Theory and multiple imputation methods. From this distribution is randomly drawn 10 so-called *plausible values* (with a value range 0-500). Each PIAAC respondent thus has 30 plausible values (3 times 10) which have been calculated by The International PIAAC Consortium. Based on the plausible values, proficiency estimates (skills in a CFS domain) can be calculated for (larger) groups but, in principle, not for individuals. This design has been chosen to make the testing time short (about 1 hour) for economic and practical reasons - analogous with the reasons behind interviewing a sample of the population rather than the whole population.

The average of the 10 plausible values for a *group* of persons represents an unbiased estimate of the proficiency for the group. This estimate is necessarily, due to the test-design, inflicted with a certain *measurement error,* which has to be estimated. Neither STATA nor other standard programming packages can handle measurements error AND errors due to sampling design and non-response (cf. above) simultaneously, in the same analysis, without additional coding/programming. STATA's multiple imputation (mi) commands can, properly used, handle measurement errors. STATA's svy commands can handle errors due to sampling design and non-response, cf. above. But mi and svy cannot be combined in the same analysis in the present standard version of STATA; maybe this will be possible in some future version.

Therefore, to analyze PIAAC data with plausible values, special programs are needed, cf. the reference to SPSS based, SAS-based, and STATA-based special programs in *section 3* of this paper.

5. PISA-PIAAC

The last SFI report from PIAAC in Denmark was published in June 2014:

Anders Rosdahl: Fra 15 til 27 år. PISA 2000-eleverne i 2011/12. København: SFI – Det Nationale Forskningscenter for Velfærd. Rapport 14.13. (148 pp).

The report can be downloaded from:
http://www.sfi.dk/rapportoplysninger-4681.aspx?Action=1&NewsId=4348&PID=9267

The report focuses on a special subsample in the Danish PIAAC survey: 1,881 individuals who had participated in PISA 2000 as 15/16-year old students as well as in PIAAC 2011/12 as young adults. The report investigates the correlation between reading skills at the end of compulsory schooling assessed in PISA 2000 and the young people's educational and labour market situation at the age of 27. It analyzes the factors that have had a positive or negative impact on the development of their reading skills from 2000 to 2011/12.

This special PISA-PIAAC part of PIAAC in Denmark was undertaken at the initiative of the funding ministries in early 2011[1], as PIAAC provided a unique opportunity to follow up on PISA 2000 students and their education, work, and reading skills in 2011/2012.

As mentioned in section 2 above, the combined PISA-PIAAC data have been placed in open access. In the following we describe some main features of the data and analysis.

The persons in the data: The point of departure for the data set is the 4,235 persons who participated in PISA 2000. At the outset it was intended to re-interview and re-test all these persons in PIAAC. However, the following persons could not be contacted by SFI-Survey collecting PIAAC data:

- 308 persons with unknown Person Identification Number (CPR), meaning that the persons' current addresses could not be found.
- 1,074 persons with so-called "researcher protection". Due to legislation SFI-Survey was not allowed to contact these persons.
- 118 persons who had died or emigrated or were institutionalized in 2011/12.

Of the remaining 2,735 persons, PIAAC interviews were obtained with 1,881 persons. For these persons both PISA and PIAAC data are available. The 1,881 persons are a sub-set of the total number of participants in PISA 2000 (4,235 persons), who are all included in the data.

[1] As mentioned the funding ministries include the Danish Ministry of Education, the Danish Ministry of Employment, the Danish Ministry of Higher Education and Science, the Danish Ministry of Business and Growth, as well as the Danish Ministry of Children, Gender Equality, Integration and Social Affairs.

A special variable in the data (called *resultcodePISA*) includes information on the outcome of the PIAAC data collection among PISA 2000 participants. The codes of this variable are:

1 Respondents (are both included in PISA 2000 and in PIAAC, 1,881 persons)
2 Researcher protection (were not allowed to be interviewed in PIAAC, 1,074 persons)
3 Dead, emigrated, institutionalized in 2011/12(could not be interviewed, 118 persons)
4 Unknown address, interviewer could not obtain contact (278 persons)
5 Did not want to be interviewed (526 persons)
6 Language problem (2 persons)
7 Disability, reading/writing difficulty (48 persons)
8 No information on person identification number (CPR) (308 persons)

Information (variables) in the data: The original PISA 2000 data are organized in different data sets. The point of departure for the PISA-PIAAC data is the so-called reading file (OECD terminology). All 4,235 PISA 2000 students completed reading skills tests, whereas only 2,382 and 2,346 persons were tested in mathematics and science. Therefore, the more comprehensive data concern reading skills. The PISA-PIAAC data focus almost exclusively on these skills among the PISA 2000 students[2].

In addition to the skills tests, PISA 2000 students responded to a comprehensive questionnaire about their social background, experience at school, computer use, leaning strategies and their assessment of their own abilities and performance at school. Questionnaire data were also collected from school managers at the students' schools. This questionnaire was about organization of the school, the pupils, staff, evaluation forms, as well as teaching resources. Data from both of these questionnaires are included in the data about the 4,235 PISA 2000 students.

For the subcategory who also participated in PIAAC (1,881 of the 4,235 respondents) the data also include the same variables as in the PIAAC Main Survey, cf. section 4 in this paper.

[2] Technical note: The PISA 2000 reading file has in our data been supplied with warm estimates in mathematics (wlemath) and science (wlescie). However, the weights in the dataset cannot be used for calculating unbiased estimates of mean and variance for these two measured skills. For technical reasons only persons where information on Danish Person Identification Number (CPR) is available can have a valid code in these two variables. This also holds for the variables ST29Q01 ST29Q02 ST29Q03 – also due to technical matters. Information on wlemath and wlescie are available for 2,055 persons and 2,017 persons, respectively. 3,721 persons have valid answers in both ST29Q01, ST29Q02, and ST29Q03.

Sampling design, weighting, and analysis: PISA 2000 used two-stage sampling. First, schools were selected and then pupils born in 1984 were randomly selected from these schools. PISA 2000 covers a total of 225 schools. The selection procedure was designed such that the population which the 4,235 pupils was to represent included all young people born in 1984 who, at the time of testing in March to April 2000, were living in Denmark and enrolled in education. A small group of young people who had not enrolled at an educational institution were not included in the population from which the PISA 2000 pupils were selected. Pupils at schools for children with special needs were also excluded. At the time of testing, pupils were aged between 15 years and 2 months and 16 years and 4 months. The vast majority of participants in PISA 2000 were pupils at the Danish Folkeskole.

The PISA 2000 data include both population and replicate weights. The population weight for a given person indicates, in popular terms, the number of persons in the population represented by this person. The sum of the population weights constitutes 47,786 persons, which is thus the built-in estimate in PISA 2000 data of the number of persons the 4,235 PISA 2000 respondents represent.

In addition to the population weight, 80 replicate weights per person (Balanced Repeated Replication, BRR) were calculated in the PISA 2000 data. These weights are the basis for estimates in PISA 2000 of the variance due to sampling design.

Data with population and replicate weights can be handled by standard programming packages, for example STATA, which was used by SFI (see also section 4 in this paper). The population and replicate weights in the data should be used in analyses where all 4,235 PISA 2000 students are included. The name of the population weight in the data is: w_fstuwt. The 80 replicate weights are: w_fstr1-80. In the same way as in PIAAC (cf. section 4) you open the STATA PISA-PIAAC data and tell STATA about the weights by issuing the "svyset" command. Then you proceed with the analysis using STATA survey estimation commands (cf. the STATA manual on survey data).

SFIs analysis of the subset of 1,881 persons participating in both PISA 2000 and PIAAC has used corrected population and replicate weights. SFI calculated each of the 81 corrected weights as the original weight times 1.08114 times the variable *bortfaldsvgt* in the data. This simple method of calculation is not 100 percent correct but it is assessed that the bias is minimal.

The variable *bortfaldsvgt* has been calculated by Statistics Denmark. It is a weight that up-weigh each of the 1,881 respondents to the subgroup of the 4,235 PISA 2000 respondents for whom the Person Identification Number (CPR) is known and who lived in Denmark and were not institutionalized in 2011-12, cf. code 3 of the variable *resultcodePISA*, cf. above. The number 1.08114 (cf. above) corrects for the fact that person identification number (CPR) is not known for all PISA 2000 respondents.
The idea behind the calculation of corrected weights is that the 1,881 PISA-PIAAC respondents are intended to represent the (total) population of (all) young people born in 1984 who went to school in Denmark in the spring of 2000 and who (still or again) lived in Denmark in 2011/12 when the PIAAC survey took place. This population is, of course, a little less than the population represented by the 4,235 original PIAAC respondents. The sum of the corrected population weights is 46,881 or about 46,500 persons, comprising about 90 per cent of the 1984 birth cohort in Denmark.

Standard STATA to handle population weights and replicate weights (and standard SAS in an analogous way) can always be used for PISA-PIAAC data when the analysis does not include plausible values.

However, if the analysis includes plausible values, standard programming packages should not be used, cf. the arguments presented in section 4 above.

Therefore, to analyze PISA-PIAAC data with plausible values special programs are needed, cf. the reference to SPSS based, SAS-based, and STATA-based special PISA-programs in *section 3* of this paper.

Observe that the PISA-PIAAC data do include PISA 2000 population and replicate weights but not PIAAC population and replicate weights. Only the PIAAC Main Survey includes PIAAC population- and replicate weights, cf. section 4. Therefore, to analyze PISA-PIAAC data special programs to handle PISA data (not PIAAC data) should be used.

Reading skills in PISA 2000 and literacy in PIAAC 2011-12 are at the theoretical level defined in nearly the same way, but the concrete tests (texts and questions) in PISA and PIAAC are different. We cannot be sure that the reading test in PISA 2000 measures exactly the same construct as the

literacy test in PIAAC 2011-12. Furthermore, the scaling is not the same in PISA as in PIAAC. It is not possible to convert a given PISA reading score to a corresponding PIAAC literacy score in the same way as you convert a temperature in Celsius to Fahrenheit and vice versa. But the PISA-PIAAC data make it possible to analyze changes in the respondents' placement in the skills distribution from 2000 to 2011-12; e.g. movements from the best third in reading in 2000 to the poorest third in 2011-12.