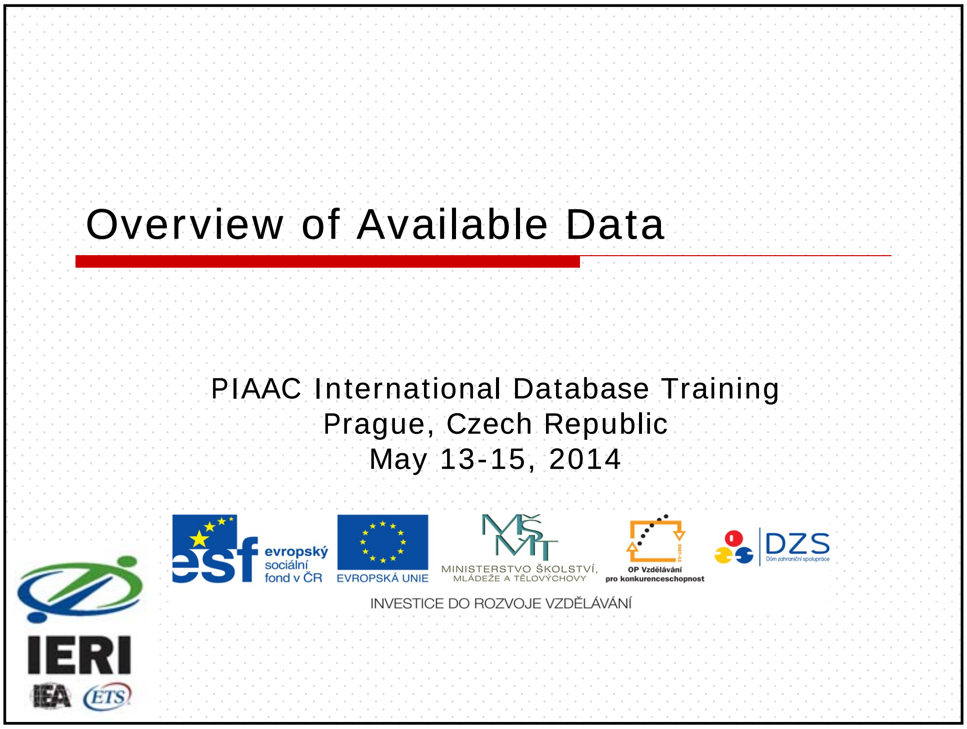


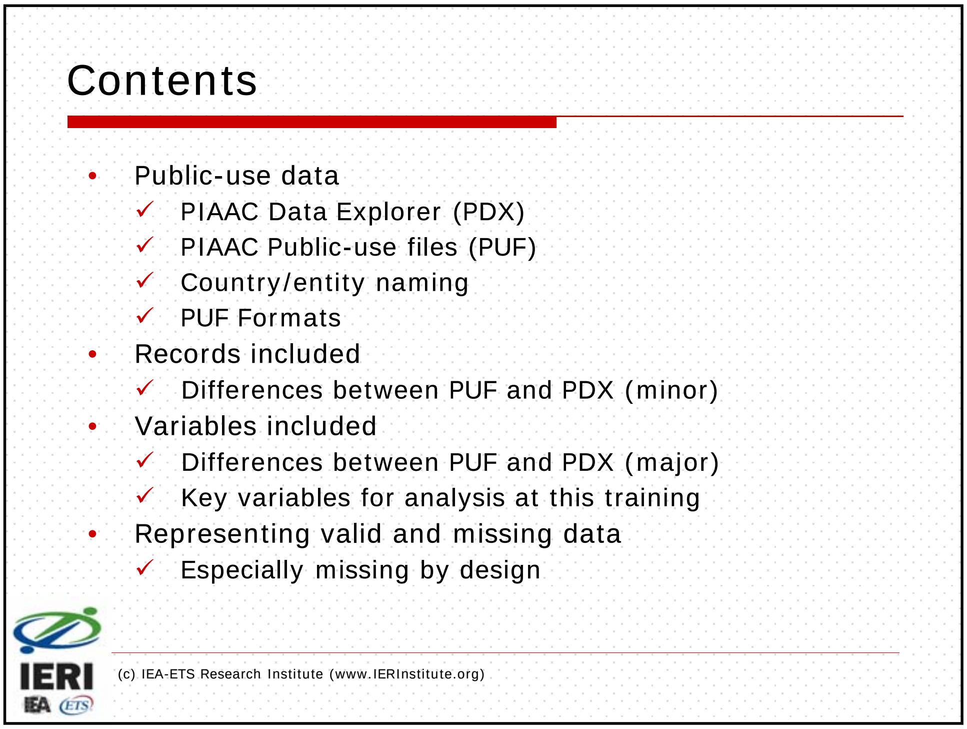
Overview of Available Data

PIAAC International Database Training
Prague, Czech Republic
May 13-15, 2014

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
DZS

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

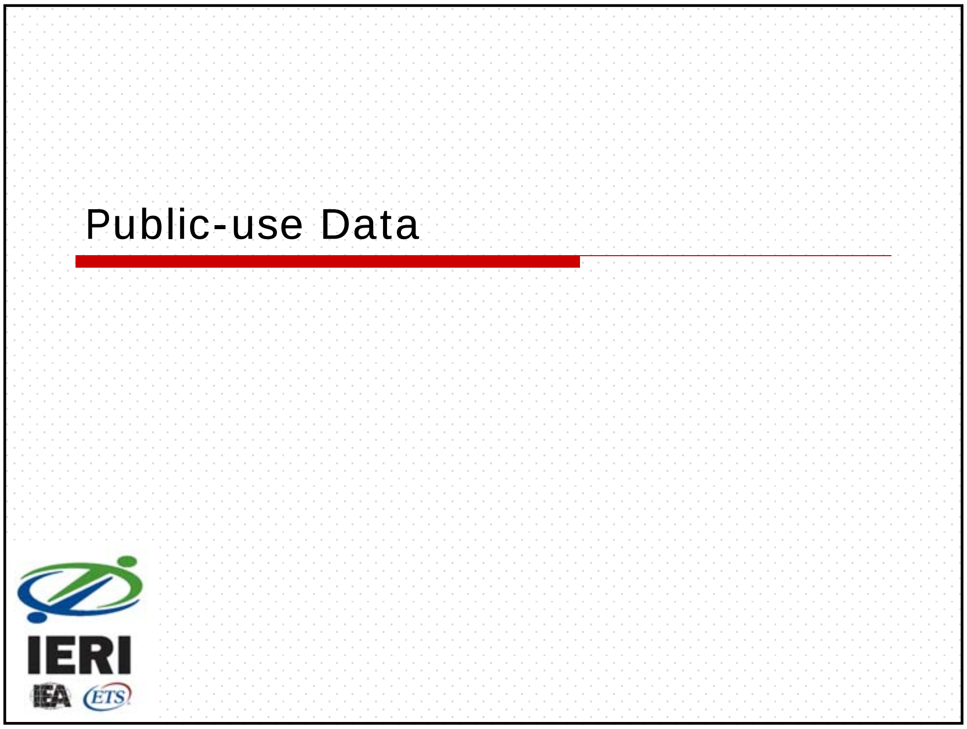
IERI
IEA ETS



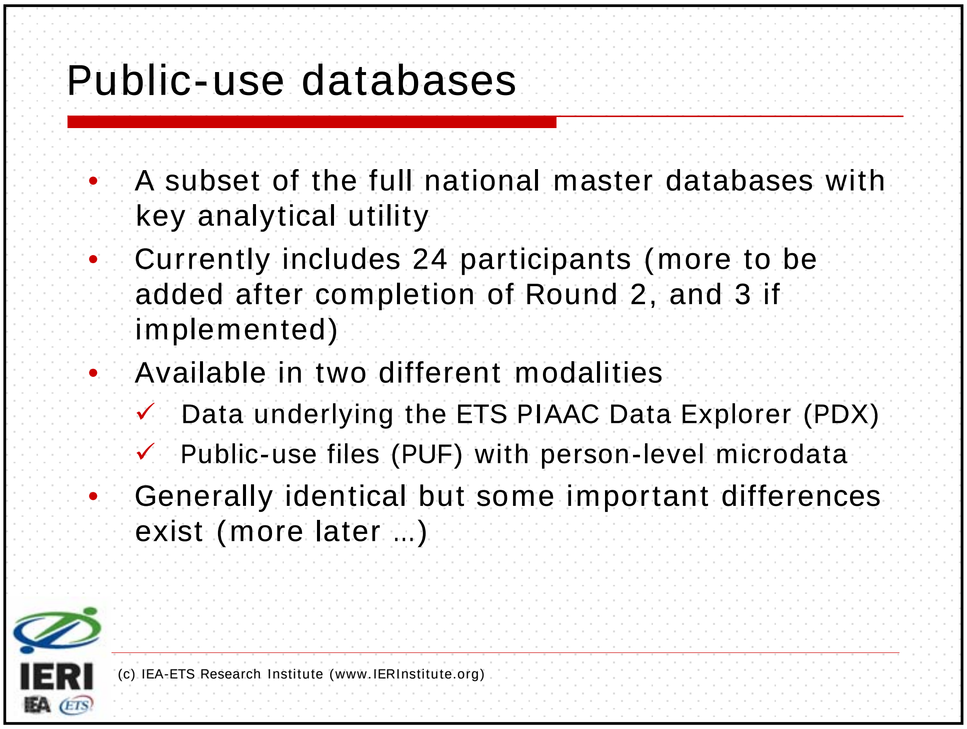
Contents

- Public-use data
  - ✓ PIAAC Data Explorer (PDX)
  - ✓ PIAAC Public-use files (PUF)
  - ✓ Country/entity naming
  - ✓ PUF Formats
- Records included
  - ✓ Differences between PUF and PDX (minor)
- Variables included
  - ✓ Differences between PUF and PDX (major)
  - ✓ Key variables for analysis at this training
- Representing valid and missing data
  - ✓ Especially missing by design

IERI
IEA ETS

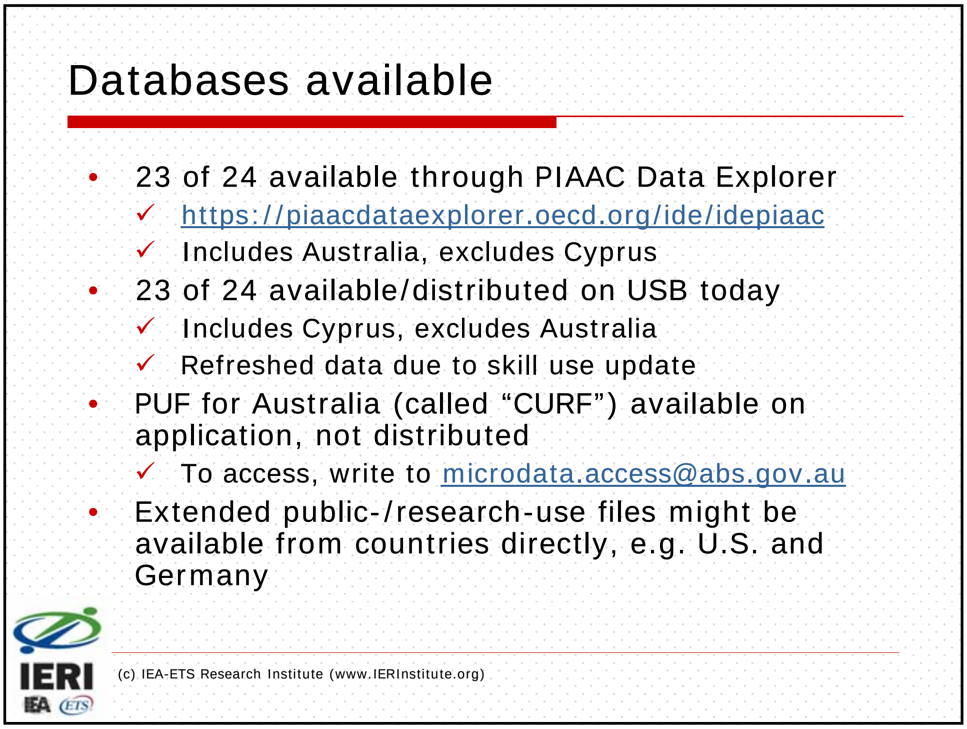


Public-use Data



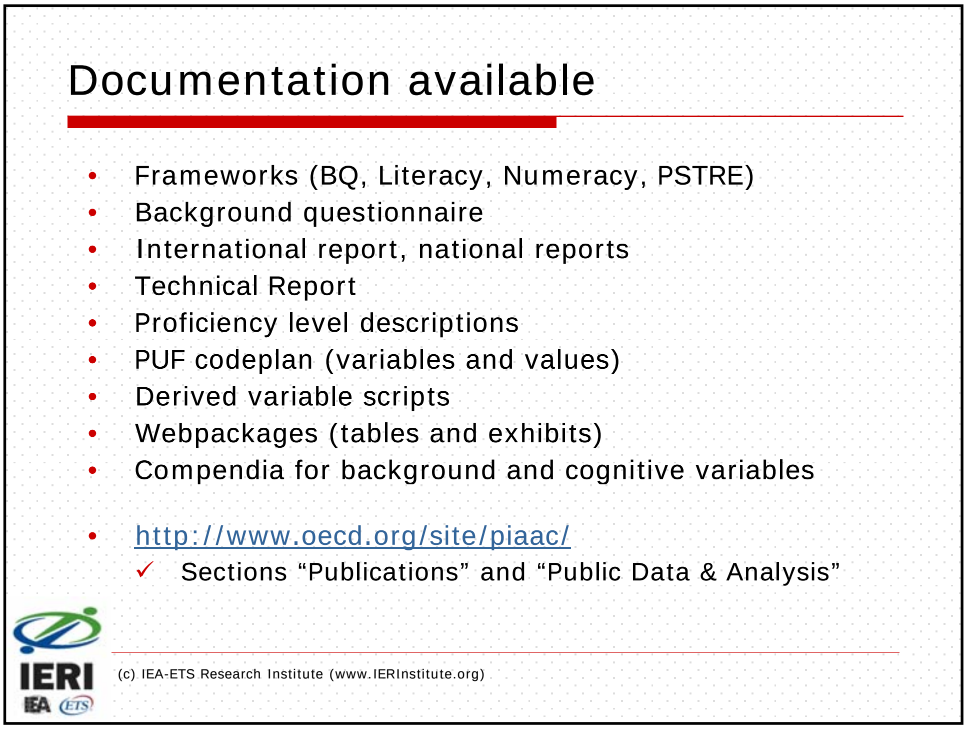
Public-use databases

- A subset of the full national master databases with key analytical utility
- Currently includes 24 participants (more to be added after completion of Round 2, and 3 if implemented)
- Available in two different modalities
  - ✓ Data underlying the ETS PIAAC Data Explorer (PDX)
  - ✓ Public-use files (PUF) with person-level microdata
- Generally identical but some important differences exist (more later …)



Databases available

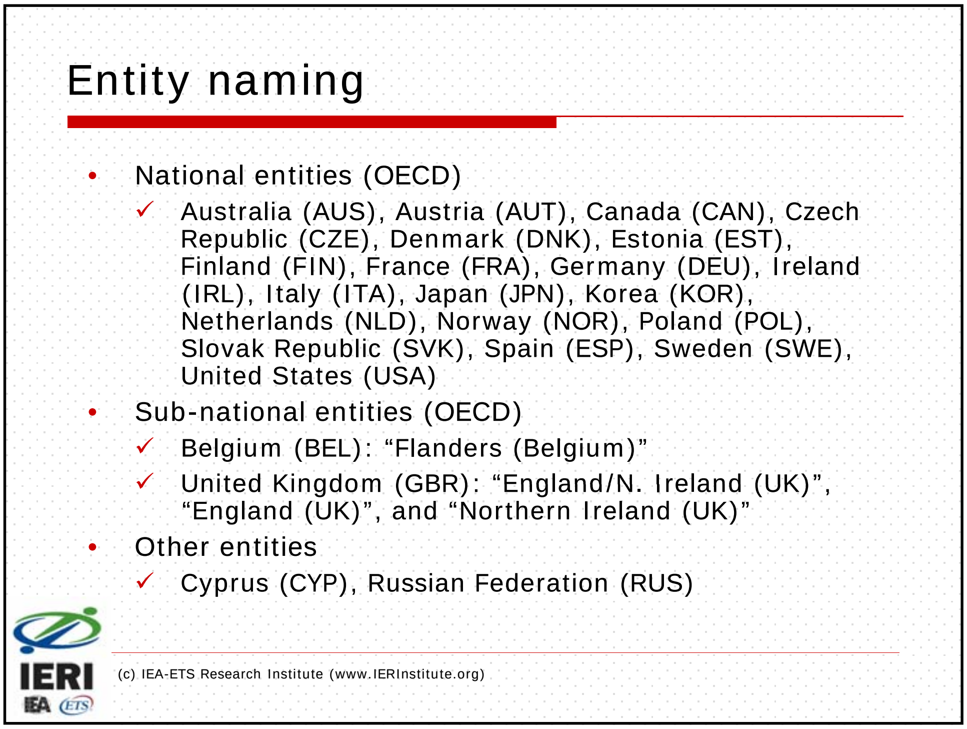
- 23 of 24 available through PIAAC Data Explorer
  - ✓ https://piaacdataexplorer.oecd.org/ide/idepiaac
  - ✓ Includes Australia, excludes Cyprus
- 23 of 24 available/distributed on USB today
  - ✓ Includes Cyprus, excludes Australia
  - ✓ Refreshed data due to skill use update
- PUF for Australia (called “CURF”) available on application, not distributed
  - ✓ To access, write to microdata.access@abs.gov.au
- Extended public-/research-use files might be available from countries directly, e.g. U.S. and Germany



Documentation available

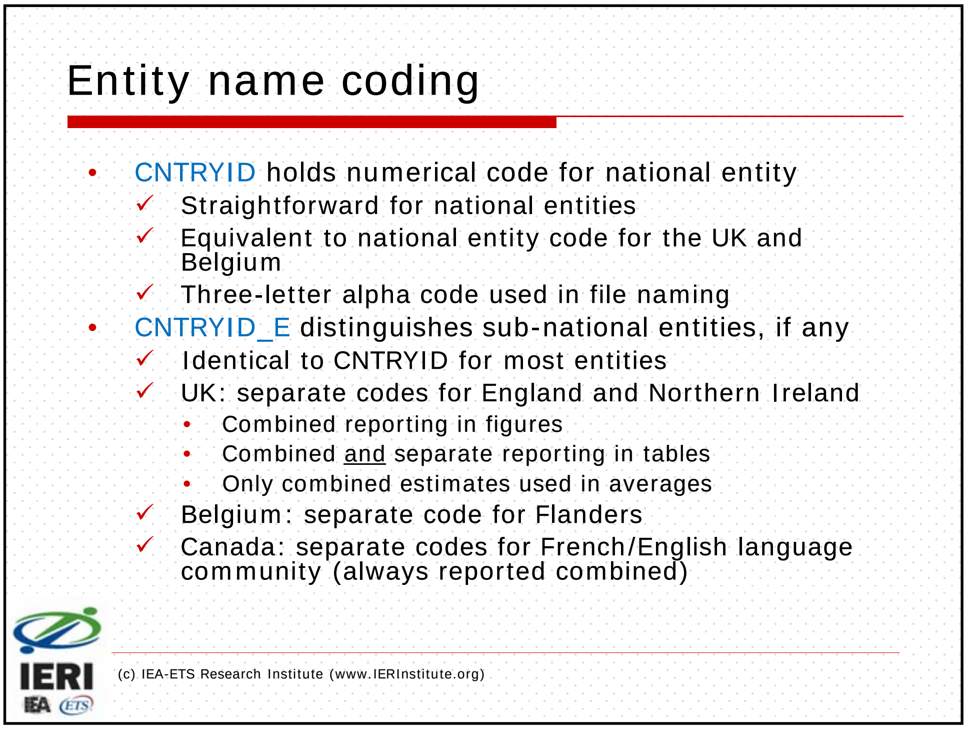
- Frameworks (BQ, Literacy, Numeracy, PSTRE)
- Background questionnaire
- International report, national reports
- Technical Report
- Proficiency level descriptions
- PUF codeplan (variables and values)
- Derived variable scripts
- Webpackages (tables and exhibits)
- Compendia for background and cognitive variables

- http://www.oecd.org/site/piaac/
  - ✓ Sections “Publications” and “Public Data & Analysis”



Entity naming

- National entities (OECD)
  - ✓ Australia (AUS), Austria (AUT), Canada (CAN), Czech Republic (CZE), Denmark (DNK), Estonia (EST), Finland (FIN), France (FRA), Germany (DEU), Ireland (IRL), Italy (ITA), Japan (JPN), Korea (KOR), Netherlands (NLD), Norway (NOR), Poland (POL), Slovak Republic (SVK), Spain (ESP), Sweden (SWE), United States (USA)
- Sub-national entities (OECD)
  - ✓ Belgium (BEL): "Flanders (Belgium)"
  - ✓ United Kingdom (GBR): "England/N. Ireland (UK)", "England (UK)", and "Northern Ireland (UK)"
- Other entities
  - ✓ Cyprus (CYP), Russian Federation (RUS)



Entity name coding

- CNTRYID holds numerical code for national entity
  - ✓ Straightforward for national entities
  - ✓ Equivalent to national entity code for the UK and Belgium
  - ✓ Three-letter alpha code used in file naming
- CNTRYID_E distinguishes sub-national entities, if any
  - ✓ Identical to CNTRYID for most entities
  - ✓ UK: separate codes for England and Northern Ireland
    - Combined reporting in figures
    - Combined <u>and</u> separate reporting in tables
    - Only combined estimates used in averages
  - ✓ Belgium: separate code for Flanders
  - ✓ Canada: separate codes for French/English language community (always reported combined)

Example exhibit (Table A2.1)

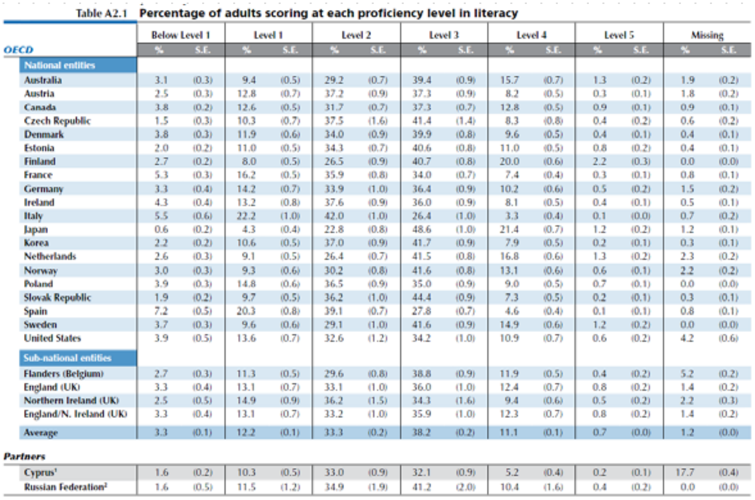


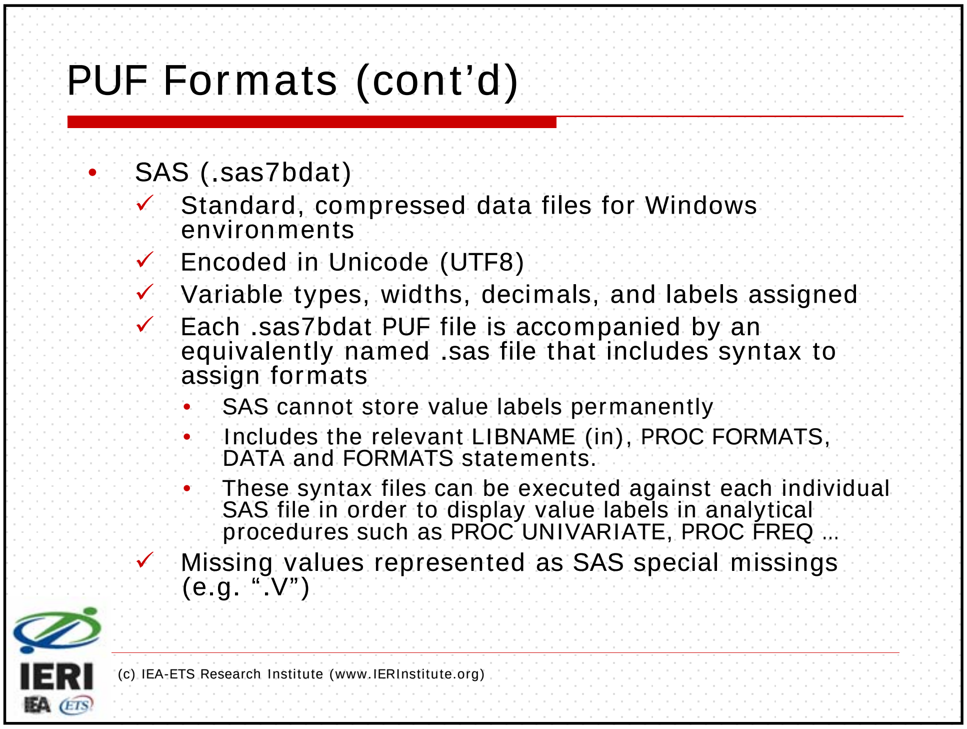
Table A2.1 **Percentage of adults scoring at each proficiency level in literacy**

	Below Level 1		Level 1		Level 2		Level 3		Level 4		Level 5		Missing	
OECD	%	S.E.	%	S.E.	%	S.E.	%	S.E.	%	S.E.	%	S.E.	%	S.E.
National entities														
Australia	3.1	(0.3)	9.4	(0.5)	29.2	(0.7)	39.4	(0.9)	15.7	(0.7)	1.3	(0.2)	1.9	(0.2)
Austria	2.5	(0.3)	12.8	(0.7)	37.2	(0.9)	37.3	(0.9)	8.2	(0.5)	0.3	(0.1)	1.8	(0.2)
Canada	3.8	(0.2)	12.6	(0.5)	31.7	(0.7)	37.3	(0.7)	12.8	(0.5)	0.9	(0.1)	0.9	(0.1)
Czech Republic	1.5	(0.3)	10.3	(0.7)	37.5	(1.6)	41.4	(1.4)	8.3	(0.8)	0.4	(0.2)	0.6	(0.2)
Denmark	3.8	(0.3)	11.9	(0.6)	34.0	(0.9)	39.9	(0.8)	9.6	(0.5)	0.4	(0.1)	0.4	(0.1)
Estonia	2.0	(0.2)	11.0	(0.5)	34.3	(0.7)	40.6	(0.8)	11.0	(0.5)	0.8	(0.2)	0.4	(0.1)
Finland	2.7	(0.2)	8.0	(0.5)	26.5	(0.9)	40.7	(0.8)	20.0	(0.6)	2.2	(0.3)	0.0	(0.0)
France	5.3	(0.3)	16.2	(0.5)	35.9	(0.8)	34.0	(0.7)	7.4	(0.4)	0.3	(0.1)	0.8	(0.1)
Germany	3.3	(0.4)	14.2	(0.7)	33.9	(1.0)	36.4	(0.9)	10.2	(0.6)	0.5	(0.2)	1.5	(0.2)
Ireland	4.3	(0.4)	13.2	(0.8)	37.6	(0.9)	36.0	(0.9)	8.1	(0.5)	0.4	(0.1)	0.5	(0.1)
Italy	5.5	(0.6)	22.2	(1.0)	42.0	(1.0)	26.4	(1.0)	3.3	(0.4)	0.1	(0.0)	0.7	(0.2)
Japan	0.6	(0.2)	4.3	(0.4)	22.8	(0.8)	48.6	(1.0)	21.4	(0.7)	1.2	(0.2)	1.2	(0.1)
Korea	2.2	(0.2)	10.6	(0.5)	37.0	(0.9)	41.7	(0.9)	7.9	(0.5)	0.2	(0.1)	0.3	(0.1)
Netherlands	2.6	(0.3)	9.1	(0.5)	26.4	(0.7)	41.5	(0.8)	16.8	(0.6)	1.3	(0.2)	2.3	(0.2)
Norway	3.0	(0.3)	9.3	(0.6)	30.2	(0.8)	41.6	(0.8)	13.1	(0.6)	0.6	(0.1)	2.2	(0.2)
Poland	3.9	(0.3)	14.8	(0.6)	36.5	(0.9)	35.0	(0.9)	9.0	(0.5)	0.7	(0.1)	0.0	(0.0)
Slovak Republic	1.9	(0.2)	9.7	(0.5)	36.2	(1.0)	44.4	(0.9)	7.3	(0.5)	0.2	(0.1)	0.3	(0.1)
Spain	7.2	(0.5)	20.3	(0.8)	39.1	(0.7)	27.8	(0.7)	4.6	(0.4)	0.1	(0.1)	0.8	(0.1)
Sweden	3.7	(0.3)	9.6	(0.6)	29.1	(1.0)	41.6	(0.9)	14.9	(0.6)	1.2	(0.2)	0.0	(0.0)
United States	3.9	(0.5)	13.6	(0.7)	32.6	(1.2)	34.2	(1.0)	10.9	(0.7)	0.6	(0.2)	4.2	(0.6)
Sub-national entities														
Flanders (Belgium)	2.7	(0.3)	11.3	(0.5)	29.6	(0.8)	38.8	(0.9)	11.9	(0.5)	0.4	(0.2)	5.2	(0.2)
England (UK)	3.3	(0.4)	13.1	(0.7)	33.1	(1.0)	36.0	(1.0)	12.4	(0.7)	0.8	(0.2)	1.4	(0.2)
Northern Ireland (UK)	2.5	(0.5)	14.9	(0.9)	36.2	(1.5)	34.3	(1.6)	9.4	(0.6)	0.5	(0.2)	2.2	(0.3)
England/N. Ireland (UK)	3.3	(0.4)	13.1	(0.7)	33.2	(1.0)	35.9	(1.0)	12.3	(0.7)	0.8	(0.2)	1.4	(0.2)
Average	3.3	(0.1)	12.2	(0.1)	33.3	(0.2)	38.2	(0.2)	11.1	(0.1)	0.7	(0.0)	1.2	(0.0)
Partners														
Cyprus[1]	1.6	(0.2)	10.3	(0.5)	33.0	(0.9)	32.1	(0.9)	5.2	(0.4)	0.2	(0.1)	17.7	(0.4)
Russian Federation[2]	1.6	(0.5)	11.5	(1.2)	34.9	(1.9)	41.2	(2.0)	10.4	(1.6)	0.4	(0.2)	0.0	(0.0)

Source: OECD (2013), *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills,* OECD Publishing. Revised version, November 2013

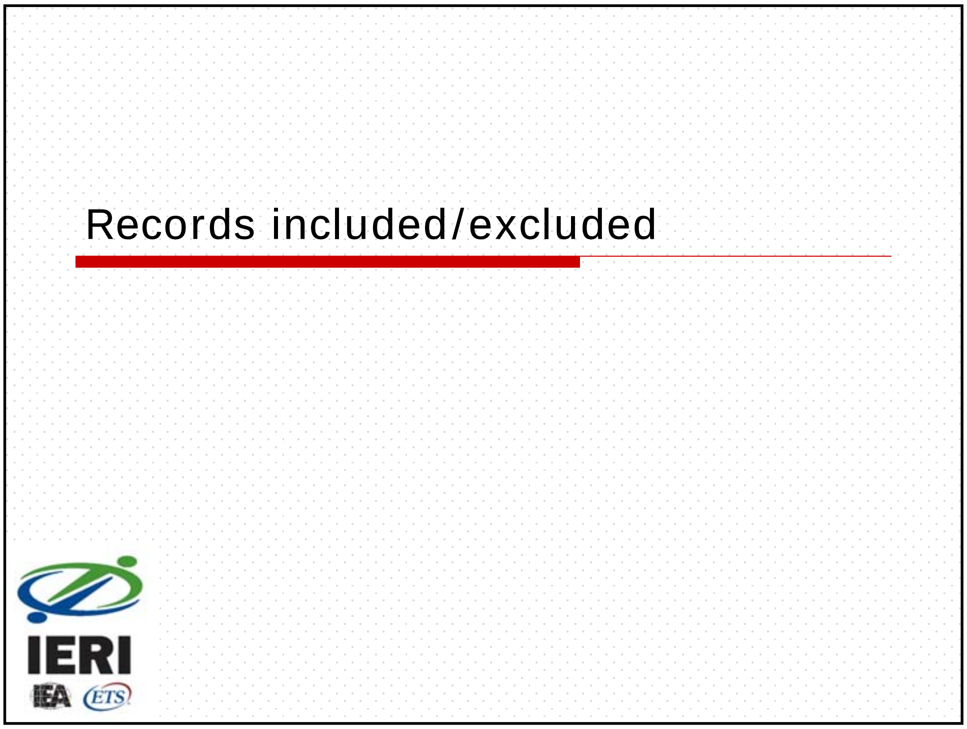
PUF Formats

- PUFs available in two standard formats
- SPSS (.sav) for version 11 or later
  - ✓ Unicode (UTF8) encoded to preserve national strings
  - ✓ Full dictionary information
    - Variable types and formats
    - Variable labels
    - Value labels (including any missing value labels)
    - Missing value definitions (except for strings)
    - Variable measurement levels
  - ✓ Missing values represented numerically (more later)

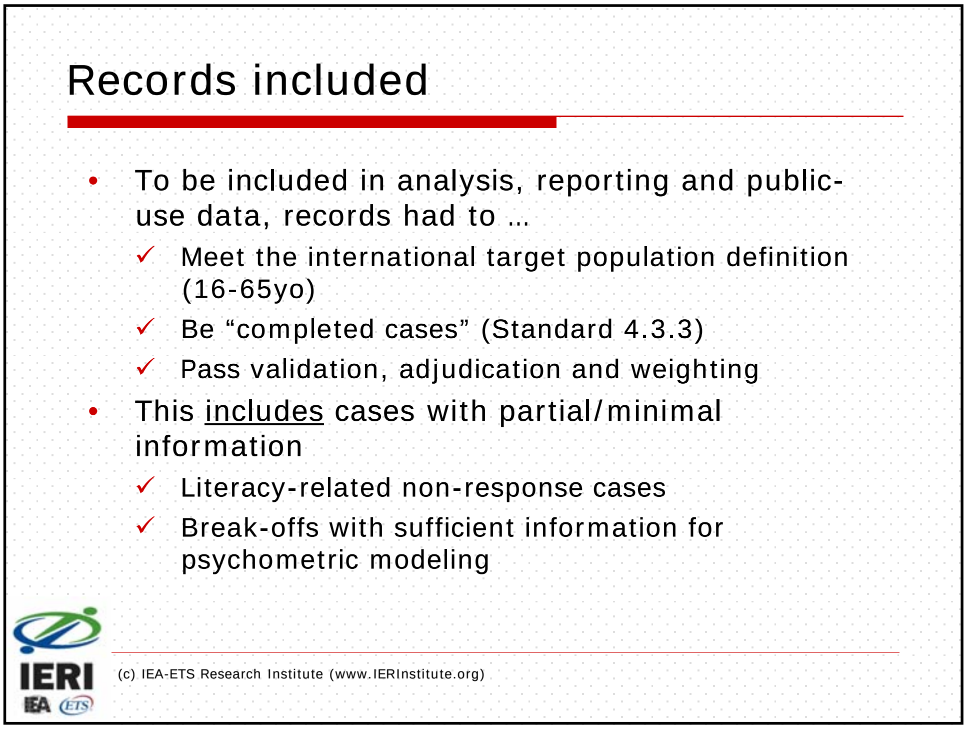


PUF Formats (cont'd)

- SAS (.sas7bdat)
  - ✓ Standard, compressed data files for Windows environments
  - ✓ Encoded in Unicode (UTF8)
  - ✓ Variable types, widths, decimals, and labels assigned
  - ✓ Each .sas7bdat PUF file is accompanied by an equivalently named .sas file that includes syntax to assign formats
    - SAS cannot store value labels permanently
    - Includes the relevant LIBNAME (in), PROC FORMATS, DATA and FORMATS statements.
    - These syntax files can be executed against each individual SAS file in order to display value labels in analytical procedures such as PROC UNIVARIATE, PROC FREQ …
  - ✓ Missing values represented as SAS special missings (e.g. ".V")

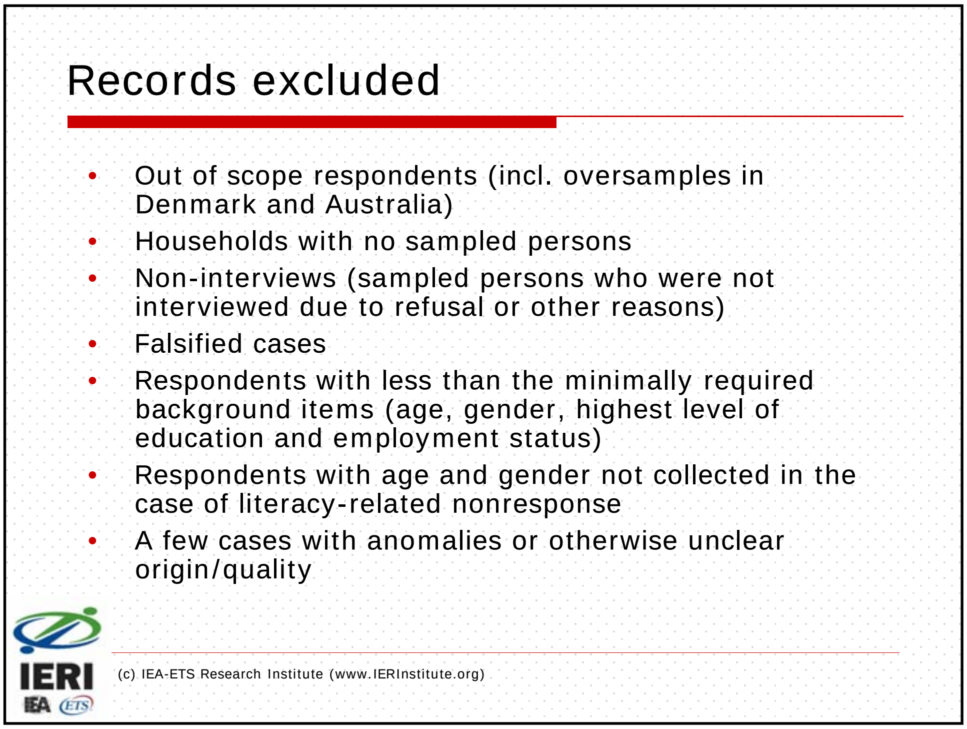


Records included/excluded



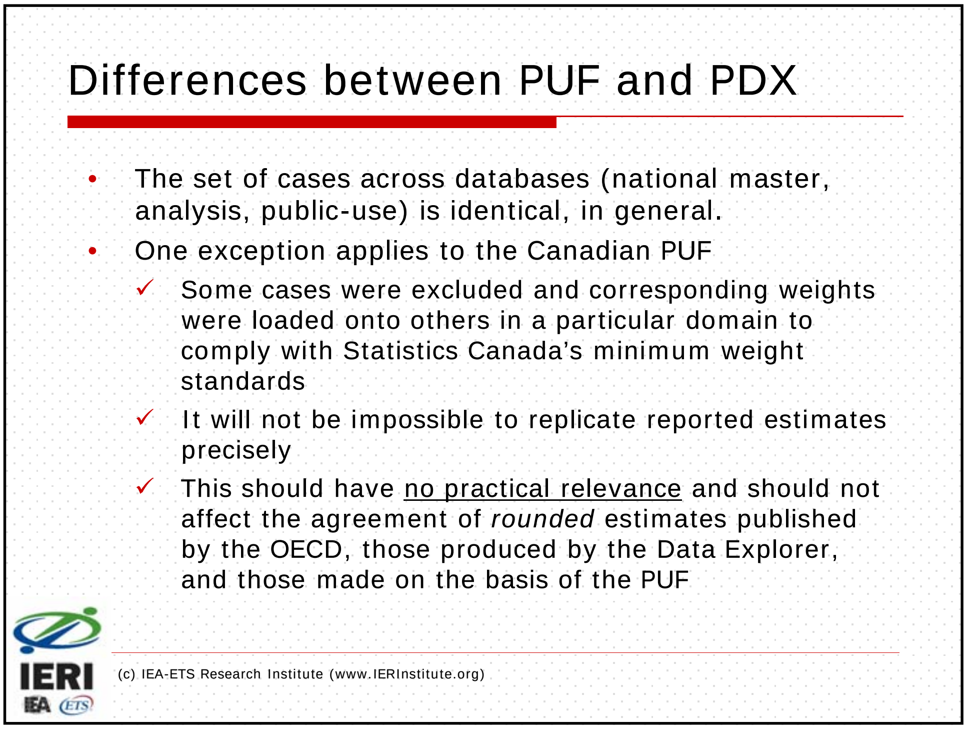
Records included

- To be included in analysis, reporting and public-use data, records had to …
  - ✓ Meet the international target population definition (16-65yo)
  - ✓ Be “completed cases” (Standard 4.3.3)
  - ✓ Pass validation, adjudication and weighting
- This <u>includes</u> cases with partial/minimal information
  - ✓ Literacy-related non-response cases
  - ✓ Break-offs with sufficient information for psychometric modeling



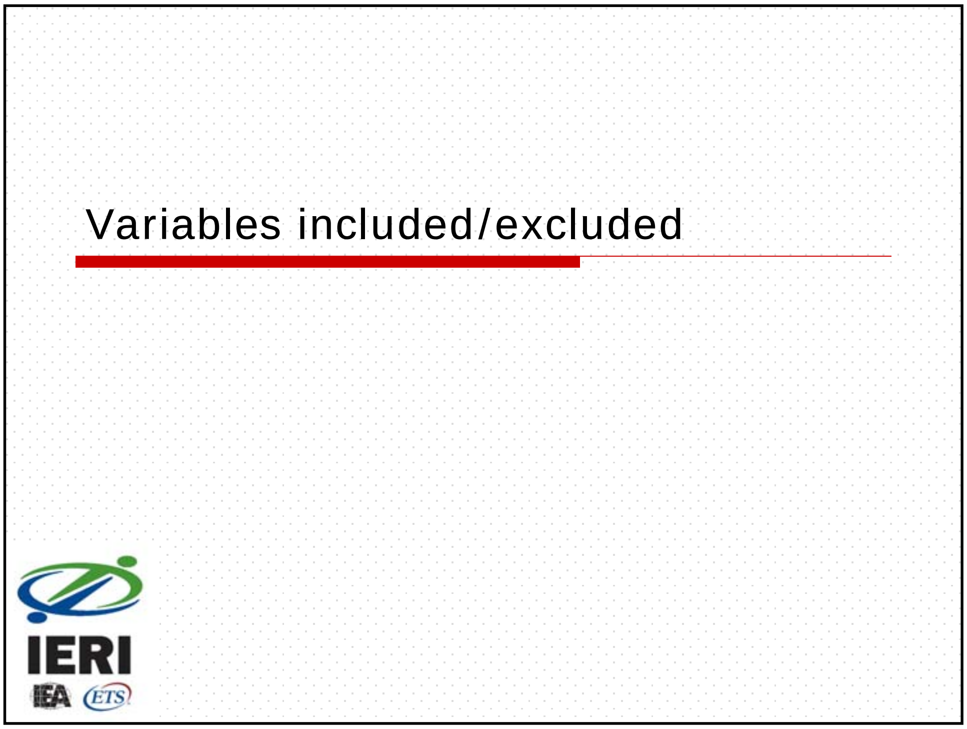
Records excluded

- Out of scope respondents (incl. oversamples in Denmark and Australia)
- Households with no sampled persons
- Non-interviews (sampled persons who were not interviewed due to refusal or other reasons)
- Falsified cases
- Respondents with less than the minimally required background items (age, gender, highest level of education and employment status)
- Respondents with age and gender not collected in the case of literacy-related nonresponse
- A few cases with anomalies or otherwise unclear origin/quality

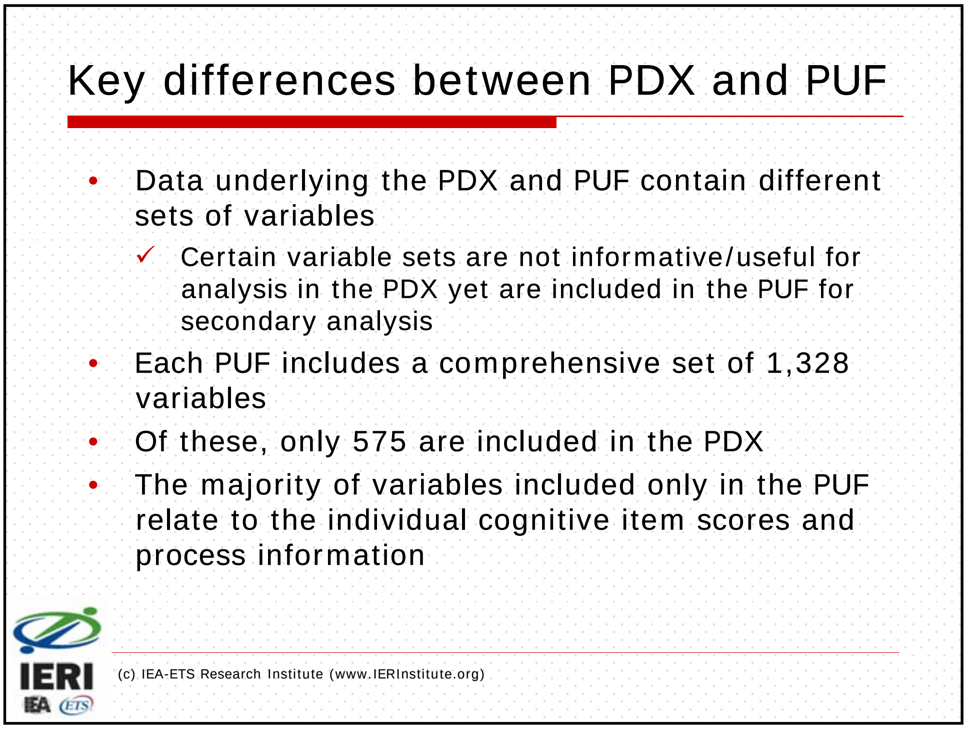


Differences between PUF and PDX

- The set of cases across databases (national master, analysis, public-use) is identical, in general.
- One exception applies to the Canadian PUF
  - ✓ Some cases were excluded and corresponding weights were loaded onto others in a particular domain to comply with Statistics Canada’s minimum weight standards
  - ✓ It will not be impossible to replicate reported estimates precisely
  - ✓ This should have no practical relevance and should not affect the agreement of *rounded* estimates published by the OECD, those produced by the Data Explorer, and those made on the basis of the PUF

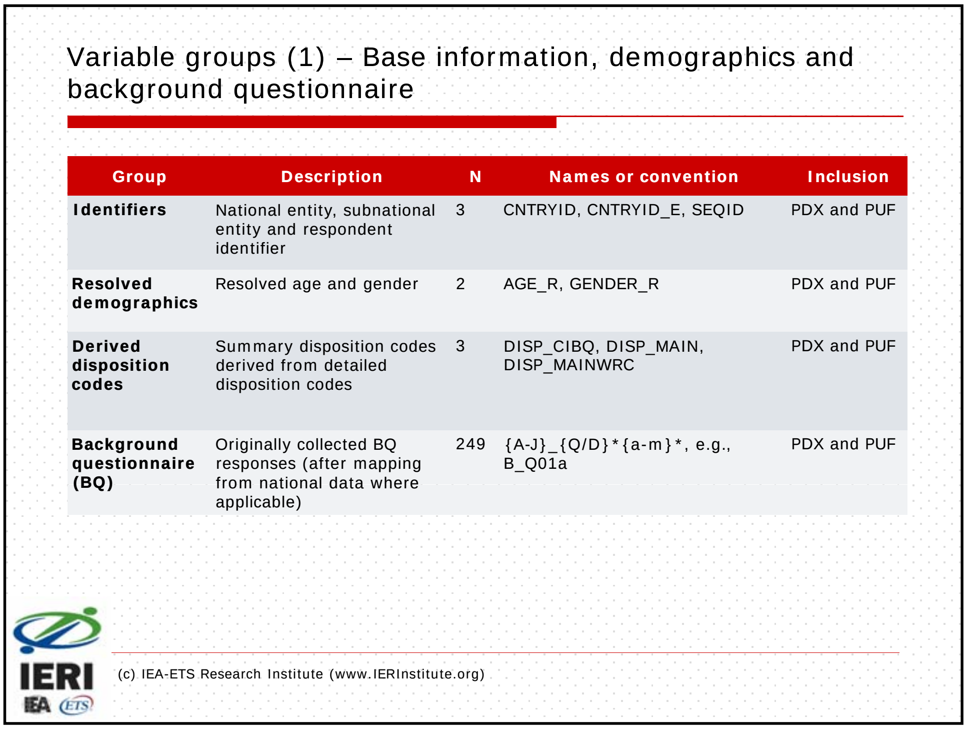


Variables included/excluded



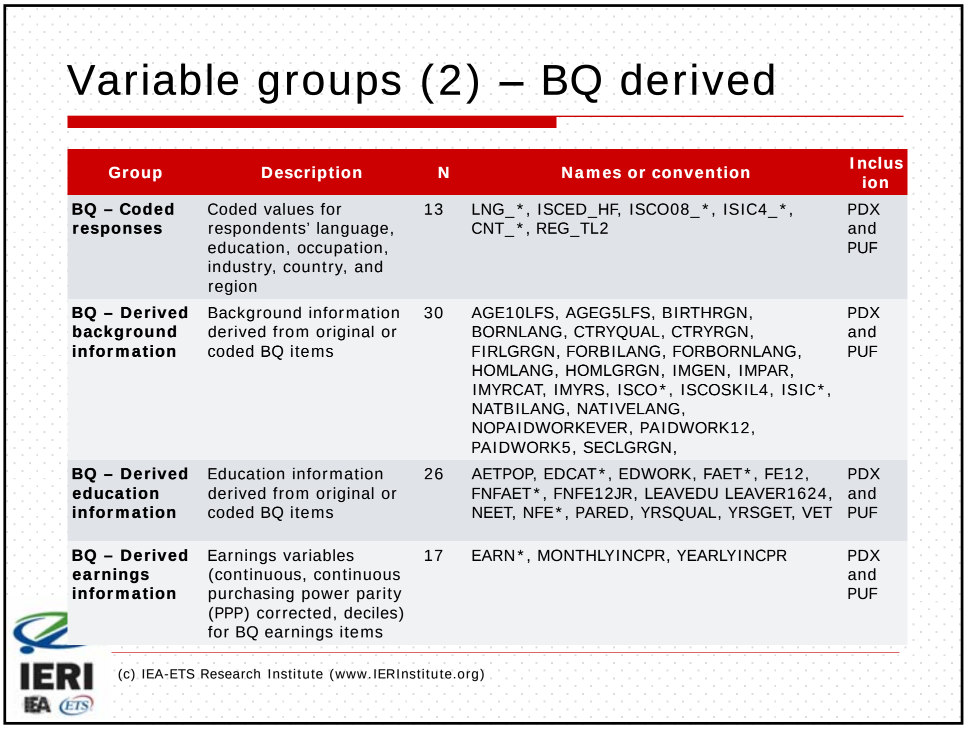
Key differences between PDX and PUF

- Data underlying the PDX and PUF contain different sets of variables
  - ✓ Certain variable sets are not informative/useful for analysis in the PDX yet are included in the PUF for secondary analysis
- Each PUF includes a comprehensive set of 1,328 variables
- Of these, only 575 are included in the PDX
- The majority of variables included only in the PUF relate to the individual cognitive item scores and process information



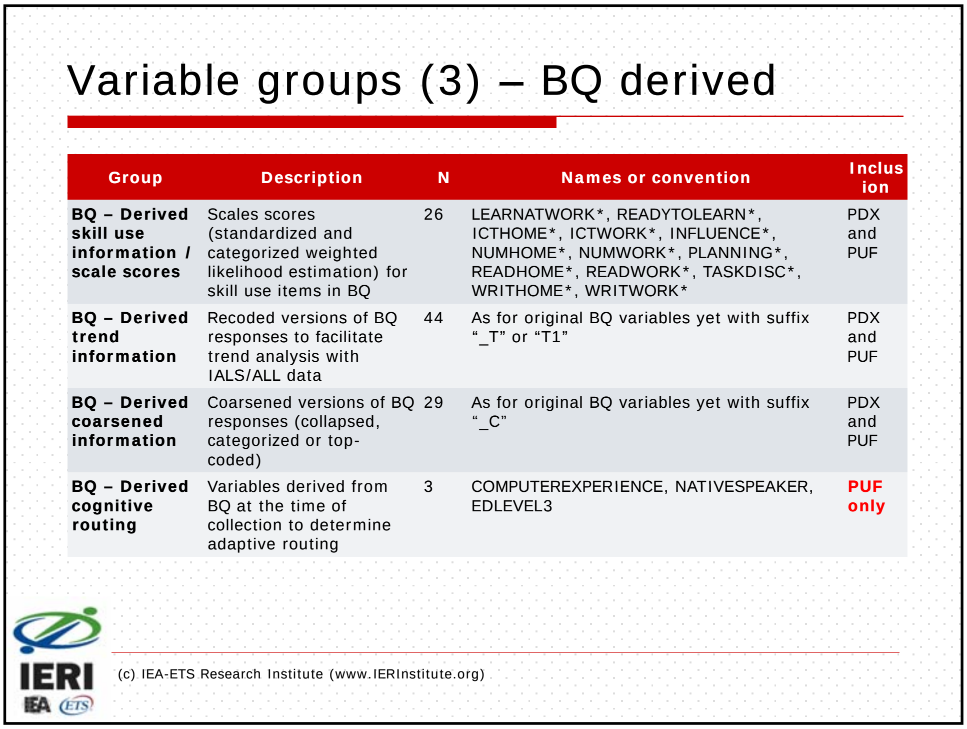
Variable groups (1) – Base information, demographics and background questionnaire

Group	Description	N	Names or convention	Inclusion
Identifiers	National entity, subnational entity and respondent identifier	3	CNTRYID, CNTRYID_E, SEQID	PDX and PUF
Resolved demographics	Resolved age and gender	2	AGE_R, GENDER_R	PDX and PUF
Derived disposition codes	Summary disposition codes derived from detailed disposition codes	3	DISP_CIBQ, DISP_MAIN, DISP_MAINWRC	PDX and PUF
Background questionnaire (BQ)	Originally collected BQ responses (after mapping from national data where applicable)	249	{A-J}_{Q/D}*{a-m}*, e.g., B_Q01a	PDX and PUF



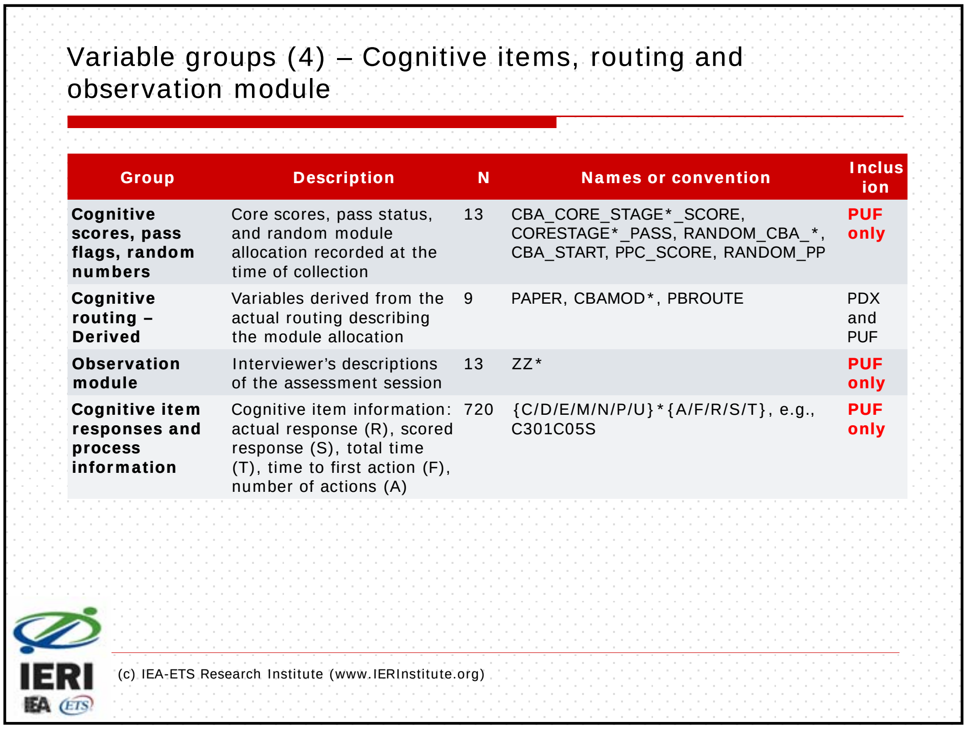
Variable groups (2) – BQ derived

Group	Description	N	Names or convention	Inclusion
BQ – Coded responses	Coded values for respondents' language, education, occupation, industry, country, and region	13	LNG_*, ISCED_HF, ISCO08_*, ISIC4_*, CNT_*, REG_TL2	PDX and PUF
BQ – Derived background information	Background information derived from original or coded BQ items	30	AGE10LFS, AGEG5LFS, BIRTHRGN, BORNLANG, CTRYQUAL, CTRYRGN, FIRLGRGN, FORBILANG, FORBORNLANG, HOMLANG, HOMLGRGN, IMGEN, IMPAR, IMYRCAT, IMYRS, ISCO*, ISCOSKIL4, ISIC*, NATBILANG, NATIVELANG, NOPAIDWORKEVER, PAIDWORK12, PAIDWORK5, SECLGRGN,	PDX and PUF
BQ – Derived education information	Education information derived from original or coded BQ items	26	AETPOP, EDCAT*, EDWORK, FAET*, FE12, FNFAET*, FNFE12JR, LEAVEDU LEAVER1624, NEET, NFE*, PARED, YRSQUAL, YRSGET, VET	PDX and PUF
BQ – Derived earnings information	Earnings variables (continuous, continuous purchasing power parity (PPP) corrected, deciles) for BQ earnings items	17	EARN*, MONTHLYINCPR, YEARLYINCPR	PDX and PUF



Variable groups (3) – BQ derived

Group	Description	N	Names or convention	Inclusion
BQ – Derived skill use information / scale scores	Scales scores (standardized and categorized weighted likelihood estimation) for skill use items in BQ	26	LEARNATWORK*, READYTOLEARN*, ICTHOME*, ICTWORK*, INFLUENCE*, NUMHOME*, NUMWORK*, PLANNING*, READHOME*, READWORK*, TASKDISC*, WRITHOME*, WRITWORK*	PDX and PUF
BQ – Derived trend information	Recoded versions of BQ responses to facilitate trend analysis with IALS/ALL data	44	As for original BQ variables yet with suffix "_T" or "T1"	PDX and PUF
BQ – Derived coarsened information	Coarsened versions of BQ responses (collapsed, categorized or top-coded)	29	As for original BQ variables yet with suffix "_C"	PDX and PUF
BQ – Derived cognitive routing	Variables derived from BQ at the time of collection to determine adaptive routing	3	COMPUTEREXPERIENCE, NATIVESPEAKER, EDLEVEL3	**PUF only**

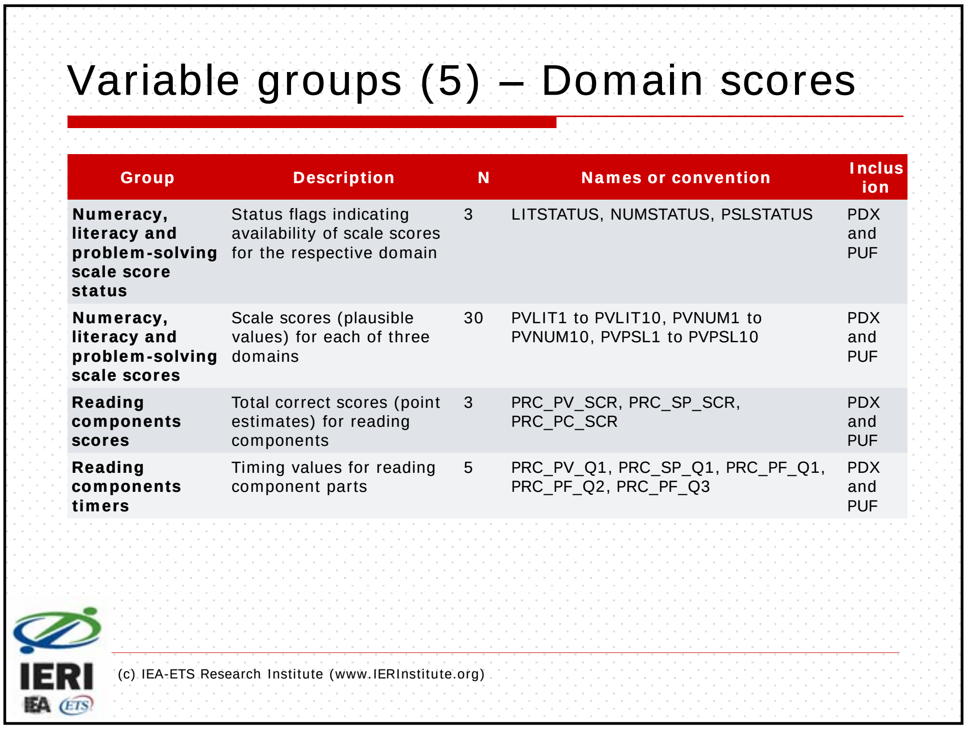


Variable groups (4) – Cognitive items, routing and observation module

Group	Description	N	Names or convention	Inclusion
Cognitive scores, pass flags, random numbers	Core scores, pass status, and random module allocation recorded at the time of collection	13	CBA_CORE_STAGE*_SCORE, CORESTAGE*_PASS, RANDOM_CBA_*, CBA_START, PPC_SCORE, RANDOM_PP	**PUF only**
Cognitive routing – Derived	Variables derived from the actual routing describing the module allocation	9	PAPER, CBAMOD*, PBROUTE	PDX and PUF
Observation module	Interviewer's descriptions of the assessment session	13	ZZ*	**PUF only**
Cognitive item responses and process information	Cognitive item information: actual response (R), scored response (S), total time (T), time to first action (F), number of actions (A)	720	{C/D/E/M/N/P/U}*{A/F/R/S/T}, e.g., C301C05S	**PUF only**

IERI IEA ETS

(c) IEA-ETS Research Institute (www.IERInstitute.org)

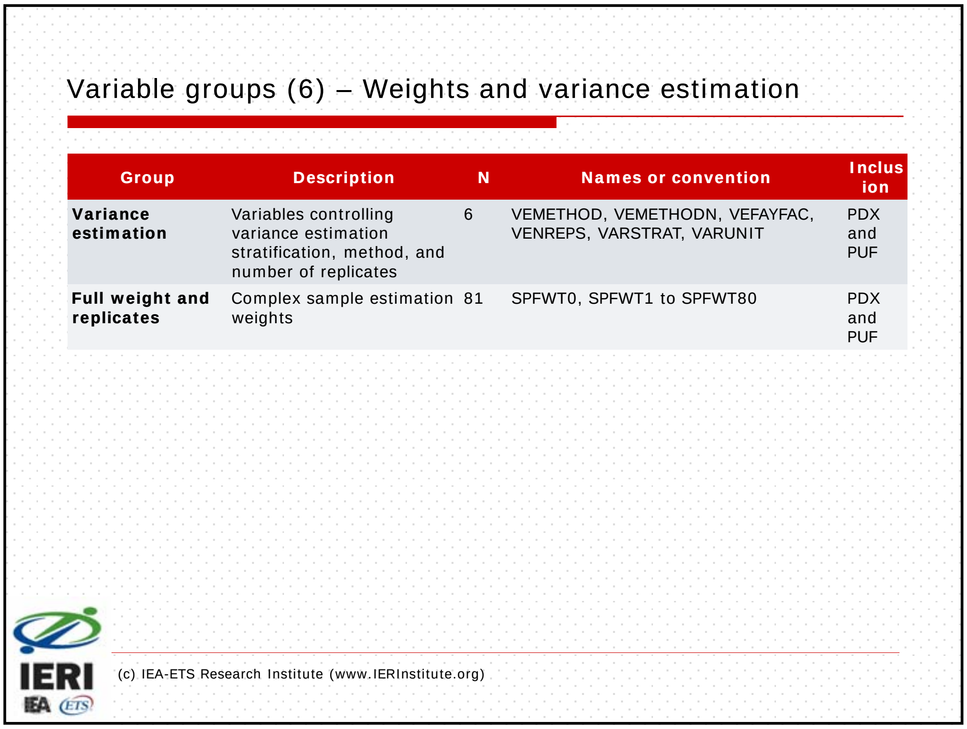


Variable groups (5) – Domain scores

Group	Description	N	Names or convention	Inclusion
Numeracy, literacy and problem-solving scale score status	Status flags indicating availability of scale scores for the respective domain	3	LITSTATUS, NUMSTATUS, PSLSTATUS	PDX and PUF
Numeracy, literacy and problem-solving scale scores	Scale scores (plausible values) for each of three domains	30	PVLIT1 to PVLIT10, PVNUM1 to PVNUM10, PVPSL1 to PVPSL10	PDX and PUF
Reading components scores	Total correct scores (point estimates) for reading components	3	PRC_PV_SCR, PRC_SP_SCR, PRC_PC_SCR	PDX and PUF
Reading components timers	Timing values for reading component parts	5	PRC_PV_Q1, PRC_SP_Q1, PRC_PF_Q1, PRC_PF_Q2, PRC_PF_Q3	PDX and PUF

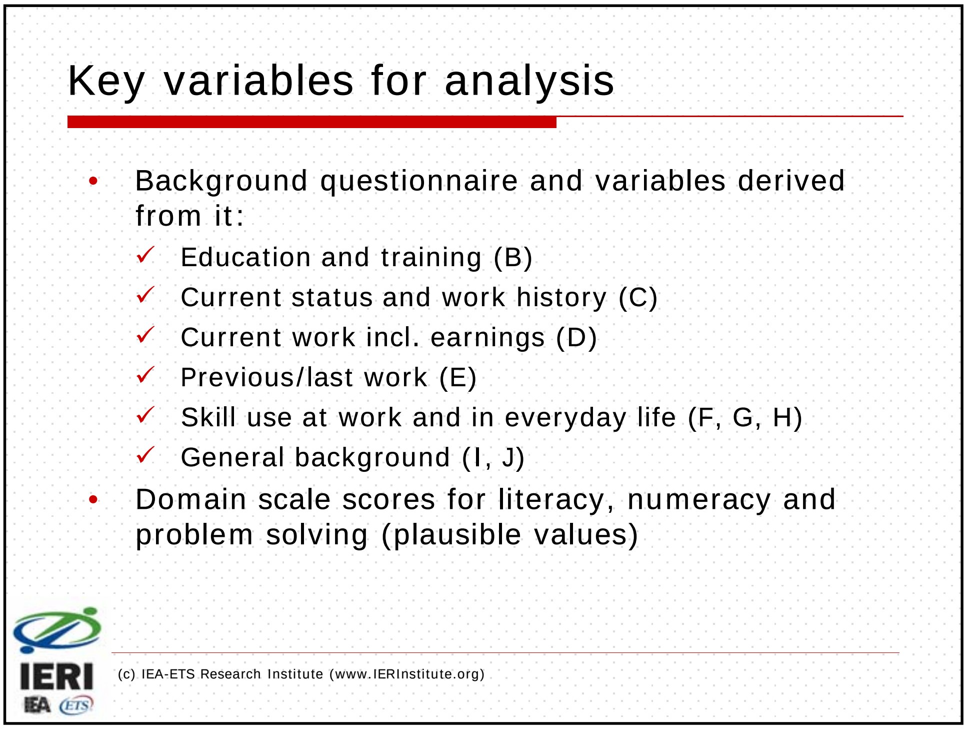
IERI IEA ETS

(c) IEA-ETS Research Institute (www.IERInstitute.org)



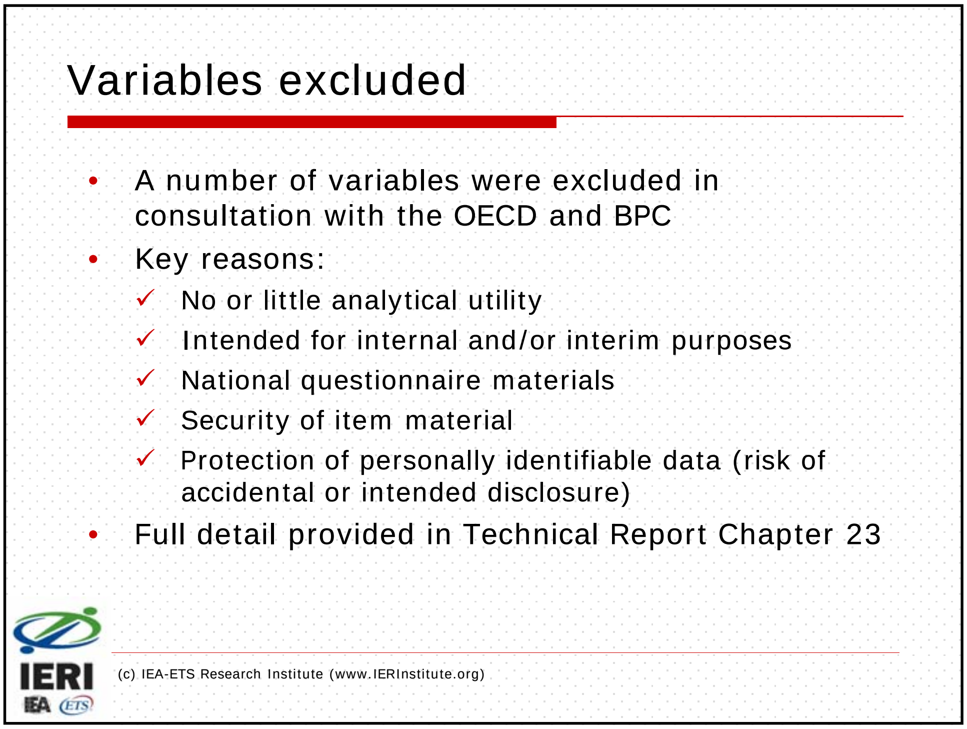
Variable groups (6) – Weights and variance estimation

Group	Description	N	Names or convention	Inclusion
Variance estimation	Variables controlling variance estimation stratification, method, and number of replicates	6	VEMETHOD, VEMETHODN, VEFAYFAC, VENREPS, VARSTRAT, VARUNIT	PDX and PUF
Full weight and replicates	Complex sample estimation weights	81	SPFWT0, SPFWT1 to SPFWT80	PDX and PUF



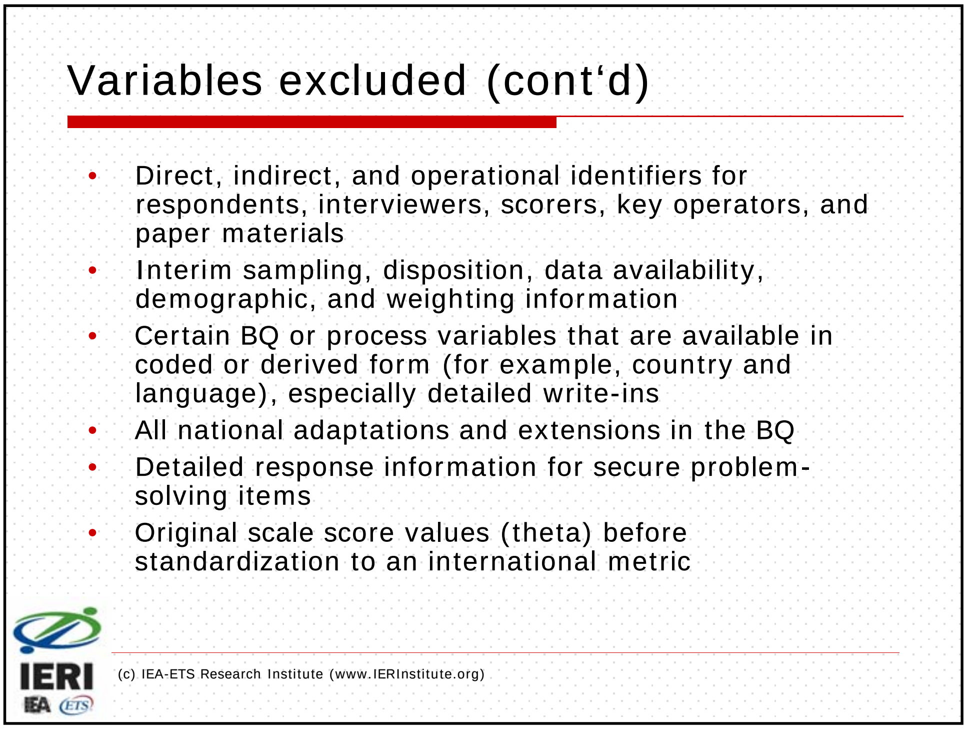
Key variables for analysis

- Background questionnaire and variables derived from it:
  - ✓ Education and training (B)
  - ✓ Current status and work history (C)
  - ✓ Current work incl. earnings (D)
  - ✓ Previous/last work (E)
  - ✓ Skill use at work and in everyday life (F, G, H)
  - ✓ General background (I, J)
- Domain scale scores for literacy, numeracy and problem solving (plausible values)



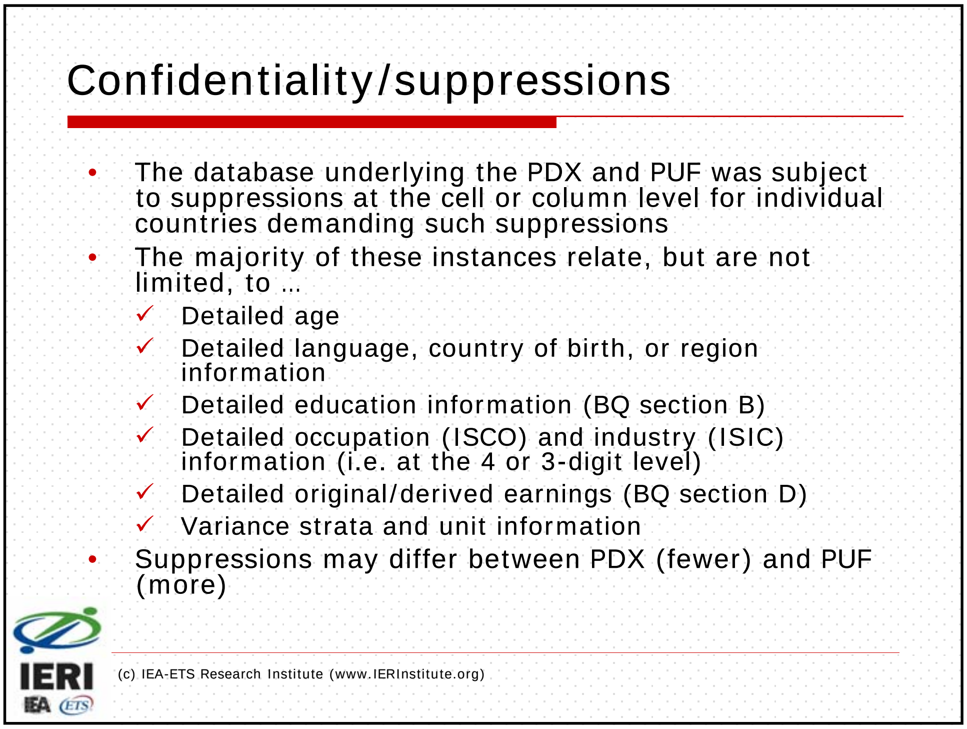
Variables excluded

- A number of variables were excluded in consultation with the OECD and BPC
- Key reasons:
  - ✓ No or little analytical utility
  - ✓ Intended for internal and/or interim purposes
  - ✓ National questionnaire materials
  - ✓ Security of item material
  - ✓ Protection of personally identifiable data (risk of accidental or intended disclosure)
- Full detail provided in Technical Report Chapter 23



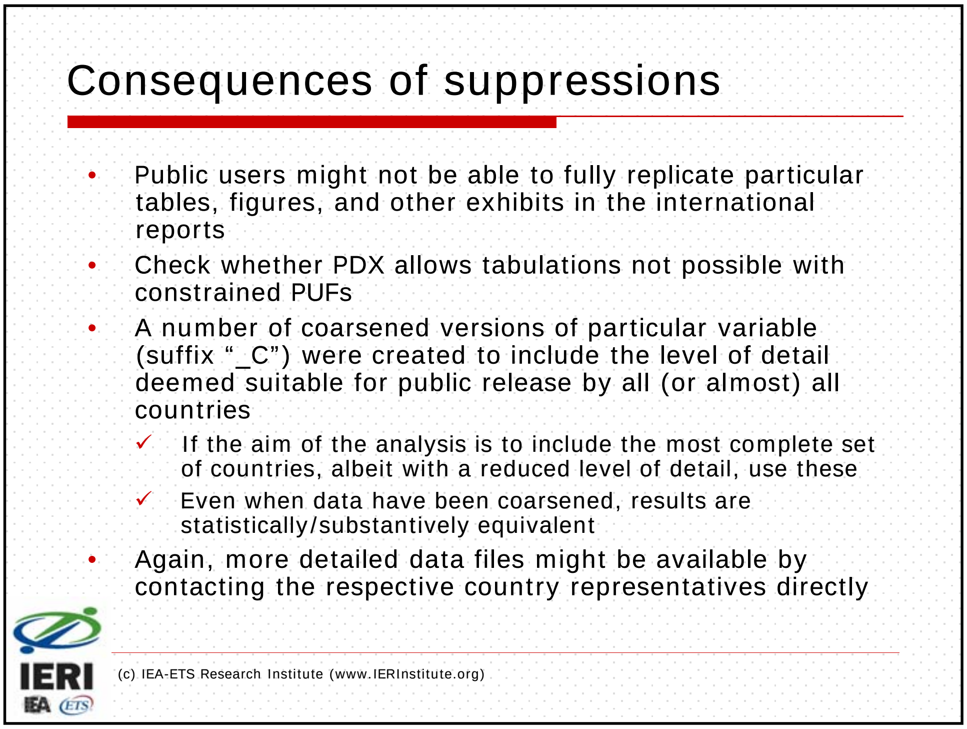
Variables excluded (cont‘d)

- Direct, indirect, and operational identifiers for respondents, interviewers, scorers, key operators, and paper materials
- Interim sampling, disposition, data availability, demographic, and weighting information
- Certain BQ or process variables that are available in coded or derived form (for example, country and language), especially detailed write-ins
- All national adaptations and extensions in the BQ
- Detailed response information for secure problem-solving items
- Original scale score values (theta) before standardization to an international metric



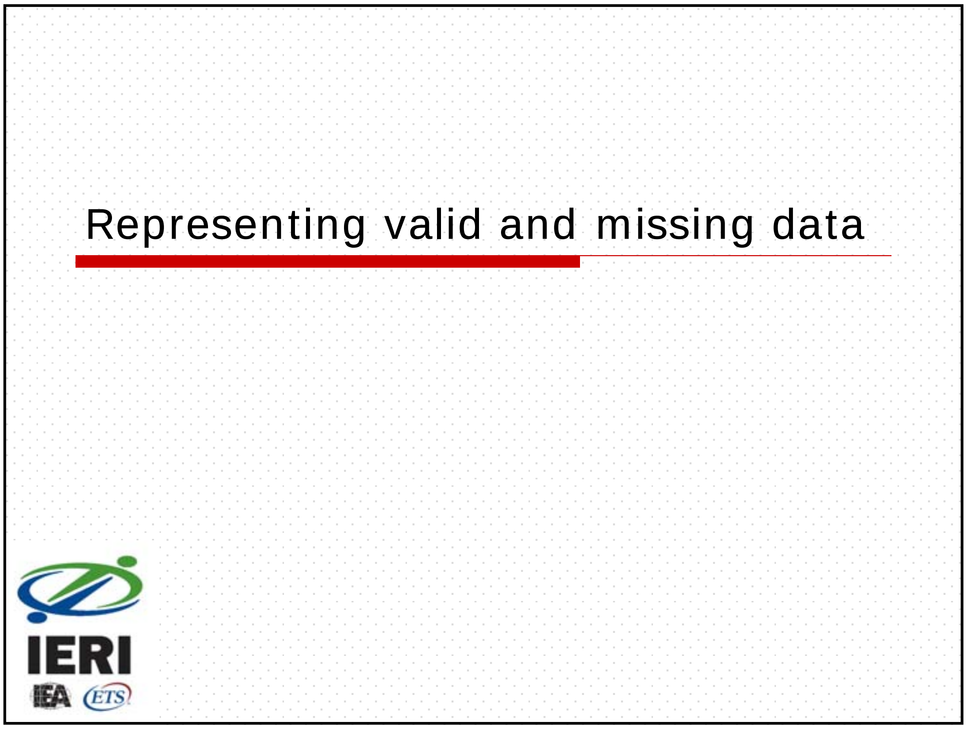
Confidentiality/suppressions

- The database underlying the PDX and PUF was subject to suppressions at the cell or column level for individual countries demanding such suppressions
- The majority of these instances relate, but are not limited, to …
  - ✓ Detailed age
  - ✓ Detailed language, country of birth, or region information
  - ✓ Detailed education information (BQ section B)
  - ✓ Detailed occupation (ISCO) and industry (ISIC) information (i.e. at the 4 or 3-digit level)
  - ✓ Detailed original/derived earnings (BQ section D)
  - ✓ Variance strata and unit information
- Suppressions may differ between PDX (fewer) and PUF (more)

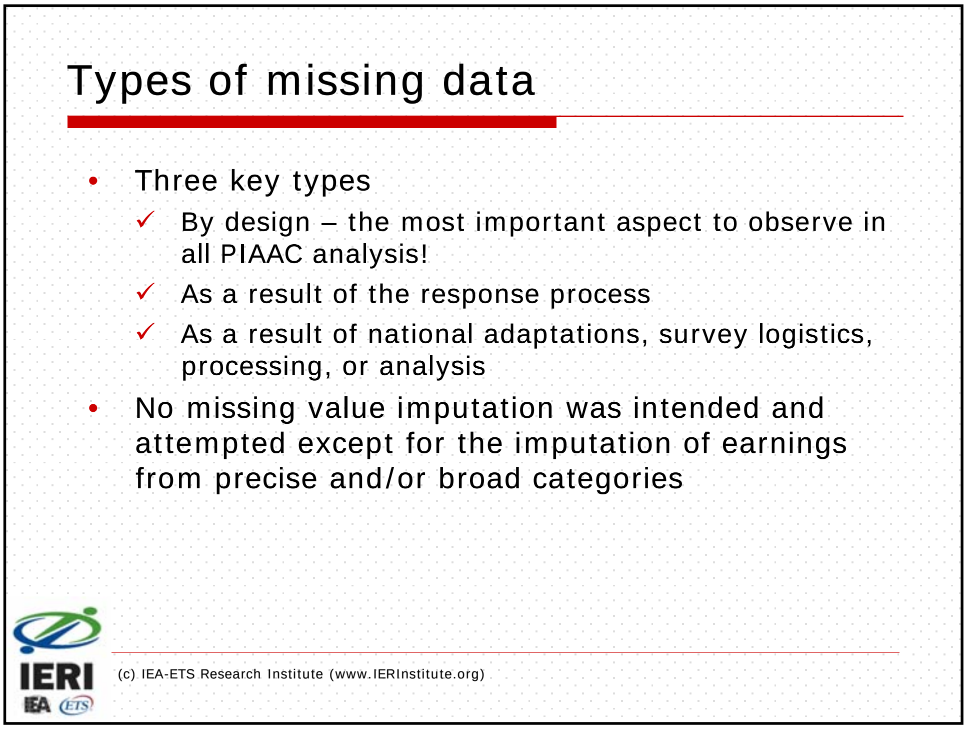


Consequences of suppressions

- Public users might not be able to fully replicate particular tables, figures, and other exhibits in the international reports
- Check whether PDX allows tabulations not possible with constrained PUFs
- A number of coarsened versions of particular variable (suffix "_C") were created to include the level of detail deemed suitable for public release by all (or almost) all countries
  - ✓ If the aim of the analysis is to include the most complete set of countries, albeit with a reduced level of detail, use these
  - ✓ Even when data have been coarsened, results are statistically/substantively equivalent
- Again, more detailed data files might be available by contacting the respective country representatives directly

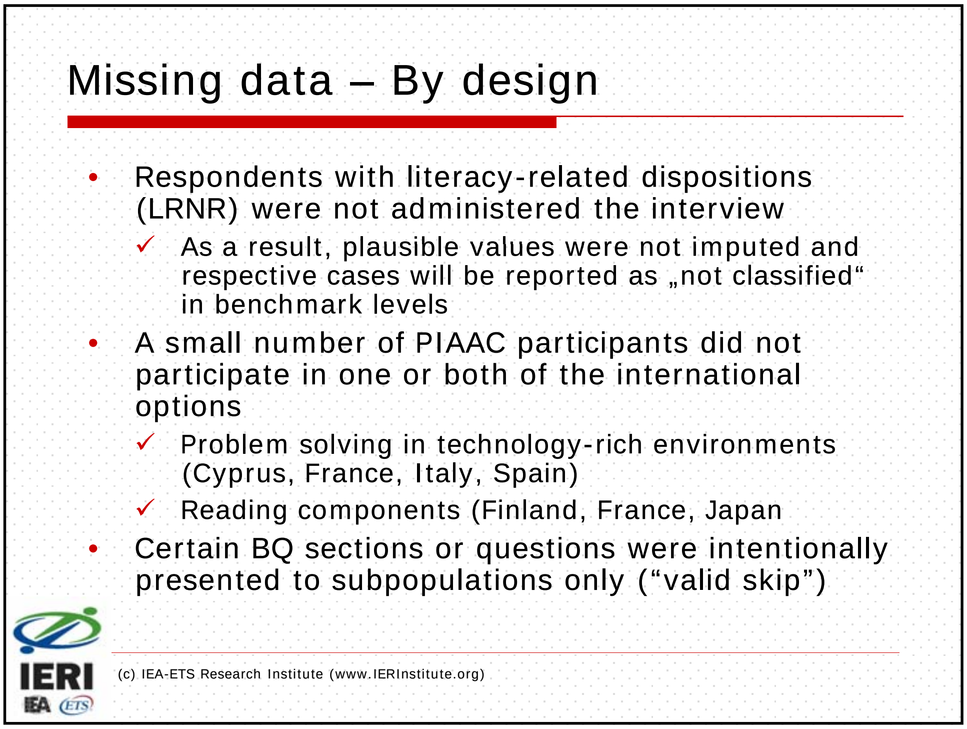


Representing valid and missing data



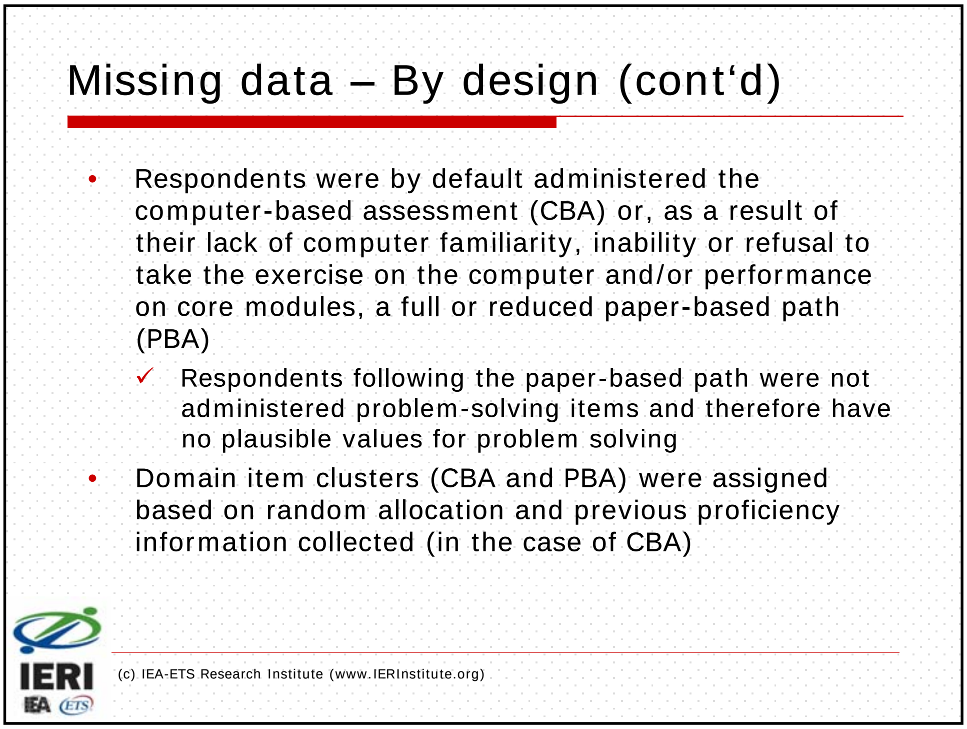
Types of missing data

- Three key types
  - ✓ By design – the most important aspect to observe in all PIAAC analysis!
  - ✓ As a result of the response process
  - ✓ As a result of national adaptations, survey logistics, processing, or analysis
- No missing value imputation was intended and attempted except for the imputation of earnings from precise and/or broad categories



Missing data – By design

- Respondents with literacy-related dispositions (LRNR) were not administered the interview
  - ✓ As a result, plausible values were not imputed and respective cases will be reported as „not classified“ in benchmark levels
- A small number of PIAAC participants did not participate in one or both of the international options
  - ✓ Problem solving in technology-rich environments (Cyprus, France, Italy, Spain)
  - ✓ Reading components (Finland, France, Japan
- Certain BQ sections or questions were intentionally presented to subpopulations only (“valid skip”)



Missing data – By design (cont‘d)

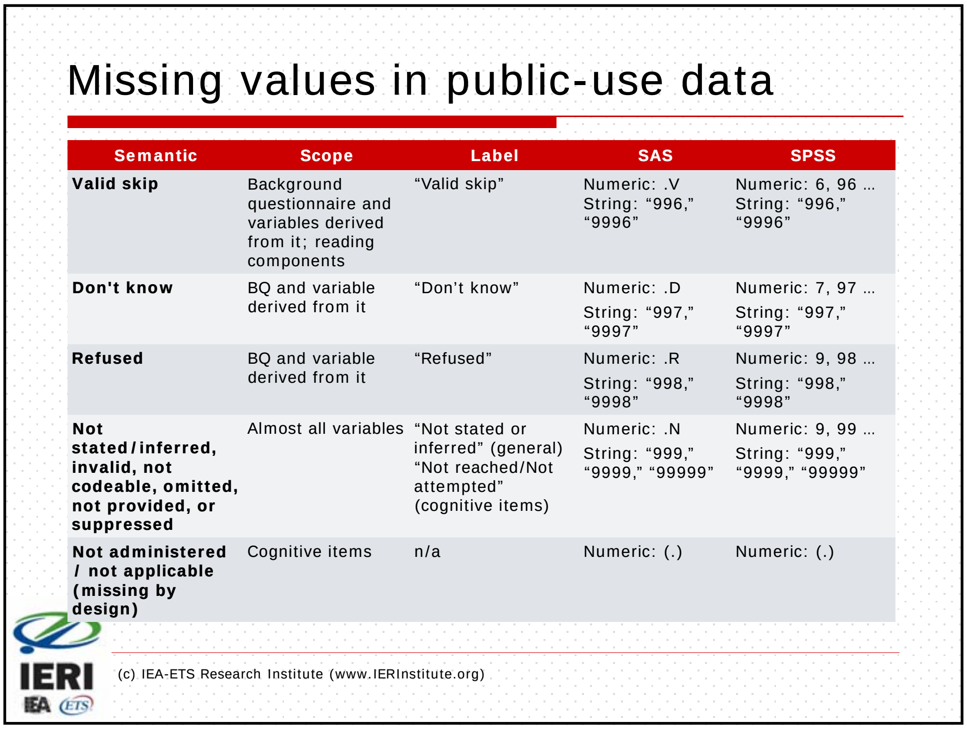
- Respondents were by default administered the computer-based assessment (CBA) or, as a result of their lack of computer familiarity, inability or refusal to take the exercise on the computer and/or performance on core modules, a full or reduced paper-based path (PBA)
  - ✓ Respondents following the paper-based path were not administered problem-solving items and therefore have no plausible values for problem solving
- Domain item clusters (CBA and PBA) were assigned based on random allocation and previous proficiency information collected (in the case of CBA)

Missing data – Response process

- Respondents may have broken off the interview after it was started as a function of, for example, time, motivation, fatigue, or sensitive questions being asked
- Respondents may have explicitly refused (“refused”) to respond to questions in the BQ or they may not have known the answer to a question with sufficient certainty (“don’t know”)

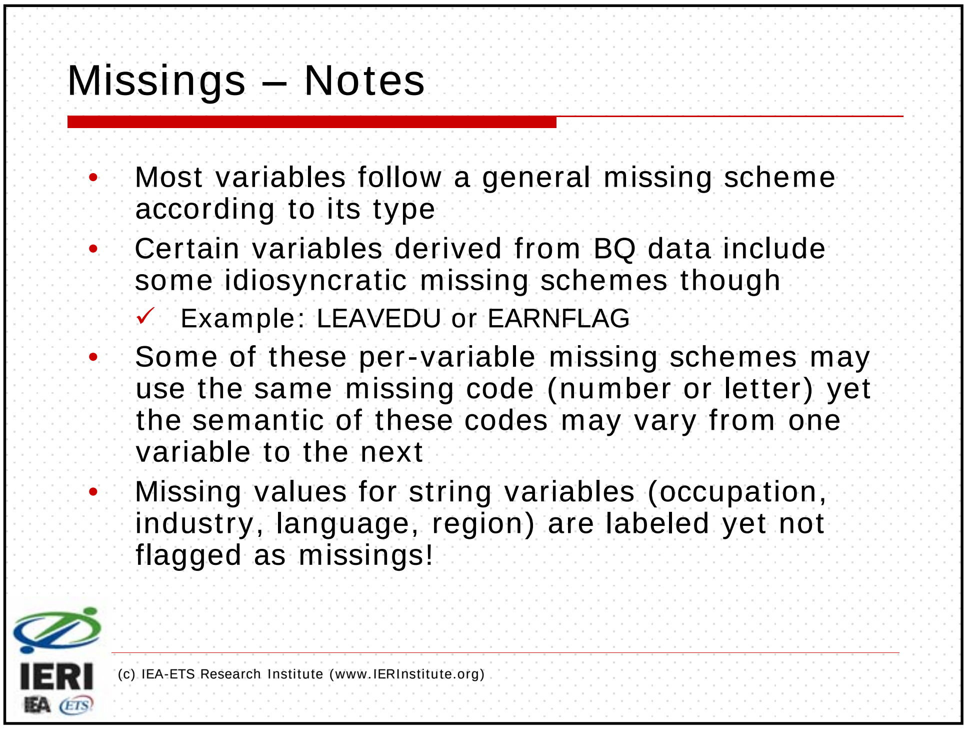
Missing data – Logistics, processing, analysis

- Erroneous routing in national versions of the BQ collected fewer data items for particular respondents than intended (very few instances)
- Certain data items (variables and/or a subset of values) were not provided or suppressed due to confidentiality concerns
- A small number of values were obvious outliers, otherwise useless, or erroneously coded in the original national databases



Missing values in public-use data

Semantic	Scope	Label	SAS	SPSS
Valid skip	Background questionnaire and variables derived from it; reading components	“Valid skip”	Numeric: .V String: “996,” “9996”	Numeric: 6, 96 … String: “996,” “9996”
Don't know	BQ and variable derived from it	“Don’t know”	Numeric: .D String: “997,” “9997”	Numeric: 7, 97 … String: “997,” “9997”
Refused	BQ and variable derived from it	“Refused”	Numeric: .R String: “998,” “9998”	Numeric: 9, 98 … String: “998,” “9998”
Not stated/inferred, invalid, not codeable, omitted, not provided, or suppressed	Almost all variables	“Not stated or inferred” (general) “Not reached/Not attempted” (cognitive items)	Numeric: .N String: “999,” “9999,” “99999”	Numeric: 9, 99 … String: “999,” “9999,” “99999”
Not administered / not applicable (missing by design)	Cognitive items	n/a	Numeric: (.)	Numeric: (.)



Missings – Notes

- Most variables follow a general missing scheme according to its type
- Certain variables derived from BQ data include some idiosyncratic missing schemes though
  - ✓ Example: LEAVEDU or EARNFLAG
- Some of these per-variable missing schemes may use the same missing code (number or letter) yet the semantic of these codes may vary from one variable to the next
- Missing values for string variables (occupation, industry, language, region) are labeled yet not flagged as missings!

Thanks for your attention!

Questions?

IERI

IEA ETS

evropský sociální fond v ČR

EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

DZS Dům zahraniční spolupráce

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ