USC Dornsife
Center for Economic and Social Research
Program on Global Aging, Health, and Policy

Data Harmonization

December 3, 2019

Gateway to Global Aging Data Team

Outline

- Introduction to harmonization
  - Gateway to Global Aging Data
  - Types of data
  - Basic principles
- Building a metadata library
  - Extracting and indexing metadata
  - Building search engine
  - Communicating data comparability
- Harmonizing phenotype data
  - Pre-statistical harmonization
  - Missing data analysis and imputation
  - Statistical harmonization

Introduction to Harmonization

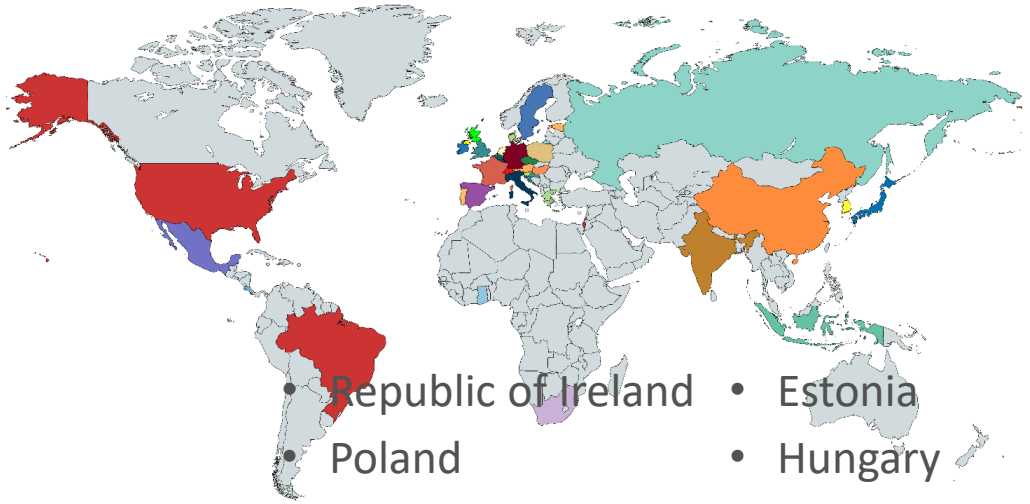
Gateway to Global Aging Data

The platform for population survey data on aging around the world

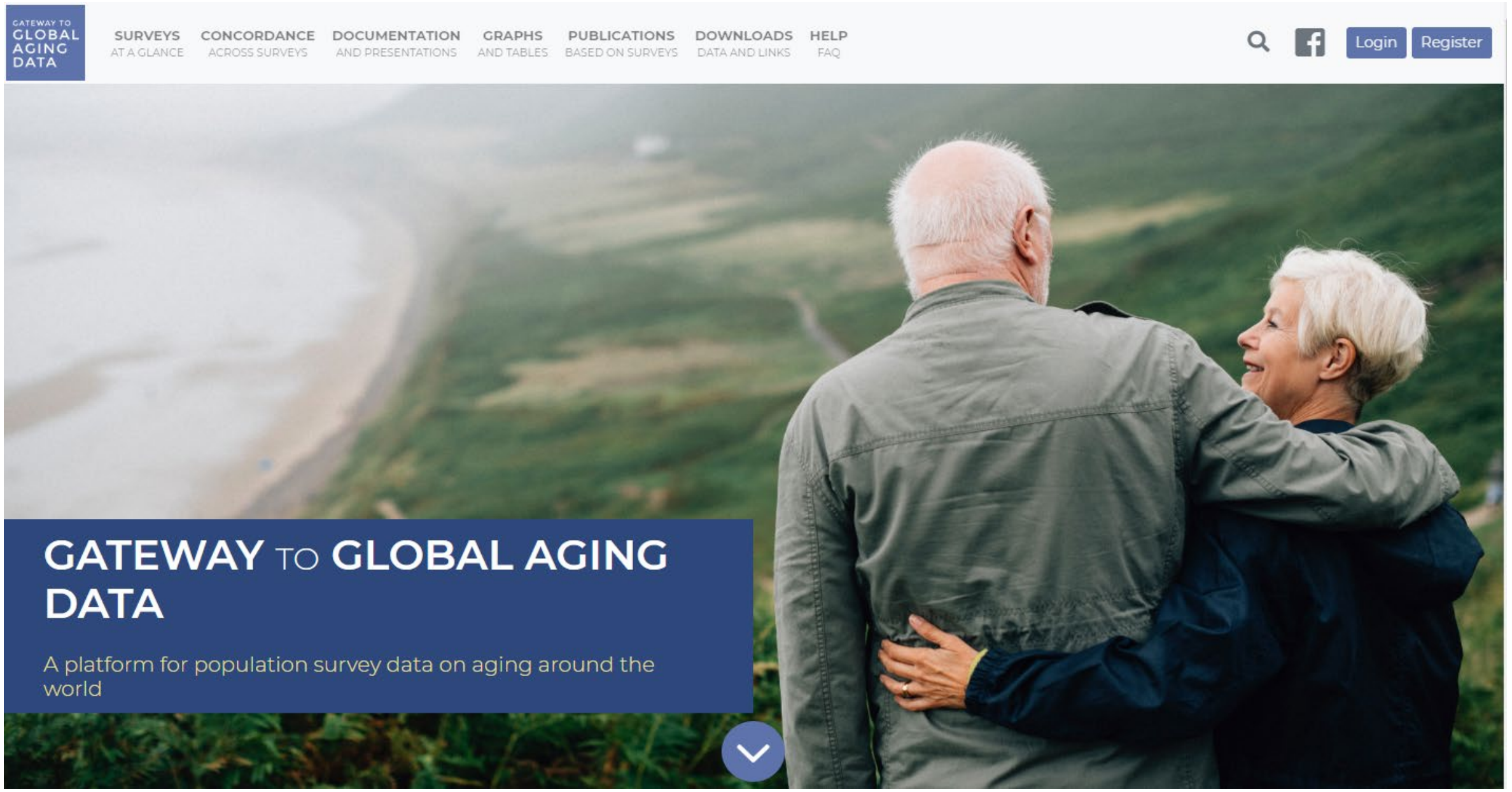
The Gateway is a free public resource designed to facilitate cross-national and longitudinal studies on aging using the family of Health and Retirement Studies around the world.

GATEWAY TO GLOBAL AGING DATA

Health and Retirement Studies Around the World



- United States
- Mexico
- England
- Austria
- Belgium
- Denmark
- France
- Germany
- Greece
- Israel
- Italy
- Netherlands
- Spain
- Sweden
- Switzerland
- Costa Rica
- South Korea
- Czech Republic
- Republic of Ireland
- Poland
- Japan
- Indonesia
- China
- Ghana
- India
- Russia
- South Africa
- Estonia
- Hungary
- Portugal
- Slovenia
- Luxembourg
- Croatia
- Brazil
- Scotland
- Northern Ireland
- Bulgaria
- Cyprus
- Finland
- Latvia
- Lithuania
- Malta
- Romania
- Slovakia
- India


GATEWAY TO GLOBAL AGING DATA
SURVEYS AT A GLANCE
CONCORDANCE ACROSS SURVEYS
DOCUMENTATION AND PRESENTATIONS
GRAPHS AND TABLES
PUBLICATIONS BASED ON SURVEYS
DOWNLOADS DATA AND LINKS
HELP FAQ
Login
Register
GATEWAY TO GLOBAL AGING DATA
A platform for population survey data on aging around the world

What the Gateway Offers

- Overview of HRS family surveys
  - Survey questionnaires
  - Flow-charts illustrating questionnaire skip patterns
  - Search engine by keyword or by topic
- Cross-study concordance tables of specific survey topics
- In-depth documentation of cross-study comparability
- Harmonized data
- Interactive graphs and tables
  - Survey statistics
  - Contextual data
- Search of publications based on HRS family surveys

Types of Data

- Metadata
  - Survey questionnaire, data collection protocol
  - Para data – time-stamps, interviewer notes
- Microdata
  - Phenotype data – survey data, biological markers
  - Genomic data – polygenic risk scores
- Contextual data
  - Macro-level data – unemployment rates, no of hospital beds
  - Environmental data – pollution, temperature
  - Institutional data – pension, long-term care policies

Basic Principles

- Accuracy
- Transparency
- Ease of use

With the ultimate goal of advancing science by supporting data users

Research Collaborators

- Sara Adar, ScD, MHS
- Jennifer Ailshire, PhD
- Marco Angrisani, PhD
- Tiago Cravooliveira, PhD
- Eileen Crimmins, PhD
- Alexandra Crosswell, PhD
- Christian Deindl, PhD
- Elissa Epel, PhD
- Alden Gross, PhD
- Hideki Hashimoto, MD, DrPH
- Cristian Herrera, PhD
- Peifeng Hu, MD, PhD
- David Knapp, PhD
- Maciej Lis, PhD
- Ana Llenanozal, PhD
- Erik Meijer, PhD
- Giacomo Pasini, PhD
- Luis Rosero-Bixby, PhD, MPH
- Lindsay Ryan, PhD
- Gavin Shaddick, PhD
- Jacqui Smith, PhD
- Andrew Steptoe, DSC, DPhil
- Elina Suzuki, PhD
- Morten Wahrendorf, PhD
- David Weir, PhD
- Rebeca Wong, PhD
- Paola Zaninotto, PhD

Collaborating Institutions

- Harvard University
- Irish Social Science Data Archive, Ireland
- NatCen Social Research, UK
- National Institute of Development Administration, Thailand
- Organization for Economic Cooperation and Development, France
- Peking University, China
- Queens' University Belfast, UK
- RAND Corporation
- Research Institute of Economy, Trade & Industry, Japan
- Trinity College Dublin, Ireland
- UK Data Service, UK
- University College London, UK
- Universita' Ca' Foscari – Venezia, Italy
- University of California, Berkeley
- University of California, Los Angeles
- University of California, San Francisco
- University of Costa Rica, Costa Rica
- University of Duesseldorf, Germany
- University of Malaysia, Malaysia
- University of Michigan
- University of Stirling, UK
- University of Texas, Medical Branch
- University of Tokyo, Japan
- World Health Organization

Basic Principles – Accuracy

- We work closely with study teams to devise a harmonization plan
- All documentation and programming are reviewed through a standardized quality control process

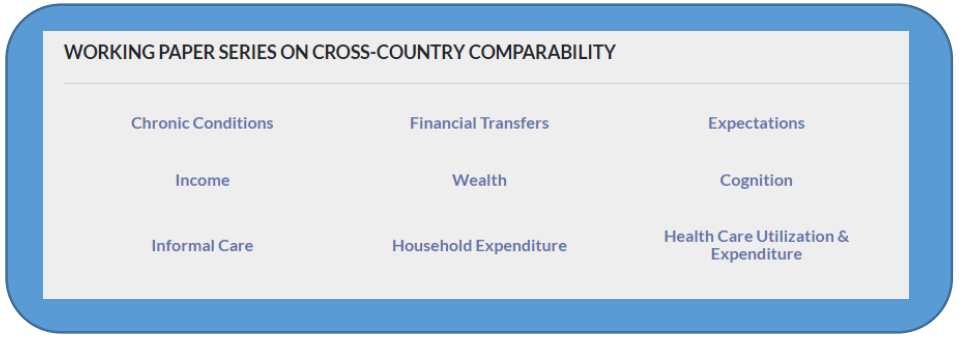
Survey Measure	Study 1	Study 2	Harmonization Rating
Self-rated health	5 point scale, excellent to poor	5 point scale, excellent to poor	Comparable
Smoking quantity	Packs smoked per day	Packs smoked per week	Can be adjusted
Whether any vigorous physical activity	3 or more time per week	More than 10 minutes per week	Cannot be adjusted
Word recall test	10 word recall	8 word recall	Requires statistical calibration

Basic Principles – Transparency

Concordance Tables

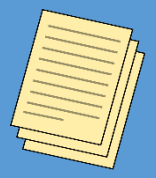
Measure	Study 1	Study 2	Study 3	Study 4
Word recall	Waves 1 -3: 20 words After wave 4: 10 words	All waves: 8 words	All waves: 10 words	All waves: 10 words

User Guides



Codebooks

Extensive Documentation



Programs

Open-source code

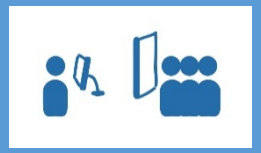
Basic Principles – Ease of Use

Intuitive Search

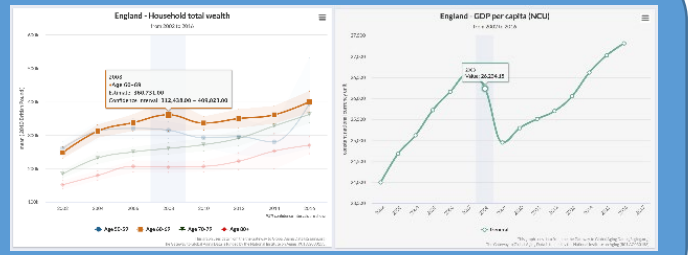
- Study and time specific
- Across studies
- Across repeated observations
- Time point comparisons

Researcher Trainings

Introductory & Advanced webinars



Graphs & Tables



Links and Instructions

Download

Building a Metadata Library

Collecting Survey Metadata

Survey metadata include:

- Data collection protocol
- Order of modules and order of questions within modules
- The location of all survey items inside the interview
- How the measure was collected and from whom
- Question text and interviewer prompts
- Answer types and choices and how the values are formatted
- Interviewer notes, time-stamps, and other para-data

Study 1
- Country: United States
- Sample: Nationally representative for people over age of 50

Baseline
- Time: Fielded in 1992
- Method: In-person interviews

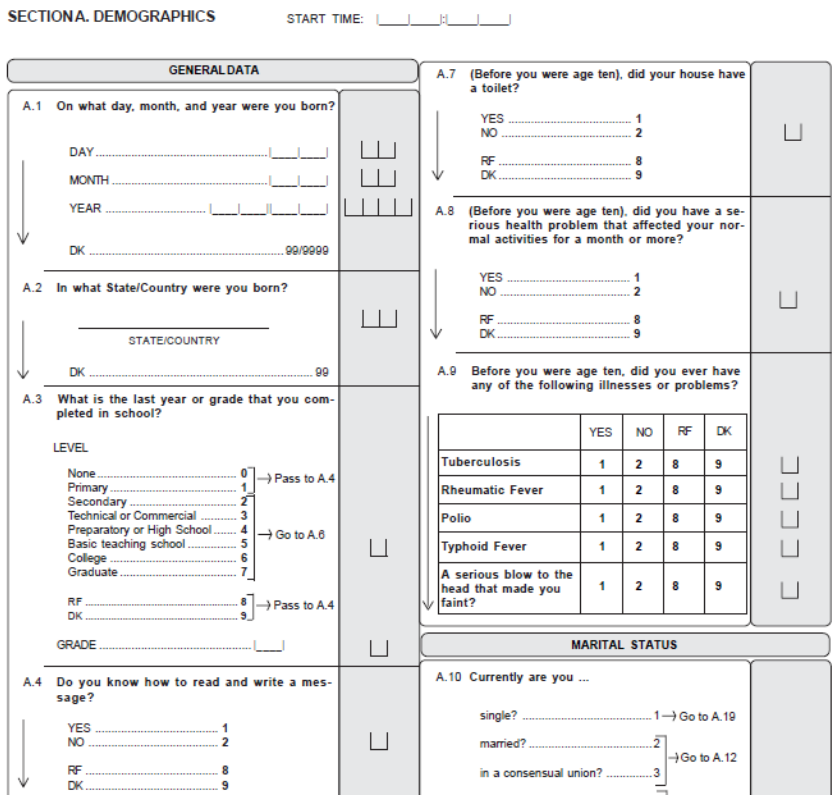
Module 3
- Asked to one person per household

- Beginning time stamp
- Survey measure 10
  - Type
  - Question Text
  - Response Type
  - Answer Choices
  - Answer Value Codes
- Ending time stamp
- Interviewer notes

Collecting Survey Metadata

- Survey metadata can be obtained using:

Optical character recognition (OCR) from paper versions of survey instruments



SECTION A. DEMOGRAPHICS START TIME: |__|__|:|__|__|

GENERAL DATA

A.1 On what day, month, and year were you born?

DAY ..... |__|__| ..... |__|
MONTH ..... |__|__| ..... |__|
YEAR ..... |__|__|__|__| ..... |__|__|__|__|
DK 99/9999

A.2 In what State/Country were you born?

STATE/COUNTRY
DK 99

A.3 What is the last year or grade that you completed in school?

LEVEL
None ..... 0 → Pass to A.4
Primary 1
Secondary 2
Technical or Commercial 3
Preparatory or High School ..... 4 → Go to A.6
Basic teaching school 5
College 6
Graduate 7
RF ..... 8 → Pass to A.4
DK 9
GRADE ..... |__|

A.4 Do you know how to read and write a message?

YES 1
NO 2
RF 8
DK 9

A.7 (Before you were age ten), did your house have a toilet?

YES 1
NO 2
RF 8
DK 9

A.8 (Before you were age ten), did you have a serious health problem that affected your normal activities for a month or more?

YES 1
NO 2
RF 8
DK 9

A.9 Before you were age ten, did you ever have any of the following illnesses or problems?

	YES	NO	RF	DK
Tuberculosis	1	2	8	9
Rheumatic Fever	1	2	8	9
Polio	1	2	8	9
Typhoid Fever	1	2	8	9
A serious blow to the head that made you faint?	1	2	8	9

MARITAL STATUS

A.10 Currently are you ...

single? ..... 1 → Go to A.19
married? 2
in a consensual union? ..... 3 → Go to A.12

Parsing the code of a computer-assisted personal interviewing (CAPI) survey instrument

```
{ This code was generated by a tool on Wednesday, December 04, 2013 at 3:10:19 AM. }
{ Colectica® 4.1.3341 Release }
{ }
{ Changes to this file may cause incorrect behavior and will be lost if the code is regenerated. }


DATAMODEL _2010Instrument "2010 United States Census Questionnaire"

INCLUDE "BBlock1.inc.bla"


TYPE
    TadditionalPeopleCodes =
    (
        C1 (1) "Children, such as newborn babies or foster children",
        C2 (2) "Relatives, such as adult children, cousins, or in-laws",
        C3 (3) "Nonrelatives, such as roommates or live-in baby sitters",
        C4 (4) "People staying here temporarily ...",
        C5 (5) "No additional people"
    )
    ThouseOwnershipCodes =
    (
        C1 (1) "Owned by you or someone in this household with a mortgage or loan? Include home equity loans.",
        C2 (2) "Owned by you or someone in this household free and clear (without a mortgage or loan)?",
        C3 (3) "Rented?",
        C4 (4) "Occupied without payment of rent?"
    )

FIELDS
    { Name: Q1 }
    Q1 "How many people were living or staying in this house, apartment, or mobile home on April 1, 2010?" : Integer

    { Name: Q2 }
    Q2 "Were there any additional people staying here April 1, 2010 that you did not include in Question 1?" : TadditionalPeopleCodes

    { Name: Q3 }
    Q3 "Is this house, apartment, or mobile home" : ThouseOwnershipCodes

    { Name: Q4 }
    Q4 "What is your telephone number? We may call if we don't understand an answer." : String[10]
```

Indexing Survey Metadata

- Collected survey metadata information are indexed into an object-oriented database where all internal components are connected to each other

- This database defines the structure of all survey items

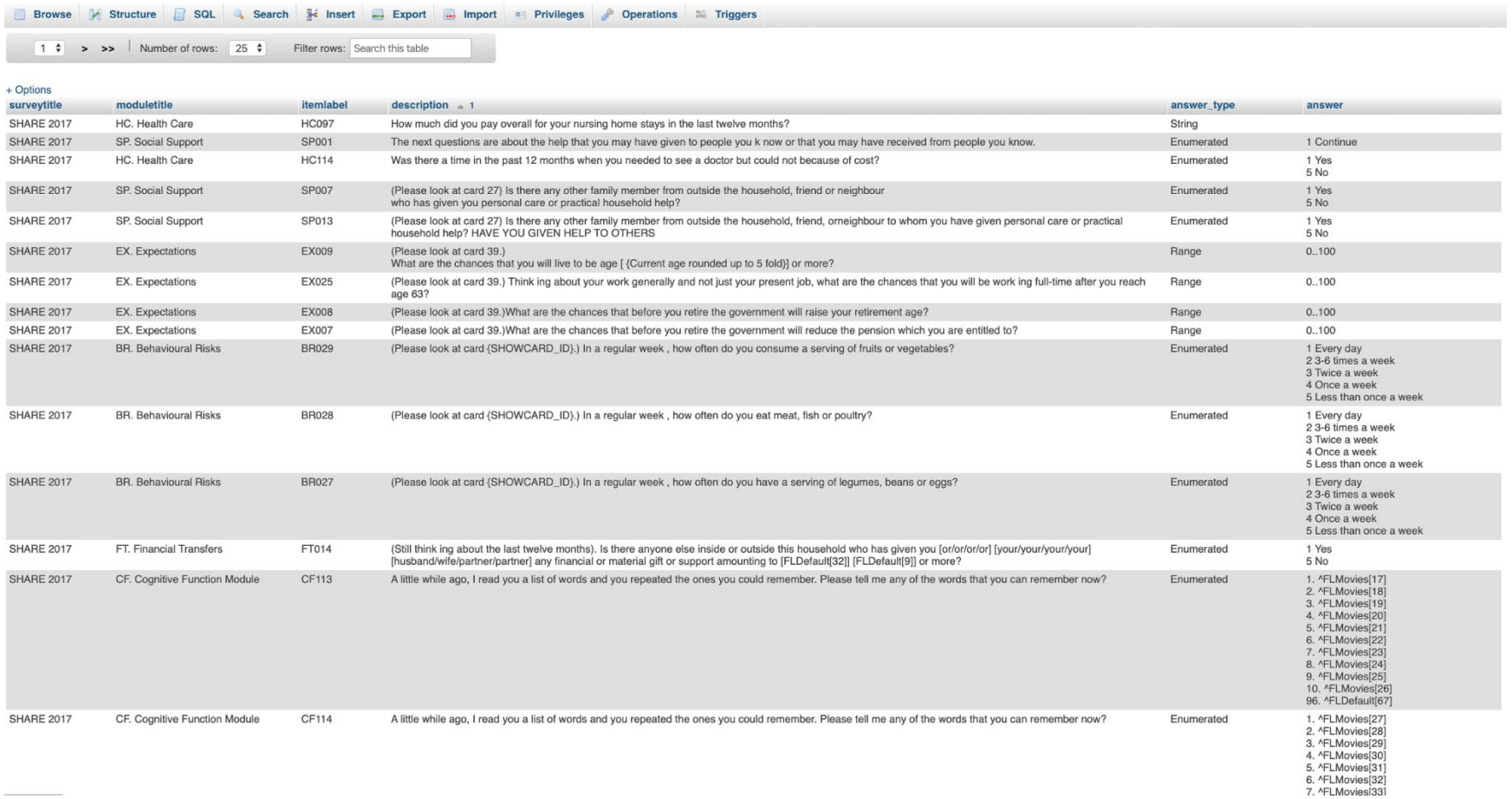
- The database also allows for instant querying by any indexed dimension of the survey metadata

♦Survey measure 1648
- •Part of module 298
  - ▪In survey 83 (conducted in 2012)
    - ★From of study 3

SELECT all survey metadata
TYPE: Question
CONTAINING TEXT: “smoking”
IN THE TIME PERIOD: 2010-2015

Indexing Survey Metadata

Object-oriented database



Browse | Structure | SQL | Search | Insert | Export | Import | Privileges | Operations | Triggers

1 > >> | Number of rows: 25 | Filter rows: Search this table

+ Options

surveytitle	moduletitle	itemlabel	description ▲ 1	answer_type	answer
SHARE 2017	HC. Health Care	HC097	How much did you pay overall for your nursing home stays in the last twelve months?	String	
SHARE 2017	SP. Social Support	SP001	The next questions are about the help that you may have given to people you k now or that you may have received from people you know.	Enumerated	1 Continue
SHARE 2017	HC. Health Care	HC114	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?	Enumerated	1 Yes 5 No
SHARE 2017	SP. Social Support	SP007	(Please look at card 27) Is there any other family member from outside the household, friend or neighbour who has given you personal care or practical household help?	Enumerated	1 Yes 5 No
SHARE 2017	SP. Social Support	SP013	(Please look at card 27) Is there any other family member from outside the household, friend, orneighbour to whom you have given personal care or practical household help? HAVE YOU GIVEN HELP TO OTHERS	Enumerated	1 Yes 5 No
SHARE 2017	EX. Expectations	EX009	(Please look at card 39.) What are the chances that you will live to be age [{Current age rounded up to 5 fold}] or more?	Range	0..100
SHARE 2017	EX. Expectations	EX025	(Please look at card 39.) Think ing about your work generally and not just your present job, what are the chances that you will be work ing full-time after you reach age 63?	Range	0..100
SHARE 2017	EX. Expectations	EX008	(Please look at card 39.)What are the chances that before you retire the government will raise your retirement age?	Range	0..100
SHARE 2017	EX. Expectations	EX007	(Please look at card 39.)What are the chances that before you retire the government will reduce the pension which you are entitled to?	Range	0..100
SHARE 2017	BR. Behavioural Risks	BR029	(Please look at card {SHOWCARD_ID}.) In a regular week , how often do you consume a serving of fruits or vegetables?	Enumerated	1 Every day 2 3-6 times a week 3 Twice a week 4 Once a week 5 Less than once a week
SHARE 2017	BR. Behavioural Risks	BR028	(Please look at card {SHOWCARD_ID}.) In a regular week , how often do you eat meat, fish or poultry?	Enumerated	1 Every day 2 3-6 times a week 3 Twice a week 4 Once a week 5 Less than once a week
SHARE 2017	BR. Behavioural Risks	BR027	(Please look at card {SHOWCARD_ID}.) In a regular week , how often do you have a serving of legumes, beans or eggs?	Enumerated	1 Every day 2 3-6 times a week 3 Twice a week 4 Once a week 5 Less than once a week
SHARE 2017	FT. Financial Transfers	FT014	(Still think ing about the last twelve months). Is there anyone else inside or outside this household who has given you [or/or/or/or] [your/your/your/your] [husband/wife/partner/partner] any financial or material gift or support amounting to [FLDefault[32]] [FLDefault[9]] or more?	Enumerated	1 Yes 5 No
SHARE 2017	CF. Cognitive Function Module	CF113	A little while ago, I read you a list of words and you repeated the ones you could remember. Please tell me any of the words that you can remember now?	Enumerated	1. ^FLMovies[17] 2. ^FLMovies[18] 3. ^FLMovies[19] 4. ^FLMovies[20] 5. ^FLMovies[21] 6. ^FLMovies[22] 7. ^FLMovies[23] 8. ^FLMovies[24] 9. ^FLMovies[25] 10. ^FLMovies[26] 96. ^FLDefault[67]
SHARE 2017	CF. Cognitive Function Module	CF114	A little while ago, I read you a list of words and you repeated the ones you could remember. Please tell me any of the words that you can remember now?	Enumerated	1. ^FLMovies[27] 2. ^FLMovies[28] 3. ^FLMovies[29] 4. ^FLMovies[30] 5. ^FLMovies[31] 6. ^FLMovies[32] 7. ^FLMovies[33]

Constructing a Browse/Search Tool

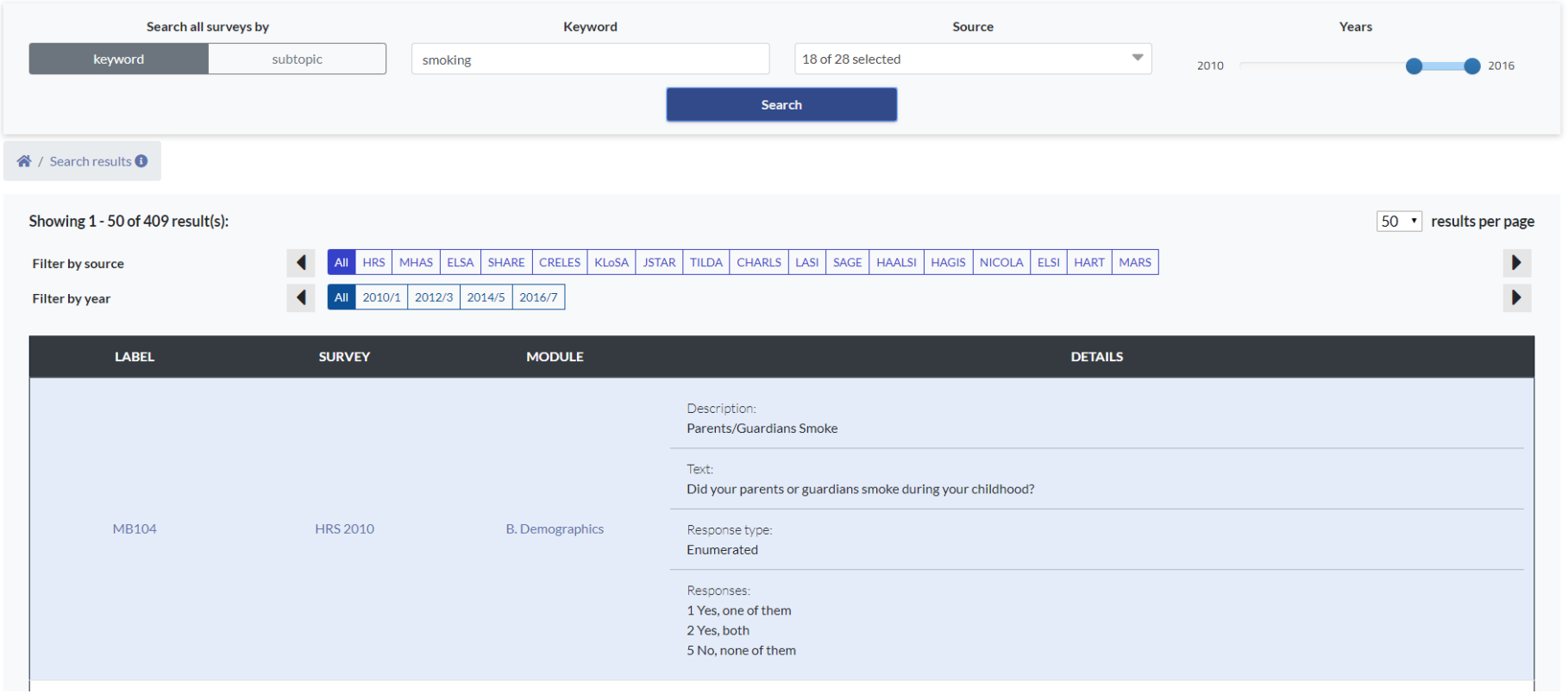
Researchers should be able to easily browse through the survey metadata while understanding the study, timing, and type of the survey

Study Overview	Core Interview	End of Life Interview	Life History	Health Assessment	Self-Completion

	HRS	MHAS	ELSA	SHARE	CRELES	KLoSA	JSTAR	TILDA	CHARLS	LASI
	United States	Mexico	England	20+ European Countries and Israel	Costa Rica	Korea	Japan	Ireland	China	India
1992-93	HRS W1 AHEAD 1993 W1									
1994-95	HRS W2 AHEAD 1995 W2									
1996-97	HRS W3									
1998-99	HRS W4									
2000-01	HRS W5	MHAS W1								
2002-03	HRS W6	MHAS W2	ELSA W1							
2004-05	HRS W7		ELSA W2	SHARE W1	CRELES W1					
2006-07	HRS W8		ELSA W3	SHARE W2	CRELES W2	KLoSA W1	JSTAR W1			
2008-09	HRS W9		ELSA W4		CRELES W3	KLoSA W2	JSTAR W2			
2010-11	HRS W10		ELSA W5	SHARE W4	CRELES W4	KLoSA W3	JSTAR W3	TILDA W1	CHARLS W1	
2012-13	HRS W11	MHAS W3	ELSA W6	SHARE W5	CRELES W5	KLoSA W4	JSTAR W4	TILDA W2	CHARLS W2	
2014-15	HRS W12 UAS HRS W1	MHAS W4	ELSA W7	SHARE W6		KLoSA W5		TILDA W3	CHARLS W4	
2016-17	HRS W13		ELSA W8	SHARE W7		KLoSA W6		TILDA W4		LASI W1

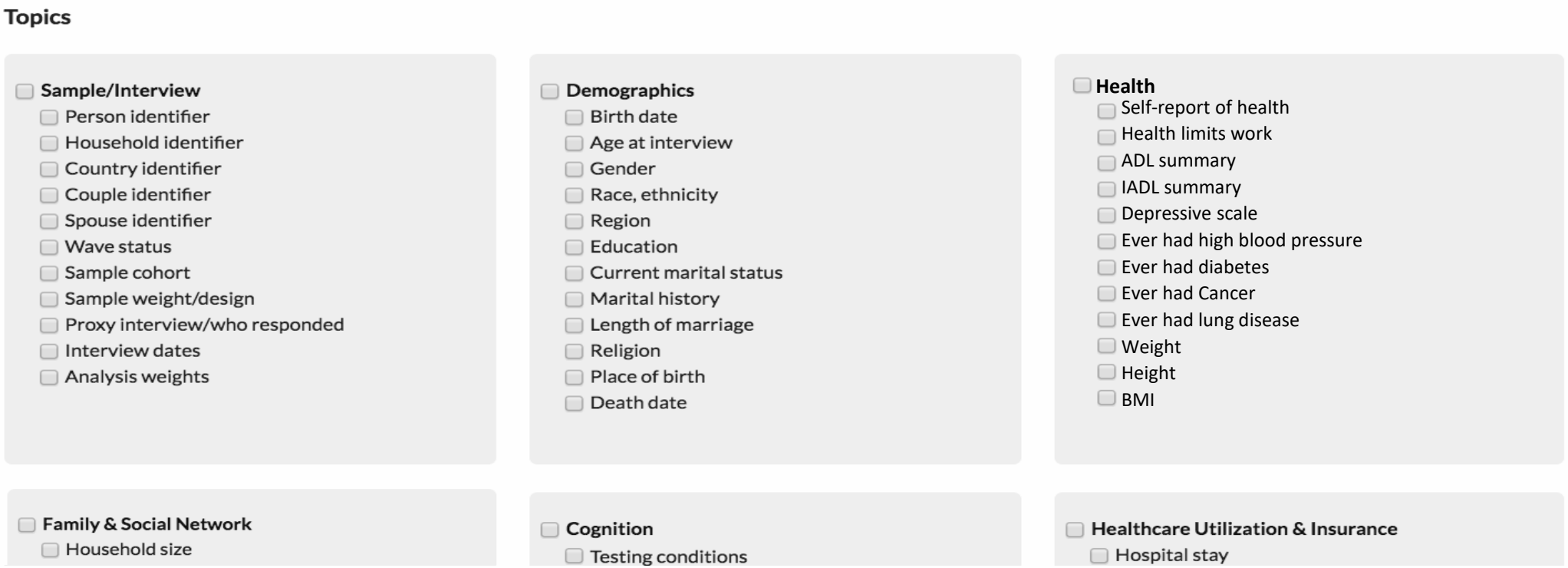
Constructing a Browse/Search Tool

Researchers should also be able to intuitively and quickly search through survey metadata to find survey items which are important for their research



Constructing a Browse/Search Tool

It is also possible to identify survey items specific to common research topics and domains from which users can see all relevant survey items



Harmonizing Phenotype Data

Data Harmonization

1. Pre-statistical harmonization
2. Missing data and imputation
3. Statistical harmonization based on Item Response Theory

Pre-statistical Harmonization

Harmonized variables are research-ready variables which easily allow researchers to conduct analysis by pooling data from multiple waves of a survey or from multiple surveys

- Variables are defined identically across all waves and surveys
- Each dataset combines all available waves from each study; each individual is one record

Study 1

Person ID	R1AGE	R2AGE	R3AGE
10000	52	54	56
10001	59	61	63

Study 2

Person ID	R1AGE	R2AGE	R3AGE
1000	71	73	75
1001	60	62	64

Pre-statistical Harmonization

Harmonized variable names ensure ease of use and transparency

- All variables use intuitive variable names, e.g. r1work – whether the respondent is currently working in wave 1
- Study specific variable names are used to indicate significant inter-study differences: e.g. RwVGACT_C – whether the respondent does any vigorous physical with different question wording

Study 1

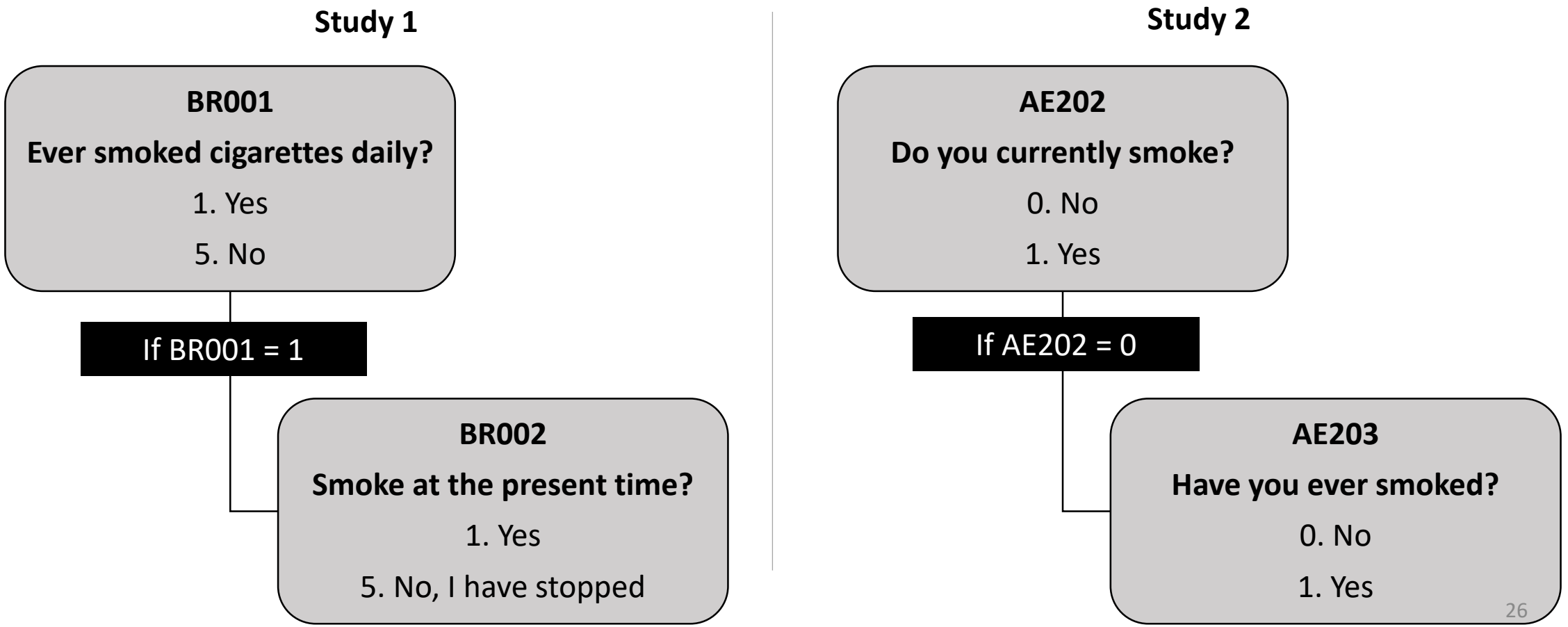
Person ID	R1VGACT	R2VGACT	R3VGACT
10000	1. Yes	0. No	1. Yes
10001	1. Yes	1. Yes	1. Yes

Study 2

Person ID	R1VGACT_C	R2VGACT_C	R3VGACT_C
1000	0. No	0. No	0.No
1001	1. Yes	1. Yes	0.No

Building Harmonized Variables

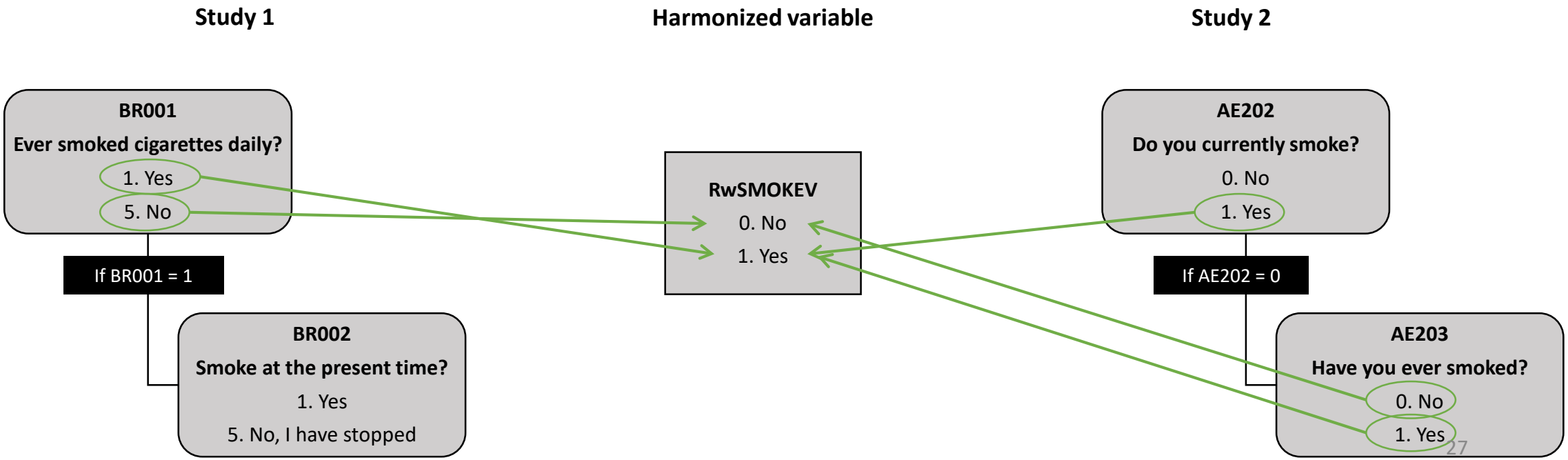
Harmonized variables have been built to account for any survey skip pattern



Building Harmonized Variables

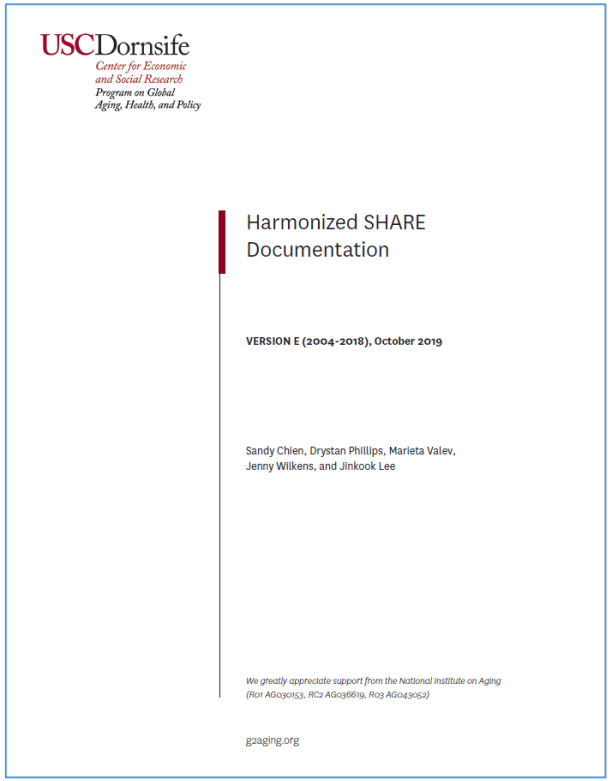
Harmonized variable name: RwSMOKEV
Harmonized variable label: Whether the respondent ever smoked
Harmonized variable codes:

- 0. No, the respondent has never smoked
- 1. Yes, the respondent has smoked



Documenting Harmonized Variables

Each harmonized dataset is accompanied by its own codebook.


USCDornsife
Center for Economic and Social Research
Program on Global Aging, Health, and Policy

Harmonized SHARE Documentation

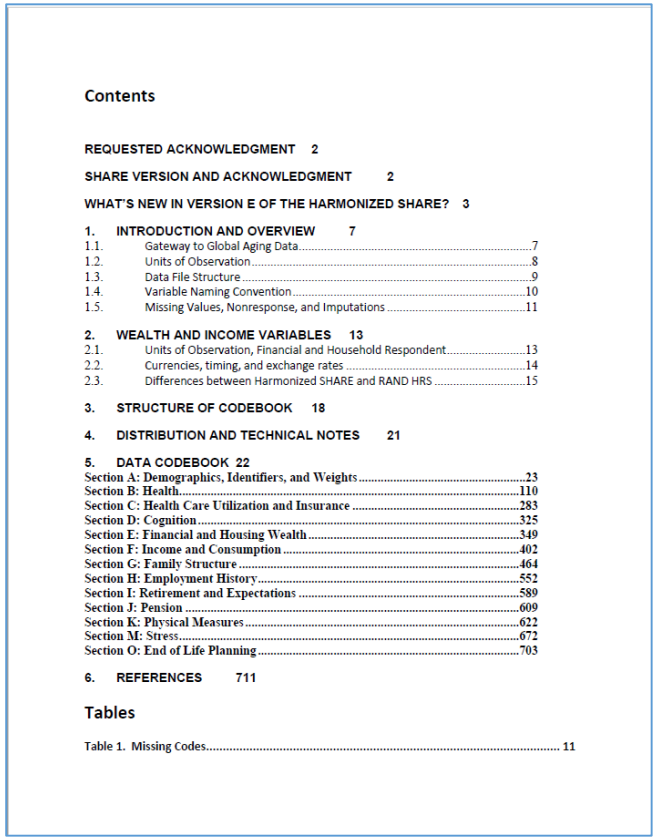
VERSION E (2004-2018), October 2019

Sandy Chien, Drystan Phillips, Marieta Valev, Jenny Wilkens, and Jinkook Lee

We greatly appreciate support from the National Institute on Aging (R01 AG030153, RC2 AG036619, R03 AG043052)

g2aging.org

- Introduces the harmonization project and study
- Overviews survey timing, survey design, and sampling framework
- Discusses weighting and imputation
- Details specifics of harmonization process
- Divides variables into sections based on research domain


Contents

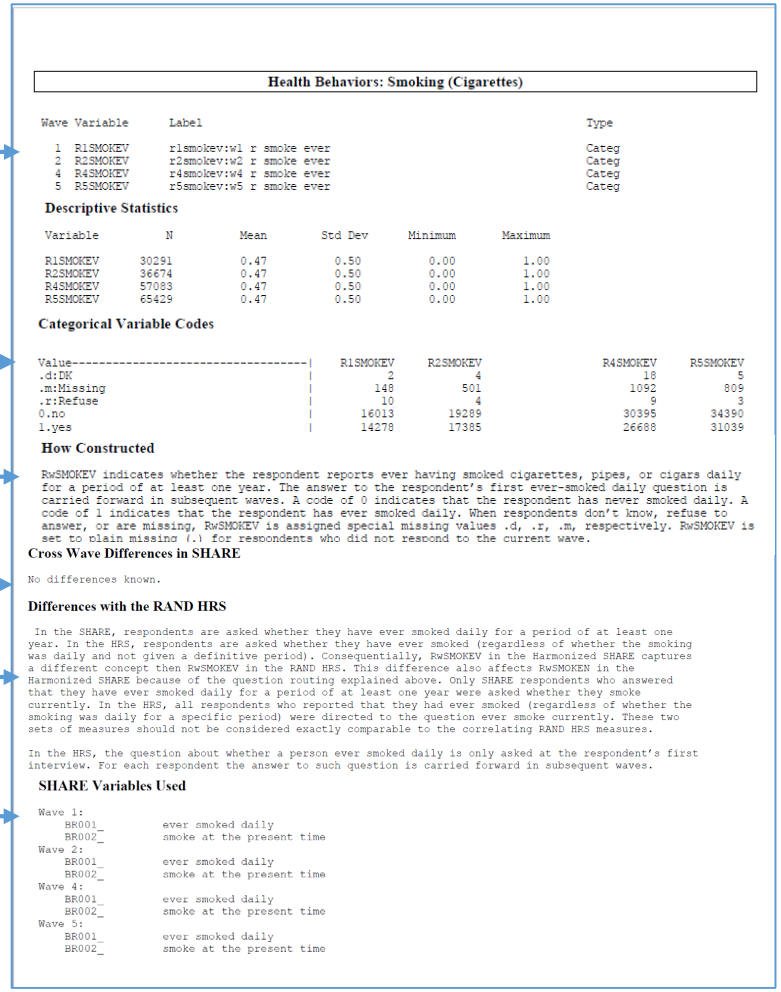
REQUESTED ACKNOWLEDGMENT 2
SHARE VERSION AND ACKNOWLEDGMENT 2
WHAT'S NEW IN VERSION E OF THE HARMONIZED SHARE? 3
1. INTRODUCTION AND OVERVIEW 7
1.1. Gateway to Global Aging Data ... 7
1.2. Units of Observation ... 8
1.3. Data File Structure ... 9
1.4. Variable Naming Convention ... 10
1.5. Missing Values, Nonresponse, and Imputations ... 11
2. WEALTH AND INCOME VARIABLES 13
2.1. Units of Observation, Financial and Household Respondent ... 13
2.2. Currencies, timing, and exchange rates ... 14
2.3. Differences between Harmonized SHARE and RAND HRS ... 15
3. STRUCTURE OF CODEBOOK 18
4. DISTRIBUTION AND TECHNICAL NOTES 21
5. DATA CODEBOOK 22
Section A: Demographics, Identifiers, and Weights ... 23
Section B: Health ... 110
Section C: Health Care Utilization and Insurance ... 283
Section D: Cognition ... 325
Section E: Financial and Housing Wealth ... 349
Section F: Income and Consumption ... 402
Section G: Family Structure ... 464
Section H: Employment History ... 552
Section I: Retirement and Expectations ... 589
Section J: Pension ... 609
Section K: Physical Measures ... 622
Section M: Stress ... 672
Section O: End of Life Planning ... 703
6. REFERENCES 711

Tables

Table 1. Missing Codes ... 11

Documenting Harmonized Variables

- Summary statistics for each set of variables
- Tabulations of all coded values
- Details about variable creation and any assumptions made in the creation
- Highlights of any intra-study differences
- Highlights of any inter-study differences
- List of all the variables from the originating dataset used in the creation of the harmonized variable



Health Behaviors: Smoking (Cigarettes)

Wave	Variable	Label	Type
1	R1SMOKEV	r1smokev:w1 r smoke ever	Categ
2	R2SMOKEV	r2smokev:w2 r smoke ever	Categ
4	R4SMOKEV	r4smokev:w4 r smoke ever	Categ
5	R5SMOKEV	r5smokev:w5 r smoke ever	Categ

Descriptive Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
R1SMOKEV	30291	0.47	0.50	0.00	1.00
R2SMOKEV	36674	0.47	0.50	0.00	1.00
R4SMOKEV	57083	0.47	0.50	0.00	1.00
R5SMOKEV	65429	0.47	0.50	0.00	1.00

Categorical Variable Codes

Value	R1SMOKEV	R2SMOKEV	R4SMOKEV	R5SMOKEV
.d:DK	2	4	18	5
.m:Missing	148	501	1092	809
.r:Refuse	10	4	9	3
0.no	16013	19289	30395	34390
1.yes	14278	17385	26688	31039

How Constructed

RwSMOKEV indicates whether the respondent reports ever having smoked cigarettes, pipes, or cigars daily for a period of at least one year. The answer to the respondent's first ever-smoked daily question is carried forward in subsequent waves. A code of 0 indicates that the respondent has never smoked daily. A code of 1 indicates that the respondent has ever smoked daily. When respondents don't know, refuse to answer, or are missing, RwSMOKEV is assigned special missing values .d, .r, .m, respectively. RwSMOKEV is set to plain missing (.) for respondents who did not respond to the current wave.

Cross Wave Differences in SHARE

No differences known.

Differences with the RAND HRS

In the SHARE, respondents are asked whether they have ever smoked daily for a period of at least one year. In the HRS, respondents are asked whether they have ever smoked (regardless of whether the smoking was daily and not given a definitive period). Consequentially, RwSMOKEV in the Harmonized SHARE captures a different concept then RwSMOKEV in the RAND HRS. This difference also affects RwSMOKEN in the Harmonized SHARE because of the question routing explained above. Only SHARE respondents who answered that they have ever smoked daily for a period of at least one year were asked whether they smoke currently. In the HRS, all respondents who reported that they had ever smoked (regardless of whether the smoking was daily for a specific period) were directed to the question ever smoke currently. These two sets of measures should not be considered exactly comparable to the correlating RAND HRS measures.

In the HRS, the question about whether a person ever smoked daily is only asked at the respondent's first interview. For each respondent the answer to such question is carried forward in subsequent waves.

SHARE Variables Used

Wave 1:		
	BR001_	ever smoked daily
	BR002_	smoke at the present time
Wave 2:		
	BR001_	ever smoked daily
	BR002_	smoke at the present time
Wave 4:		
	BR001_	ever smoked daily
	BR002_	smoke at the present time
Wave 5:		
	BR001_	ever smoked daily
	BR002_	smoke at the present time

Performing Statistical Harmonization to Address Item-level Missingness

- Complete case analysis is the default for most statistical software
- Excluding observations with missing values is potentially a huge loss of information (large standard errors; imprecise estimators)
- Excluding observations with missing values also potentially introduces large biases (people with lower levels of cognition are more likely to have missing values in cognitive tests)

Person ID	Gender	Age	Income Level	Education	Marital Status	Word Recall Score	INCLUSION
1	Male	70	High	HS graduate	Married	8	YES
2	Female	55	High	College grad	Married	10	YES
3	Male	89	Low	HS graduate	Widowed	.	NO

Performing Statistical Harmonization to Address Item-level Missingness

- Imputed values replaces missing values with draws from their (conditional) distribution, and thereby creates a complete data set for researchers
- Imputation is often economically efficient and scientifically preferable because it is done only once (individual researchers do not have to do it, and all researchers work with the same data)

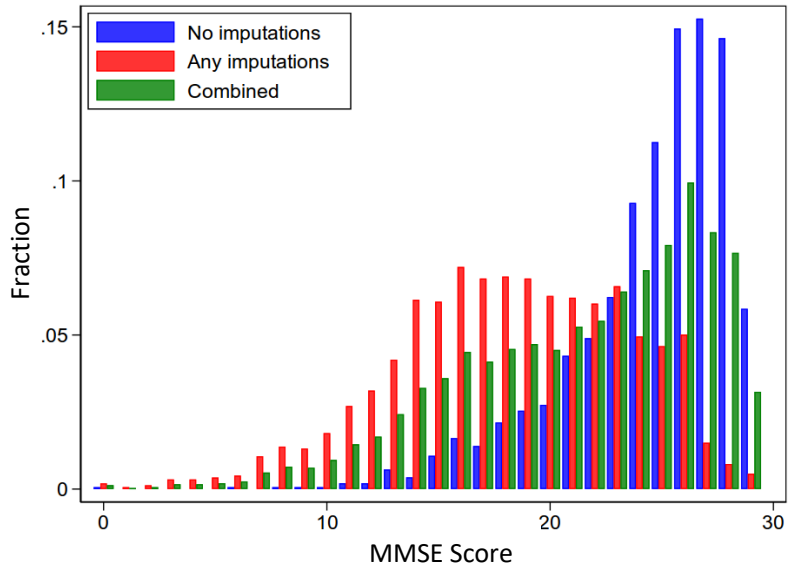
Person ID	Gender	Age	Income Level	Education	Marital Status	Word Recall Score	Word Recall Score w/ Imputations	INCLUSION
1	Male	70	High	HS graduate	Married	8	8	YES
2	Female	55	High	College grad	Married	10	10	YES
3	Male	89	Low	HS graduate	Widowed	.	3	YES

Performing Statistical Harmonization to Address Item-level Missingness

Harmonized imputation method

- Estimates the joint distribution of the variables in each dataset in the presence of missing data
- Assigns “don’t know” values for many cognitive tests as 0 values
- Estimates a regression model which specifies the conditional distribution of the variable to be imputed as a function of the regressors.
- Imputes missing values using pseudo-random draws from the conditional distribution

Person	Gender	Age	Value	Value w/ Imputations
1	Male	60	20	20
2	Male	60	.	20
3	Male	72	10	10



Performing Statistical Harmonization to Address Item-level Missingness

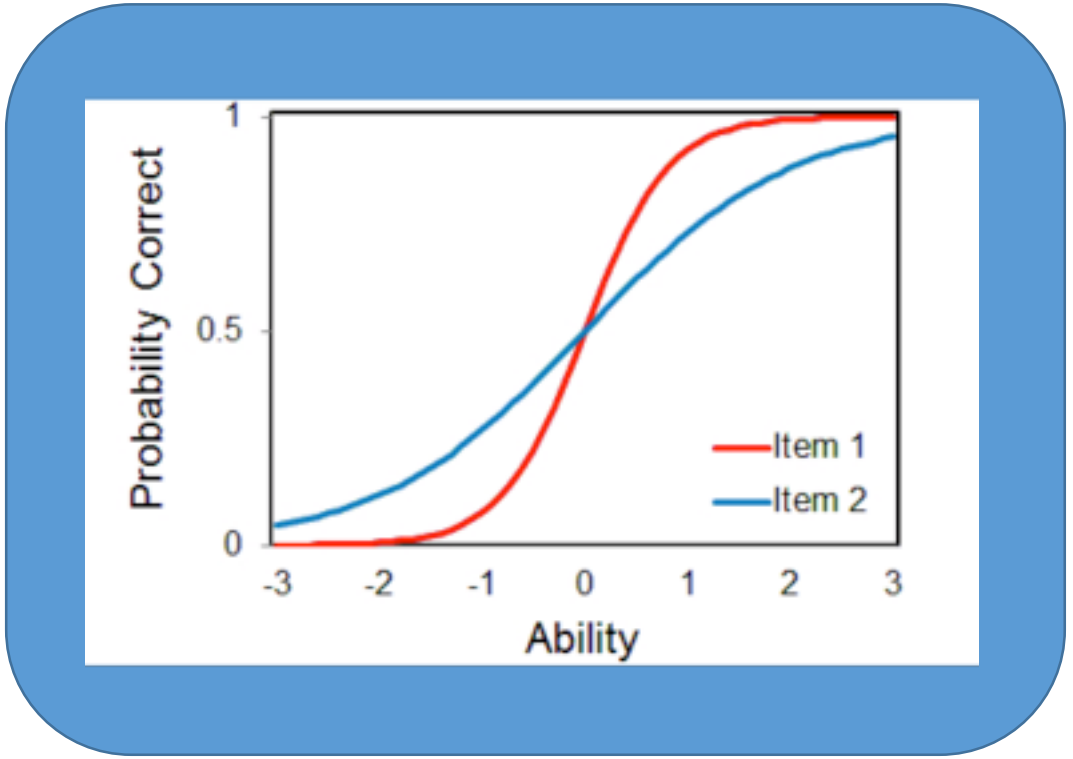
Researchers are provided with indicators of whether values were imputed and what level of information was used in the imputation procedure

Value	R1FMO
0. Not imputed	2833
1. DK – assigned 0 value	273
2. Not assessed – imputed	91
3. Refused – imputed	25
4. Missing - imputed	2

Developing Harmonized Domain-Specific Models using Item Response Theory (IRT)

Item Response Theory

- A general approach to data analysis relating responses to underlying traits
- Many related statistical models
- Broadly contained within general latent variable framework
- Developed in fields of Educational and Psychological Assessment (1930-1970's)
- Continually refined methods
- Broad applications in social and health sciences



Developing Harmonized Domain-Specific Models using Item Response Theory (IRT)

Working with psychometricians we develop harmonized domain-specific models

- Classify items according to measurement
  - Domain
  - Modality (performance, self-rated, informant-rated, expert rating)
  - Response type

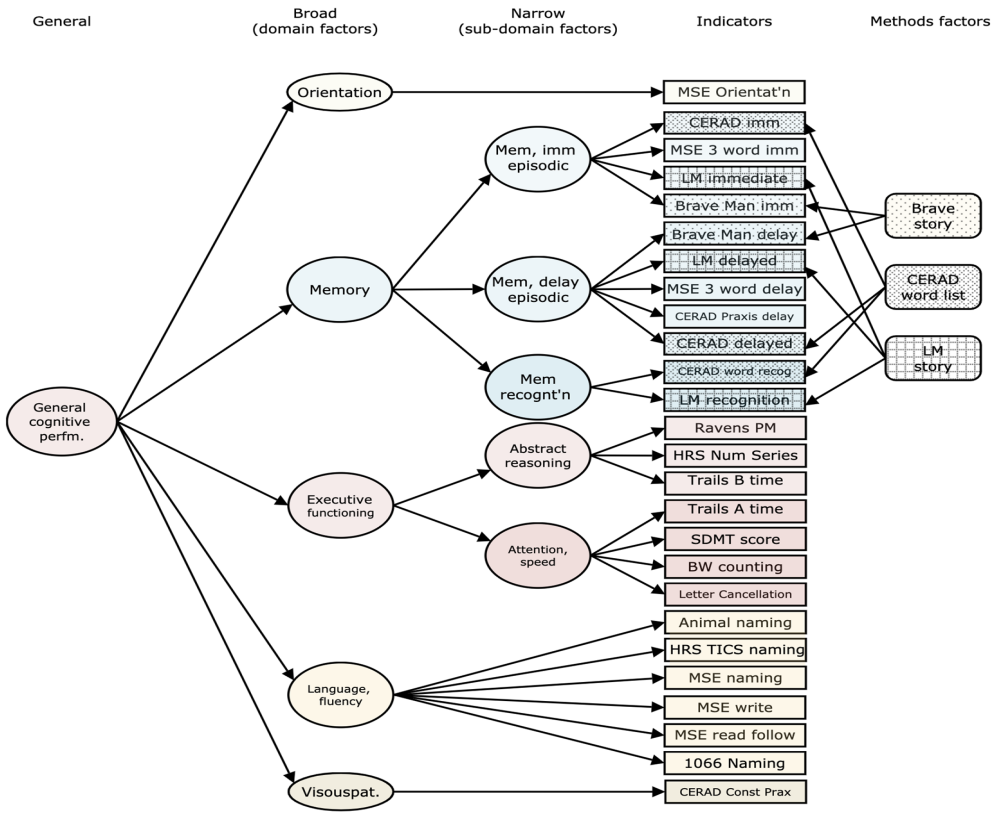
Date Naming
Domain: Time and Place Orientation
Modality: Performance
Response Type: Boolean

- Identify reference population and sample(s)

Reference population:
Clinician diagnosed individuals with moderate and severe cognitive impairment as part of study ABC

Developing Harmonized Domain-Specific Models using Item Response Theory (IRT)

- Assess cross-validation of the measurement model using IRT



Model	Descriptor
Single domain models	
Orientation	Good
Memory-Episodic-Immediate	Perfect
Memory-Episodic-Delayed	Good
Memory-Episodic-Recognition	Perfect
EF-Abstract-Reasoning	Perfect
EF-Attention-Speed	Good
Language	Good
Visuospatial	Perfect

Developing Harmonized Domain-Specific Models using Item Response Theory (IRT)

- Rank and/or classify individuals

Episodic Memory Impairment Cut Points
None: >=16
Mild: 13-15
Serious: 11-12
Severe: <= 10

- Co-calibrate across studies and time

Study	**Score Equivalence**							
Study 1	1	2	3	4	5	6	7	8
Study 2	1		2		3	4	5	6	7	8
Study 3	1	2	3		4		5	6	7	8

Thank You

National Institute on Aging, NIH

(R01 AG030153, RC2 AG036619, R03 AG043052, R24 AG048024)