LIS
Technical Working Paper

No. 12

Harmonization and Quality Assurance of Income and Wealth Data: The Case of LIS

Jörg Neugschwender, Teresa Munzi, Piotr R. Paradowski

March 2024

A revised version of this paper has been published as:
"Harmonization and Quality Assurance of Income and Wealth Data: The Case of LIS",
in Survey Data Harmonization in the Social Sciences, edited by Irina Tomescu-Dubrow, Christof Wolf, Kazimierz M. Slomczynski, and J. Craig Jenkins, Chapter 15, pp. 269-284.
Wiley, 2024. https://doi.org/10.1002/9781119712206.ch15

Luxembourg Income Study (LIS), asbl

Harmonization and Quality Assurance of Income and Wealth Data: The Case of LIS

Jörg Neugschwender, LIS

Teresa Munzi, LIS

Piotr R. Paradowski, LIS & Gdansk University of Technology

Comparability of concepts in survey data harmonization is essential for scientific analyses. LIS - also known as the Luxembourg Income Study or LIS Cross-National Data Center in Luxembourg – acquires and harmonizes income and wealth microdata to provide the scientific community with a comparable database that is unique in the world in its growing temporal and geographic breadth. Over many decades, scholars worldwide have used the Luxembourg Income Study (LIS) and Luxembourg Wealth Study (LWS) databases to compare economic and social policies and their effects on outcomes, including poverty, income inequality, employment, gender inequality, and wealth portfolios. Since source data entering LIS differ substantially in terms of collection mode, type of information collected, level of detail, and structure of the data, this chapter elaborates on the various harmonization efforts at LIS, revolving around the ex-post aspect of harmonization. The discussion of core challenges of ex-post harmonization, LIS guiding principle of operational comparability, documentation, software tools, and quality assurance procedures set in place at LIS are enriched with practical examples. The last section concludes with key lessons learned from nearly 40 years of harmonization of microdata, pointing out to other significant factors, such as the importance of interaction with scholars, data providers, and other experts in the field in order to provide reliable data for cross-national interdisciplinary research.

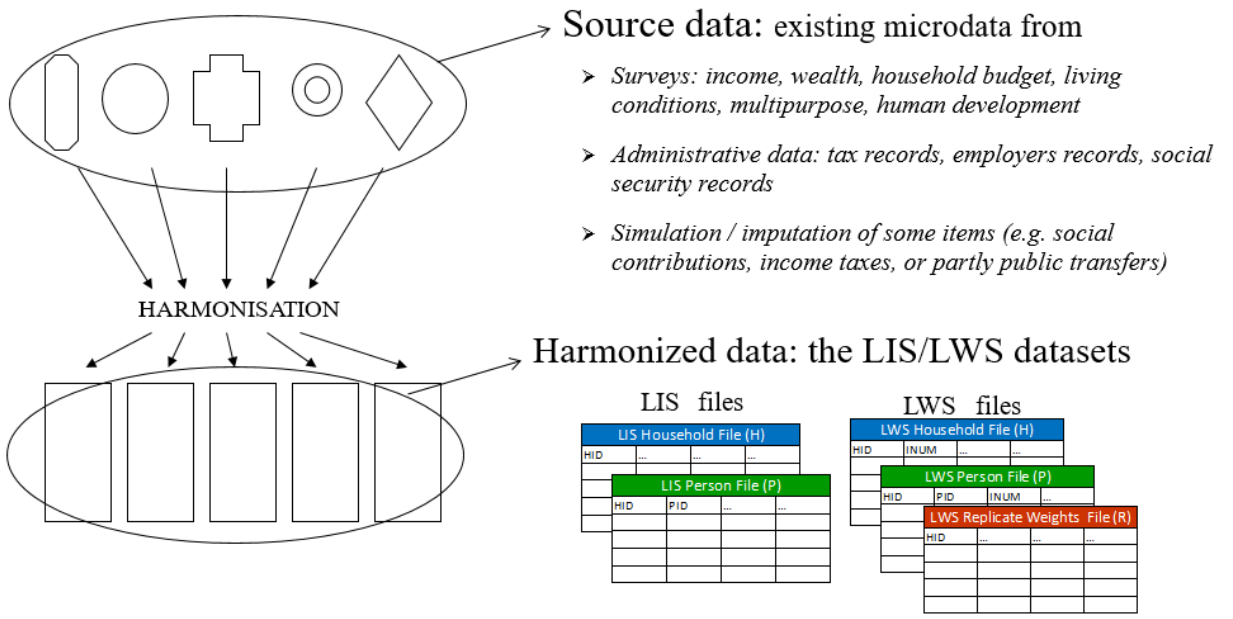
Keywords: income, wealth, microdata, ex-post harmonization, comparability

1. Introduction

LIS - also known as the *Luxembourg Income Study* or *LIS Cross-National Data Center in Luxembourg* – is an independent, non-profit cross-national microdata archive and research institute. One of the main reasons behind launching LIS is providing researchers with microdata on income (and later wealth) that facilitates comparative analyses across countries and time. Initiated at an international conference in Luxembourg in 1982, which saw the gathering of experienced researchers from various countries, the idea – rather innovative for the time – developed that "it would be possible to pool the knowledge and experience in these countries to create internally and externally consistent data sets for comparative studies which are far superior to those currently in existence." (Smeeding *et al*., 1985). The main drivers for establishing the *Luxembourg Income Project* were relatively simple. Which welfare state policies work to protect risk groups, e.g., which public and private transfers provide income in old age? Even more importantly, what is needed to compare well income data across countries? Thus, LIS was founded based on the need to carry out conceptual harmonization that allowed the subsequent analyses to compare similar concepts, an objective that is until today a crucial element of LIS' day-to-day work.

Source data entering LIS differ substantially in terms of collection mode (surveys vs. administrative data), type of information collected, level of detail, and structure of the data (Figure 1). Through the harmonization process, the *ex-post* harmonized data are joined together in a common repository, where each set of data (or *dataset*, i.e. a group of individuals and households representing the total population in one year for one country) has the same structure and each variable comparable contents. Researchers are able to access all datasets *via* a remote access system.

Figure 1. *Ex-post* harmonization at LIS



The harmonized LIS datasets are stored in two *databases,* the *Luxembourg Income Study (LIS) Database*, which includes income data, and the *Luxembourg Wealth Study (LWS) Database*, which focuses on wealth data. The LIS data target all individuals in a household unit (typically defined as a single person or a group of persons living in one dwelling and sharing a budget). Each dataset contains household- and person-level data on labor income, capital income, pensions, public social security benefits (excl. pensions), private transfers, taxes and contributions and other non-consumption expenditures, and consumption expenditures, as well as socio-demographic and labor market information. Besides the identical blocks of information from *LIS* datasets, the wealth datasets additionally provide information on assets and liabilities, contingent assets and liabilities, assets acquired in the past, and behavioral variables.

Concerning the frequency of microdata, LIS shortened over time the interval between two periods; data in the early years were provided every five years (until 2000), then every four years (from 2000-2004), and then every three years (from 2004-2016). Since 2019, various data are acquired annually, and country series are retrospectively extended following an annual frequency when feasible. As a result, LIS stores a quickly growing large pool of data, which at the time of writing consisted of around 600 datasets harmonized and documented for the scientific community, covering nearly 60 countries, spanning over 50 years.

With this immense pool of harmonized data, nowadays, researchers can easily analyze the data in a cross-national perspective, reflecting the initial idea since launching LIS. Scholars worldwide have used LIS datasets to compare economic and social policies and their effects on outcomes, including poverty, income inequality, employment status, wage patterns, gender inequality, family formation, child-wellbeing, health status, and immigration. In addition, the newer LWS datasets enable comparative research on wealth portfolios, assets and liabilities, the relation between household income and wealth, and economic behavior.

The following sections shed particular light on the various harmonization efforts (Section 2) and quality assurance procedures (Section 3) at LIS, exemplified with practical examples. Section 4 focuses on core challenges of *ex-post* harmonization. Section 5 describes quality assurance from a technical point of view. The last section concludes with key lessons learned from nearly 40 years of harmonization of microdata.

2. Applied harmonization methods

The *Luxembourg Income Study project* has been initiated by three main ideas: i) take pre-existing microdata, ii) establish a common conceptual framework for harmonized data, and iii) provide comparable microdata files commonly accessible for researchers (Smeeding *et al*., 1985). These motives determine the scope of LIS' work. This section presents the main workflow of harmonization at LIS, which is defined by i) acquisition of the data, ii) the harmonization process, and iii) data quality assurance. Section 2 focuses on the first two points; the latter is described in Section 3.

The central aspect of LIS' work revolves around the *ex-post* feature of harmonization: LIS acquires pre-existing survey and administrative data in their original shape. At the same time, LIS applies a strict policy to accept data for harmonization. There are various minimum entry conditions:

- The source data must contain income and wealth items that ensure a representative and comparable situation of well-being.
- The datasets need to contain the correct micro-level detail, i.e. information needs to be collected at the household and individual level.
- The unit of analysis should be the household; only in rare circumstances, LIS accepts alternative definitions, such as the tax unit, when there is no other dataset available in the country.
- Sampling has to guarantee coverage of the total population, and weighting[1] must ensure that the sample is representative of the total population and subgroups in the country, e.g., by age, sex, regions, or total labor force.

In the first *opening* of the microdata, the LIS team assesses the quality of the source data. Particularly the screening of the source data with respect to representativeness of income and wealth indicates the need of a conceptual framework that defines the ideal concepts underlying income and wealth items. This framework has been established gradually over many years of experience through involvement in international scientific projects and contribution to expert reports. More specifically, the conceptual framework of *disposable household income, assets*, *liabilities, and net worth* has been ensured by LIS' participation to the Canberra Group and other expert groups such as the OECD. The resulting publications (Canberra Group, 1999; UNECE, 2011; OECD, 2013a; OECD, 2013b) have shaped into internationally acknowledged guidelines for producers and users of household income distribution statistics.

A second element of the *opening* is to clarify whether the income and wealth items are collected with a low degree of missing values. LIS has established as a rule of thumb that the percentage of missing values for *disposable household income* should not exceed 20 percent. As this situation is a result of accumulating person and item unit-non response patterns from the individual income sources, it means that the main

individual income source is expected to have only about 5-10 percent of missing values. A higher percentage of missing values gets particularly problematic for representativeness of large households, where missing values are more likely to occur. When, on the other hand, the source data were already imputed by data providers (and this is typically the case with wealth data, which suffers from high numbers of missing values), the entry criterion for inclusion into the databases focuses on the quality of the imputation of missing data. Imputation has to be carried out utilizing well-documented statistical techniques such as multiple imputation, in compliance with latest scientific standards.

Once a dataset passed the entry criteria, the actual harmonization work starts. Before presenting concrete examples of day-to-day tasks of harmonization work at LIS, it is crucial to clarify general guiding principles that have influenced the development of the conceptual framework, the workflow at LIS, and the technical infrastructure for the *LIS* and *LWS* Databases. Due to cross-national differences in the source data, LIS established *operational comparability* as a guiding principle for *ex-post* harmonization. *Operational comparability* entails finding the right compromise between, on one side, creating variables following concepts that are purely comparable from the theoretical point of view, and on the other side, seeking to make sure that the harmonized variables are available for most datasets. Thus, *operational comparability* allows researchers to analyze variables across several datasets, i.e. making comparability operational.

The outcome of *operational comparability* is best shown with examples. LIS' concept of *disposable household income* excludes the value of *imputed rent* (market rent that would be paid by homeowners and subsidized or rent-free tenants for a dwelling similar to the occupied one) from its definition, although its inclusion is widely advised in various scientific guidelines. Due to scarce availability and comparability limitations of the concept of *imputed rent* in the source data, a conceptual inclusion of imputed rent would lower *operational comparability.* Thus, users of the LIS databases would only have a limited set of datasets to their disposal.

That said, after careful selection of datasets where *imputed rent* is available, researchers can still add it to the definition of disposable household income. The conceptual framework for *disposable household income* is summarized in Figure 2, where it can be seen that *social transfers in kind (STIK)* are also excluded due to similar reasoning.

Figure 2. Conceptual framework for *disposable household income*

	CASH	**NON CASH**
LABOUR INCOME	*Wages, salaries, bonuses*	*In-kind earnings*
	Profits and losses from self-employment	*Own consumption*
+ CAPITAL INCOME	*Interest and dividends*	-
	Rental income	~~*Imputed rent*~~
+ PENSIONS	*Universal and assistance*	-
	Contributory public insurance	-
	Occupational and individual	-
+ PUBLIC BENEFITS	*Family benefits*	~~*STIK*~~

	Unemployment benefits	
	Sick pay / work injury / disability benefits	
	Housing / heating benefits	
+ PRIVATE TRANSFERS	*Scholarships, charity*	*In-kind assistance*
	Alimony, remittances	*Gifts*
- DEDUCTIONS	*Income taxes*	
	Social security contributions	
= DISPOSABLE HOUSEHOLD INCOME		

Likewise, the conceptual treatment of pension assets when defining the measure of *disposable net worth* is also guided by *operational comparability*. While there is scientific consensus that *total assets* should include the totality of pension assets, the current value of all pension assets (individual, occupational and public) are rarely collected or calculated by the data providers. In order to address the gap between scientific concepts and the availability of data, LIS provides several alternative *net worth* measures that gradually include additional parts of pension assets, hence allowing researchers to compare the different wealth measures available in the source data. Figure 3 summarizes the various net worth definitions.

Figure 3. Conceptual framework for *net worth*

NON-FINANCIAL ASSETS
+ FINANCIAL ASSETS
- TOTAL LIABILITIES
= DISPOSABLE NET WORTH
+ LIFE INSURANCE AND VOLUNTARY INDIVIDUAL PENSIONS
= ADJUSTED DISPOSABLE NET WORTH
+ OTHER PENSIONS
= TOTAL NET WORTH

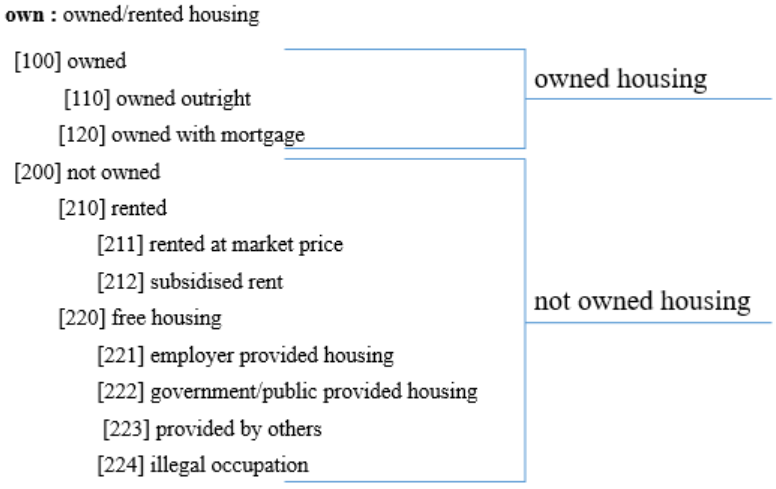
Alongside striving for *operational comparability*, LIS harmonization work aims as much as possible to reach full *standardization. Standardization* refers to the unified meaning of value codes and labels in the variables across countries and thus enables an easy understanding of the harmonized data. In the LIS databases, many categorical variables include *standardized* categories to create a high degree of *operational comparability* in the harmonized data files. Thus, LIS systematically groups the national information to more broadly defined categories, ensuring more widely available and comparable country samples – another facet of *operational comparability*.

While working on the data, the LIS team follows a set of Stata programs, the *harmonization template.* The dataset-specific programs include three elements: the documentation of variables in the source data, the coding of the final LIS variables, and any relevant discussion about country-specific decisions. Thus, once familiar with this process, each person at LIS can easily follow the harmonization practices carried out by co-workers in other datasets.

Once the LIS team has performed carefully data management, programming and documentation to reflect the survey-specific situations, various internal routines in the *harmonization template* allow an automatized creation of partial or complete LIS files. These routines automatically create some of the LIS variables, where information can be derived from other variables. The automatized routines simplify especially the construction of hierarchical sums in the income or balance sheet classifications and ensure consistency across countries and over time. Further, these procedures enable a consistent treatment of person and item-unit non-response patterns in the *LIS* and *LWS* datasets.

After presenting the harmonization methods in a rather general and abstract way, some practical examples of the harmonization work are presented below. Figure 4 illustrates two points, the complete *standardization* of most LIS categorical variables and the *nesting* of subcategories within the more aggregated levels. For example, the variable *own* (owned/rented housing) contains at the highest level of aggregation the code *owned* (100) and *not owned* (200); those are widely available across countries. The idea of *nesting* refers to the subgroups within the major groups. These codes are sublevels of *owned* (100) and *not owned* (200); all information in the source data is recoded to any of the twelve specified categories and is thus fully *standardized* and *operationally* comparable. However, depending on the information available in the source data, either sub-groups or major categories are constructed. A more detailed breakdown is available for a smaller set of countries. It is essential that users of the harmonized data consult the *Metadata Information System* (*METIS*) and tabulate the variables, so that they conclude whether in all their selected datasets sufficient level of information is available, e. g. that all datasets allow for analyzing the two subgroups *owned outright* vs. *owned with mortgage*.

Figure 4. Aggregation of subcategories into overall categories

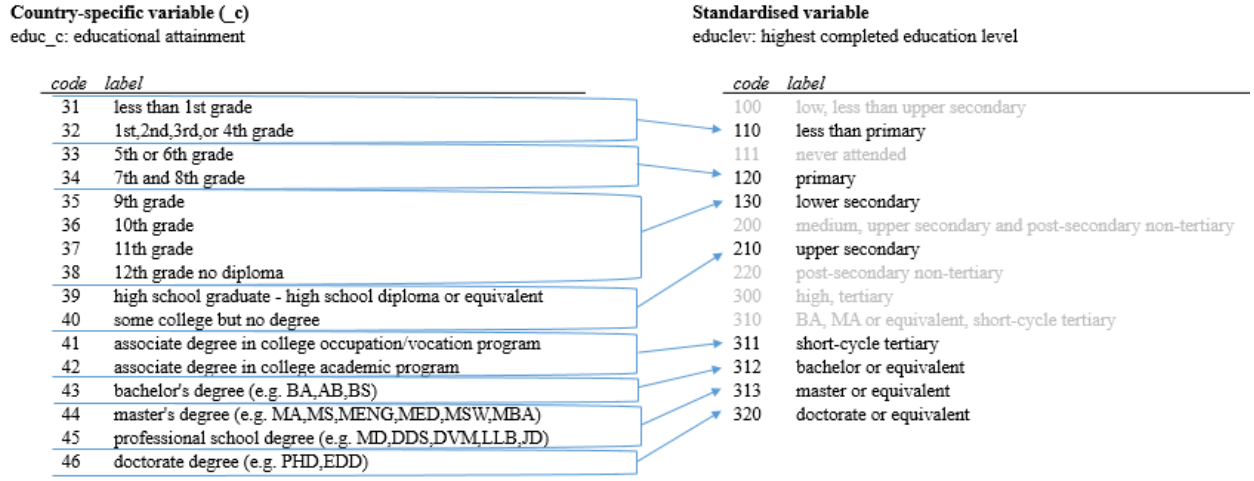


On the other hand, for variables where the national contents prove impractical to be harmonized, LIS provides dataset-specific categories and labels; technically, LIS variables ending with *_c* indicate country-specific variables. These variables derive from three specific considerations. First, geographic information follows national groupings, which cannot be further standardized to a common structure. Second, for

various classifications, such as *highest education level achieved*, *industry/occupation in the main job*, LIS provides more national detail next to the highly grouped information in the standardized variables following international classifications. These country-specific variables open up the possibility of in-depth country case studies embedded in the otherwise harmonized cross-country context. Third, variables about financial literacy, type of business, or subjective health status contain information that is collected differently across countries. Hence, a *standardization* of codes reflecting all various national measures is not possible.

Figure 5 exemplifies the harmonization for the *highest level of education achieved*. National education systems differ across countries, thus researchers are in need to study differences in educational levels when comparing educational qualifications across countries. This is where the harmonization efforts of the LIS team comes into play. By closely following the *International Standard Classification of Education (ISCED) mappings* by the UNESCO Institute for Statistics (UIS) and its Member States, the LIS databases provide a standardized variable. However, this standardized variable cannot simply follow an ideal classification, as it needs to accommodate for cross-national differences in data collection. Thus, for example, code 310 *(BA, MA or equivalent, short cycle tertiary*) was nested in the variable codes for datasets where educational qualifications were collected with less detail (e. g. the *first stage of tertiary education*). This way, without the need of further documentation, it gets relatively clear which level of detail is available in a specific dataset. At the same time, by making available the country-specific detail in variable *educ_c*, LIS provides both the documentation of the national categories collected in the source data and the re-classification to the standardized (typically less detailed) comparable categories.

Figure 5. Country-specific variable and *standardized* variable for educational qualification



Note: the above example refers to harmonization of the United States 2019 dataset.

Altogether, the various elements described above exemplify that the *ex-post* harmonization process requires a multifaceted skill set to harmonize microdata for the LIS income and wealth databases properly. For a high-quality *ex-post* harmonization, persons working on the preparation of the data need expertise on scientific concepts and need to be able to adapt these concepts to the national context. At the same time,

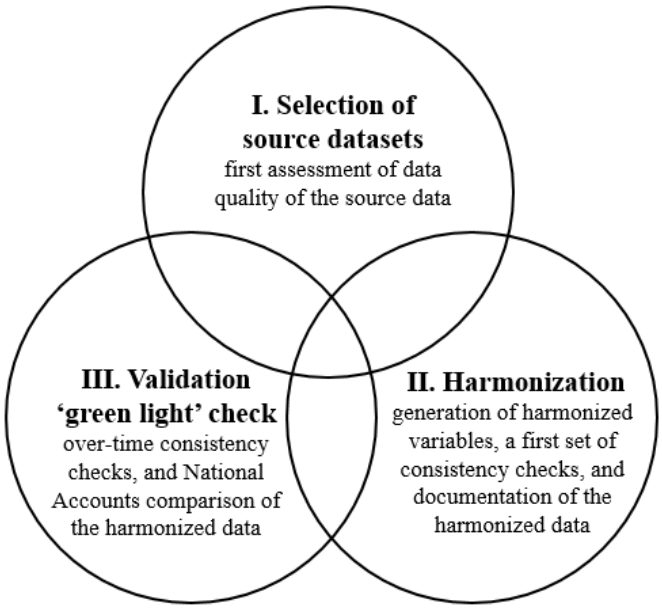
they need to understand what it means to provide comparable data. LIS operates under the assumption that a solid understanding of cross-national differences is more relevant than the national expertise to ensure maximum consistency for comparative research. Therefore, *internal* harmonization rather than outsourcing to country experts is the primary mode of working at LIS. The microdata experts pursue the whole harmonization chain and gradually build up invaluable expertise to reflect carefully about country differences and the challenges of harmonizing cross-national data. On the other hand, LIS has established a network of country experts regularly consulted while carrying out the harmonization. The broad knowledge and the consultation of national experts ensure a good reflection of the national context.

3. Documentation and quality assurance

3.1 Quality assurance

Data quality at LIS is ensured at three levels: i) the careful selection of source datasets for the LIS databases, ii) the structured harmonization process, and iii) the careful validation of the produced data before they are included in the LIS databases. Next to this, there is a highly automated infrastructure, making the microdata available in the remote access system and providing the metadata documentation in an entirely standardized way. Figure 6 summarizes the three main quality assurance stages. The choice of a Venn diagram mirrors the interdependency of each stage with the other two.

Figure 6. The three main stages of quality assurance at LIS



I. Selection of source datasets

As shown in detail in Section 2, the selection and acquisition of source data is a main element of *ex-post* harmonization. Therefore, through a careful assessment of compliance criteria of the source data for

inclusion to its databases, LIS makes sure to select only datasets, which are of suitable quality for harmonization with the LIS conceptual framework, foremost with respect to completeness and representativeness of income or wealth information in the source data.

II. Harmonization

The structured harmonization process is the crucial element of LIS' quality assurance: the series of applied harmonization methods (described in Section 2) ensure that harmonization is carried out with solid expertise, foremost through the data team's fluency with income and wealth concepts, established labor market and education classifications, and multifaceted reflection of comparability problems in *ex-post* harmonized microdata.

III. Validation – 'green light' check

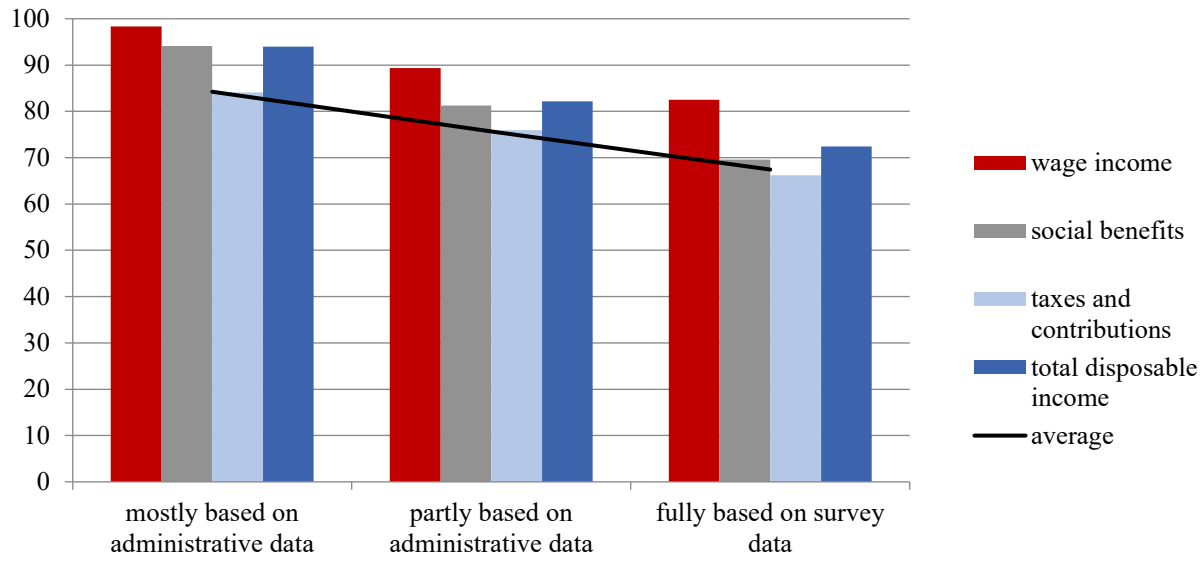
A significant element of the harmonization work at LIS is the *over-time consistency check*. To save time for the LIS team to compute over-time comparisons, LIS established a set of flexible *checking tools*, which the team can apply to several datasets simultaneously. These *tools* enable selecting specific variables or sections of variables, allowing filter conditions, and creating a final overview of the harmonized data for the final validation. A first consistency assessment is part of the individual data expert's tasks during harmonization, but also, a central role of data quality coordination for all harmonized datasets generally contributes to the validation stage of the harmonization. In this respect, data quality coordination refers to going again through all consistency files and picking up remaining comparability issues, which are discussed and solved in the team's joint effort.

The primary purpose of checks for categorical variables is to analyze frequency patterns of categories and missing value patterns over time. Automatized Excel sheet overviews for income, expenditure, and balance sheet variables enable a comprehensive view concerning the mean values, percentage of missing values, non-missing and non-zero cases, and percentage of non-missing and non-zero cases in income and wealth items. For mean and median statistics an automatized reading of consumer price indices and PPP conversion rates from the World Development Indicators (WDI), allow for exploring data in nominal values in local currency and in real values in International Dollars. All elements are critically assessed for comparability with other datasets. This process foresees going through frequency tables and the prepared documentation of the harmonized variables over time. Once all intermediate data modules of the harmonized subsections are thoroughly checked for consistency, the final person and household level files are constructed.

A second major component in the validation stage is comparing various components of income with corresponding National Accounts. This enables LIS to understand better the quality of the source and harmonized data with respect to representativeness of the information provided in the data. This is a crucial element for any data analysis aimed at evidence-based conclusions about identified patterns in the data. Figure 7 groups the harmonized datasets from 2013 to 2016. These groupings present from left to right the relevance of administrative data in the data collection phase. Thus, in countries where the microdata is

based on administrative data, the inflated numbers from the microdata best resemble the actual numbers in the corresponding National Account categories. On the contrary, where microdata solely rely on survey data, the ratios between inflated microdata against official National Accounts are typically the lowest.

Figure 7. LIS to National Accounts ratios in percent – around 2016 or latest year available



Notes: Category 'mostly based on administrative data' includes Austria, Denmark, Estonia, Finland, Iceland, Netherlands, Norway, Spain, and Sweden; category 'partly based on administrative data' includes Canada, Ireland, Lithuania, and Switzerland; category 'fully based on survey data' includes Australia, Brazil, Colombia, Czechia, Germany, Greece, Israel, Italy, Japan, South Korea, Luxembourg, Russia, Slovakia, United Kingdom, United States, and South Africa.

Source: Microdata come from the *Luxembourg Income Study (LIS) Database, National Accounts figures come from the Annual Detailed Non-financial Sector Accounts from the OECD Stats, June 2021*.

In addition, a specific *data quality protocol* that is run at the end of the harmonization allows assessing the harmonized datasets with respect to: i) sensitivity to extreme values (i. e. the impact of bottom and top coding procedures for income inequality measures), ii) internal consistency of overtime series (to make researchers aware of breaks in series due to changing methodology in the data collection), and iii) the coherence of the major aggregates derived from the microdata against corresponding National Accounts figures.

3.2 Documentation

Documentation at LIS fulfills a central role in the harmonization process. All dataset-specific documentation is provided in the *Metadata Information System* (*METIS*), an interactive selection tool for datasets and contents. *METIS* provides: i) detailed documentation of the source data entering the harmonized databases, including information on sampling, collection period, modes of collection and

instruments, various data quality aspects (non-sampling error, item non-response / imputation, weighting), and information on the collection of income, wealth, and labor force status; ii) descriptive statistics for all LIS variables in each dataset; iii) information on the exact contents of the income, consumption, expenditure, and balance sheet items, possibly with the indication of the country specificities (such as the name of the programs of the social security benefits included in the income variables); and iv) a series of *notes* to warn researchers about situations where the contents of a variable deviate from the ideal variable definition in the conceptual framework, hence giving rise to limitations in comparability.

4. Challenges to harmonization

The "*ex-post*" aspect of harmonization is what makes it most challenging. The variety of differences in the national data collection not only create the need for a substantial reworking of the source data during the *harmonization* process, but also require clear and accessible documentation. In this section, we describe key challenges of the *ex-post* harmonization work and illustrate with some examples limitations in *ex-post* harmonized data.

Harmonization of income and wealth data is especially challenging. The final measures need to be both exhaustive and without double counts. Regarding exhaustiveness, the main challenge is to decide which subcomponents to include to construct the best comparable total household income or net worth measure. For example, it is not always clear which in-kind incomes should be included in the harmonized variables. The LIS conceptual framework prescribes leaving out all *non-monetary universal transfers from government*, whereas in-kind income received within the scope of social assistance (such as food and clothes distributed to precarious households) should be included. It is the distinction between the two that requires a thorough understanding of the country's social system, as well as the availability of detailed information and documentation of the source data.

The risk of double-counting arises when information is collected in more than one part of the questionnaire. This situation appears most often in surveys that collect both income and consumption information so that some amounts can be recorded both as in-kind incomes in the income section of the questionnaire and as non-monetary consumption in the consumption part of the questionnaire. Similarly, the same subcomponents of wealth can be collected at both individual and household level, creating a danger of double counting as both head and spouse can claim the ownership of the same assets/debts. Persons working on the harmonization need to assess which amounts should be included from which section of the questionnaire.

Although LIS has put a strong effort in setting up a solid conceptual framework, practices for quality assessment and documentation, *ex-post* harmonization bears a major challenge. The various survey and administrative data from different countries do not necessarily collect perfectly comparable information. Here we pick up again the concept of *operational comparability* that we introduced in section 2), while clarifying harmonization methods. The principle is guiding LIS of what is possible to harmonize - thus defining the conceptual framework for *ex-post* harmonization, including the generic definitions for each

harmonized variable. Thus, over the years, LIS refined the variable list, to better adopt the harmonization work to refinements in the conceptual standards. Particularly the enlargement of the scope of *ex-post* harmonization to middle-income countries brought a further focus on non-monetary and in-kind benefits on the agenda of the harmonization work. However, the generic definitions assume availability of all income items in every national data source, whereas not necessarily all information is collected in such a way to fit in these variables perfectly – a major challenge to *operational comparability*.

Figures 8a and 8b illustrate two examples. Example a) shows differences in the way self-employment income is collected. In country A, all self-employment income, consisting of gains and losses from activity as employer, own-account work, and contributing work in family or farming business, is collected at the individual level. In country B, only gains and losses from activity as employer or own-account work is collected at individual level, whereas income from farming activity is collected from a section of household activity, not separating out the individual's involvement. Example b) points to the varying level of detail collected in the data. From a conceptual viewpoint occupational (2nd pillar) and individual (3rd pillar) pensions typically exist side by side in many countries, but in country A the level of detail is not available in the source data, and hence the *ex-post* harmonized cannot provide the detail either. Both examples highlight the need of comprehensive documentation of the harmonized data.

Figure 8 a) Example: availability of information person vs. household level

	Country A		Country B	
	Personal level variables	Household level variables	Personal level variables	Household level variables
wage income	✓	aggregated from personal level	✓	aggregated from personal level
self employment income	✓	aggregated from personal level	(✓) (excluding farm income)	aggregated from personal level + farm income
of which: farm income	✓	aggregated from personal level	✗	✓

Figure 8 b) Example: (non-)availability of most-detailed level

	Country A		Country B	
	Personal level variables	Household level variables	Personal level variables	Household level variables
private pensions	✓	aggregated from personal level	✓	aggregated from personal level
of which: occupational pensions	✗	✗	✓	aggregated from personal level
of which: personal pensions	✗	✗	✓	aggregated from personal level

Operational comparability is particularly difficult to achieve in the classification of social security benefits. Not only does the right conceptual placement require extensive knowledge of the national social security system, but also LIS needs to follow acceptable standards for social benefits cutting across various types of benefits provision, or various needs being relieved. For example, social assistance programs in emerging economies frequently cut across different providers, the state or non-governmental institutions. At the same time, these programs are set up to relieve individuals of the burden of different risks or needs, such as low

income, inadequate education of children, or insufficient nutrition, so that there is no unique placement where they belong to in the conceptual framework (at the same time general social assistance, family benefits, education or food benefits). Still, LIS tries to place them in the most appropriate placement for international comparison not to lose the detail for cross-country studies (in this example, general social assistance) and documents these decisions so that researchers better understand which benefits were included in LIS variables.

Besides the non-availability of information, other general limitations in comparability of income components arise; source data do not necessarily collect all income items with the same reference period, even within the same survey. In the case of actual income surveys (that collect the regular amount received at the last payment), *annualization* may imply a multiplication by the number of periods in the year if the income is received regularly (such as a weekly wage) but may need to be multiplied by a smaller factor in case the income is not regularly received throughout the year.

More generally, on the one hand, *ex-post* harmonization suffers from data providers not following precisely international standards and guidelines for definitions and collection. For instance, for wealth data, definitions of what constitutes business wealth and valuation criteria for assets are still different across countries. On the other hand, following international standards, e. g. such as ILO employment criteria, may lead to situations – especially in emerging economies - where a vast majority of the population is classified as mainly employed, but where very few people actually have a regular job, hence creating a bias when comparing emerging economies to affluent countries. As a result, even though there is not a direct challenge to the harmonization in this example, it points rather to research that needs to interpret the *ex-post* harmonized data from different regions in the world.

Last but not least, different cultures and languages cause varying ways of asking questions with differences in wording and understanding for one and the same concept. This is particularly relevant in trying to achieve a harmonized concept of disability. Besides the interpretation while translating documentation, national institutions may differ. Thus, in country A, disability might be linked to the eligibility to disability benefits, whereas in country B, a question about general limitations in daily activities, might be still influenced by the existing norms and social benefits for disability or invalidity. In country C, where such norms are possibly less common, the same question about limitations in daily activities might more generally lead to a rather subjective than objective answer. Due to these cross-national differences, LIS decided to always add clarifying remarks on the national contents for variable *disability* in *METIS*.

To sum up, despite the efforts carried out during a careful *ex-post* harmonization, not all of the challenges posed by the source data can be solved. It is imperative that persons working on harmonization recognize these limitations and that clear comparability warnings are passed on to the final data users through a sound documentation system. Likewise, it is crucial that data users bear in mind these limitations of *ex-post* harmonized data.

5. Software tools

As shown in Section 3, LIS puts a strong focus on standardization practices to ensure high-quality data and documentation. Thus, all datasets acquired by LIS to be released in the *LIS* and *LWS* Databases are entered directly to the *Metadata Information System* (*METIS*). *METIS* is a user interface consisting of a *frontend* and *backend* application that has been specifically built for the needs of LIS to provide general information on variable definitions, dataset availability, as well as dataset specific information on the source data, variable content, their availability, comparability warnings and basic statistics. The entries in the LIS internal *backend* interface allow a clear overview and easy navigation through the rich information stored in the various underlying databases. Various designated fields allow the LIS team to easily update the databases. Once synchronized with the *frontend* user interface, the information is then accessible for the public, where scholars can explore variables and documentation interactively.

Managing the data and their documentation with an internal *backend* at LIS has the advantage that datasets can be easily created, revised, selected and grouped by its current work status. Likewise, the *frontend* allows scholars to explore selected information across countries, which then can be easily exported.

For the latest harmonization *template*, LIS currently uses Stata as harmonization software. The various harmonization programs enable an automated run of all *LIS*/*LWS* Database variables. The specific country folders contain a standardized *.do*-file structure for variable construction. An internal Stata *.ado*-file allows execution of various pieces of the programs; the syntax recalls the standard folder structure for programs, intermediate and final data, which enables the possibility of a reoccurring update of the harmonized files once the original microdata is fed again in the system. During the harmonization, with the same Stata *.ado*-file, the various consistency checks (described in Section 3) can be carried out, which allows a tailor-made comparison of individual variables from datasets in progress and currently online.

For continuous variables in the income, consumption expenditure, and balance sheet, specific Excel tables were built that report for the selected datasets mean and median values, percentage of missing values, non-missing and non-zero cases, and percentage of non-missing and non-zero cases over time.

Once the harmonized data are created, the data and documentation are made available for external use. This stage includes a series of actions using various software (e.g., R, SAS, SQL), including conversion of the Stata microdata to other formats compatible with R, SAS, and SPSS so that the data can be accessed in the remote access system by different software packages.

6. Conclusion

As shown in this chapter, *ex-post* harmonization of data is not a trivial task, but it is necessary for comparative research that intends to use data from different sources and in different formats. When undertaking the task of *ex-post* harmonizing pre-existing data, it is imperative that all differences in the source data are carefully reviewed so that appropriate treatment can be applied for the best comparability

between harmonized datasets. We also highlighted that quality assurance procedures (including proper infrastructure for the data management) are necessary for harmonization that aims to construct data to conduct high-quality research.

As a pioneer in cross-national *ex-post* microdata harmonization, LIS has vast experience conducting the assignment of constructing datasets that enable comparative research on household income, consumption, net worth, portfolio composition, and income and wealth distributions. Through nearly 40 years of experience in harmonization and data dissemination for researchers, LIS has learned that to provide the best quality harmonized data is a combination of efforts that involve many actors. So far, we have not talked about the users of LIS data, who must be able to conduct state-of-the-art research with the cross-national databases. Thanks to the researchers, LIS can maintain these databases since their analyses provide feedback by pointing to some obstacles in the harmonized data and helping with the development of conceptual frameworks needed for the latest research advancements. The LIS users' numerous *LIS*/*LWS Working Papers* and peer-reviewed publications that use LIS databases can be viewed as proxies for the high quality of the harmonized data. It indicates that the researchers from various social science disciplines trust the harmonized *LIS*/*LWS* databases.

The establishment of networks with the researchers as well as mutual advice is an additional component of a broadly defined harmonization process that needs to be highlighted here. Thus, LIS also provides the annual week-long workshops in Luxembourg, which provide new and existing users with solid knowledge of LIS databases and the latest methodological advancements on welfare economics. LIS can provide such a service since the team regularly collaborates and interacts with prominent scholars in the field and the LIS team conducts research with *LIS*/*LWS* data.

The regular exchange through the extensive day-to-day user support with many of LIS databases users leads us to provide crucial general recommendations, while working with the harmonized data. Foremost, it is vital to understand differences in data context and consult each dataset's documentation. This recommendation relates to the fact that despite the efforts carried out during a thoughtful harmonization, not all of the challenges posed by the very different input data can be solved so that the final harmonized data could still have some comparability issues. Some harmonized variables might differ from the conceptual definitions set up as the standard for harmonization (due to the data collection); therefore, users of the *ex-post* harmonized databases are advised to consult the metadata documentation carefully. The remaining comparability issues might have implications on how the analysis should be carried out, as well as how the results should be interpreted.

What else have we learned through all these years as an institution that harmonizes and disseminates data for scientific inquiry? In order to provide reliable data for the cross-national interdisciplinary research, the processes of conducting LIS activities are not only related to the good source data, well-implemented harmonization procedures, quality assurance, documentation, infrastructure, and hiring highly skilled labor force. The success of the LIS databases also involves a network with other institutions and researchers to exchange knowledge and expertise that must be implemented into the overall activities of harmonizing

institutions. With this exchange, LIS goes beyond the *ex-post* harmonization procedures conducted in-house, having the objective to improve standardized practices for collecting income and wealth data. First, LIS provides international guidelines for data producers in the form of collaborative publications and trough the direct contact with data providers during harmonization. These are essential factors that help to improve not only the survey data but also the design of questionnaires. Second, through the established network and close ties to experts in the field, the LIS team easily learns about and adopts latest survey developments, advises on potential data improvements that can help comparative research, and provides feedback on quality and consistency of the source data. All these elements are the ingredients to make the efforts of our work thrive, continuing to improve the outcome of *ex-post* harmonized data.

[1] National data providers carry out the weighting, and as sampling and non-response patterns and methods in accounting for non-response differ, LIS has no influence on the construction and quality of the weighting factors, particularly for representative coverage of sub-groups in the survey. For the estimates to be representative of the whole population it is crucial that data providers calculate weighting factors that account for sampling probability, propensity to respond, and calibration to sub-groups in the sample.

References

ILO (2012). *International Standard Classification of Occupations: ISCO-08*. Geneva: ILO.

OECD (2013a). *OECD Guidelines for Micro Statistics on Household Wealth*. Paris: OECD Publishing. https://doi.org/10.1787/9789264194878-en.

OECD (2013b). *Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*. Paris: OECD Publishing. https://doi.org/10.1787/9789264194830-1-en

OECD. European Union. UNESCO Institute for Statistics (2015). ISCED 2011 *Operational Manual: Guidelines for Classifying National Education Programmes and Related Qualifications*. OECD Publishing. http://dx.doi.org/10.1787/9789264228368-en.

UNECE (2011). *Canberra Group Handbook on Household Income Statistics*. 2nd edition. New York: UN. http://digitallibrary.un.org/record/719970.

United Nations. Statistical Division. (2008). *International Standard industrial classification of all economic activities (ISIC)*. New York: United Nations.

Smeeding, T.; Schmaus; G; Allegrezza, S. (1985). An Introduction to LIS. LIS Working Paper Series. No. 1.

The Canberra Group (2001). *Expert Group on Household Income Statistics: Final Report and Recommendations*. Ottawa.