Methods Protocol for the Human Fertility Collection

O. Grigoriev, A. Jasilioniene, D.A. Jdanov, P. Grigoriev,
T. Sobotka, K. Zeman, and V.M. Shkolnikov

Introduction

The Human Fertility Collection (HFC) is part of the Human Fertility Data Project, which is a joint project of the Max Planck Institute for Demographic Research (MPIDR) and the Vienna Institute of Demography (VID). The aim of the project is to compile and maintain two companion databases based at the MPIDR: the Human Fertility Database (HFD) and the Human Fertility Collection. The HFC has been designed to supplement the HFD and to provide the international research community with free, user-friendly access to a wide range of fertility data that, for various reasons, cannot be included in the HFD.

The HFD is the primary database of the Human Fertility Data Project. The data that are entered in the HFD are expected to have high levels of quality and detail. The data are entirely based on official and detailed vital statistics, and the database organizers place a great deal of emphasis on data checking and documentation, and on ensuring data comparability across time and countries through the application of a set of comprehensive methods. Because these standards are rigorously enforced, the HFD is a valuable data source, especially for scientific fertility research, but most of its data are on Europe and other advanced countries. Additionally, the HFD focuses primarily on period and cohort fertility by age of the mother and birth order, and has limited scope for taking into account other fertility dimensions (e.g., region of residence, ethnicity, marital status). The HFC, by contrast, is intended to be more flexible. It is capable of integrating a broad variety of fertility data pertaining to national and regional populations, as well as to various sub-populations. The quality requirements for the data selected for the HFC are less strict than those for the HFD, which allows for the expansion of the geographic coverage of HFC data to less developed parts of the world.

The HFC provides fertility data assembled from different (and not necessarily official) sources, such as statistical and scientific publications, online databases of national statistical offices, and data collections compiled by individual researchers and research organizations. At present, the HFC is based on one type of primary data: the period age-specific fertility rates (ASFRs) for all birth orders combined and by birth order[1]. The original ASFRs undergo an adjustment procedure that standardizes the data with respect to the age scale and the birth order range (see section 2 for details). In cases in which the original ASFRs are available only by aggregated age groups, the detailed age schedule provided should be used with caution. Caution is needed because the results of adjustments, while seemingly very plausible, do not necessarily reflect the real (unknown) shape of the age-specific fertility curve across single-year ages. On the basis of the adjusted ASFRs, the cumulative period

[1] At present, the HFC displays fertility data by biological birth order only. This means that the child is ranked in relation to all of the previously live-born children of the mother, irrespective of her marital status at birth.

fertility rates (CPFR), the period total fertility rates (TFR), and the period mean ages at birth (MAB) are calculated; this is done for all of the birth orders combined and, when available, by birth order (see section 3 for details).

All of the output HFC data are organized in a uniform format, and are provided together with full references to their sources. The raw data[2] are also made accessible to HFC users: the HFC provides downloadable original data files, copies of publications, or the internet pages from which the data originate in PDF format. For the data from the data collections assembled by individual researchers or research organizations, descriptions of their estimation methods are supplied when available. These descriptions are placed together with the raw data files in zip archives. Detailed information on the structure of all of the data files available in the HFC is provided in the file *Data formats*[3].

In addition to female fertility data, the HFC also provides fertility data for males for countries for which such data are available. Male fertility data are processed following the same HFC data standardization and computation procedures as data for females, with a few modifications required by specific features of these data. The processing of male fertility data is addressed separately in section 4.

Compared to the data provided in the HFD, the data provided in the HFC may be of lower quality, may have breaks in the time series, and may not always be comparable across countries and time due to variability in their origins and estimation methods. Furthermore, in the HFC the original data producers and providers bear the responsibility for the quality of the data they provide. The HFC team engages in only very basic data checking to ensure that no obviously incorrect data enter the database. If we detect errors or other problems in the data, we do not include them or exclude them if they had already been published in the HFC. HFC data users are therefore advised to consider whether the use of HFC data is appropriate given the analytical purposes of and the methodology applied in their work.

[2] In this document the term "raw data" always refers to the original data before any further modifications were made using the HFC methods.
[3] The file "Data formats" is available for download on any country Data page on the HFC website.

1 General Principles and Data Processing in the HFC

1.1 Raw Data

The period unconditional age-specific fertility rates[4] (ASFR) are the only raw data that are currently collected for the HFC, and they are used as the input data after being converted into the HFC standard format. The raw data are compiled from different data sources, and mainly originate from official websites of national statistical offices, official statistical publications, and data collections assembled by individual researchers or research institutions.

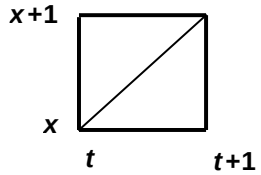
The raw data on the ASFRs vary considerably with respect to the definition of age of the mother, the age scale, and the range of available birth orders. The HFC integrates these kinds of data with a few exceptions. For example, cases in which the first age interval in the data is "20 and younger" or the last age interval is "30+" are not accepted in the HFC. In terms of the birth order, it is stipulated that at least the birth orders 1, 2, and 3+ are available in the raw data.

1.2. Age Definition

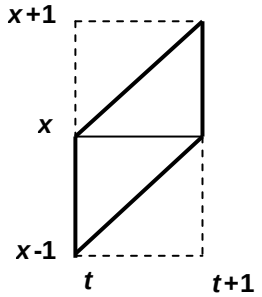
Two age definitions are used in the raw data on the ASFRs: the age in completed years (ACY) and the age reached during the year (ARDY). ACY, or the age at last birthday, represents a woman being at the age of x completed years within the time interval $[t, t+1)$. When the birth rates are classified by ACY, it is implied that the data at a given age x include information from two cohorts of mothers born in years t-x and t-x-1. Its configuration corresponds to the square on the Lexis diagram (Figure 1; see Jasilioniene et al. (2015) for details). The ARDY data do not mix different cohorts. All of the women aged x-1 at the beginning of the year t reach the age x during this year. Thus, ARDY is equal to the difference t-c, where c is the year of birth. On the Lexis diagram, the birth rates classified by ARDY have a configuration of vertical Lexis parallelograms (Figure 1). Information on the age definition for every data series is given in the field "Age definition" (ACY or ARDY) in the data files (see *Data formats* for more details).

[4] Unconditional age-specific fertility rates are obtained by dividing the births to women at age x in a given year t by the person-years lived in that year by all of the women of this age, irrespective of their parity status.

Figure 1. Lexis shapes of the ASFRs available in the HFC



Square (or rectangle)



Vertical parallelogram

1.3 Adjustments to Raw Data

The HFC methodology includes procedures that transform the varying raw data into a format with a standard age scale and a standard birth order range. The standard age scale in the HFC includes the ages ≤14, 15, …, 49, and 50+. Thus, the raw data on the ASFRs that are classified by aggregated age groups (e.g., five-year age groups) are split into single-year ages by means of interpolation (see sections 2.1 and 2.2 for details). In the cases in which the raw data are classified by single years of age, but the original age scale goes beyond the standard age limits of the HFC (e.g., 12, 13, 14, 15,…, 49, 50,…, 55+), the age groups "≤14" and/or "50+" are created by aggregating the corresponding single ages (section 2.3). If the original age scale is shorter than the HFC standard and there are no open-ended age intervals, the rates missing at the age tail(s) are assigned the missing values ("."), and in the calculation of the TFRs and the MABs are assumed to be equal to zero.

Regarding the birth order, the HFC provides the following standard categories: 1, 2, 3, 3+, 4, 4+, and 5+. Depending on the available raw data, appropriate higher-level birth orders can be aggregated (see section 2.3), but lower-level birth orders (e.g., 3+ or 4+) are not split. For example, if the original ASFRs are classified by the birth orders 1 through 5+ or higher, then the values for the birth orders 3+, 4+, and 5+ are being calculated by summation. If the original source provides the birth rates up to the birth order 4+, then fertility rates for the birth order 3+ are being additionally calculated. Finally, if the original ASFRs are available for the birth orders up to 3+, no additional calculations are performed.

1.4 Aggregated Fertility Indicators

On the basis of the ASFRs in the standard age scale and with the standard birth order, the following period fertility indicators are calculated:

- cumulative period fertility rates (CPFR),
- period total fertility rates (TFR), and
- period mean ages at birth (MAB).

These fertility indicators are calculated for all of the birth orders combined, and, when available, by birth order. The respective computational procedures are described in section 3.

Data users should be aware that although the values of the aggregated indicators (except of the TFRs) estimated by the HFC team are consistent within the collection, they do not always exactly match the corresponding officially reported estimates.

1.5 Data Presentation on the Website

There are two major groups of output data files available on the HFC website:

1. **Pooled data files** (or multiple-source data files), which combine data from all data sources; and
2. **Single-source data files**, in which each file contains data from one particular data source.

The data in the pooled as well as in the single-source data files are presented in a uniform HFC format as comma-delimited text files (see sections 1.5.1 and 1.5.2 for further details). The exception are the **raw data files,** which are also available for download on the website. These raw data files—which can be downloaded as Excel, PDF, or other file formats—show the ASFRs exactly as they are in the original data source. A detailed description of the data file formats can be found in *Data formats*.

1.5.1 Pooled (multiple-source) data files

There are three different pooled data files provided in the HFC. They separately display data for all birth orders combined, data by birth order, and data on male fertility. The three files are as follows:

1. ***Adjusted ASFRs, with a standardized age scale and, when possible,* a *standardized birth order range***. Cumulative fertility rates (CPFR) calculated on the basis of the adjusted ASFRs are also included in these files.
    - ✓ Pooled data files on the adjusted ASFRs are available both for the entire HFC and for each country separately.
2. ***Original ASFRs***, with varying original age scales and birth order ranges (up to the birth order 5+).
    - ✓ Pooled data files on the original ASFRs are available only for the entire HFC. These are all single-source data files compiled in a single file (see also section 1.5.2).
3. ***TFRs and MABs,*** calculated on the basis of the adjusted ASFRs.
    - ✓ Pooled data files on the TFRs and the MABs, together with the PDF files that graphically illustrate the trends in these indicators (for all birth orders combined only), are available both for the entire HFC and for each country separately.

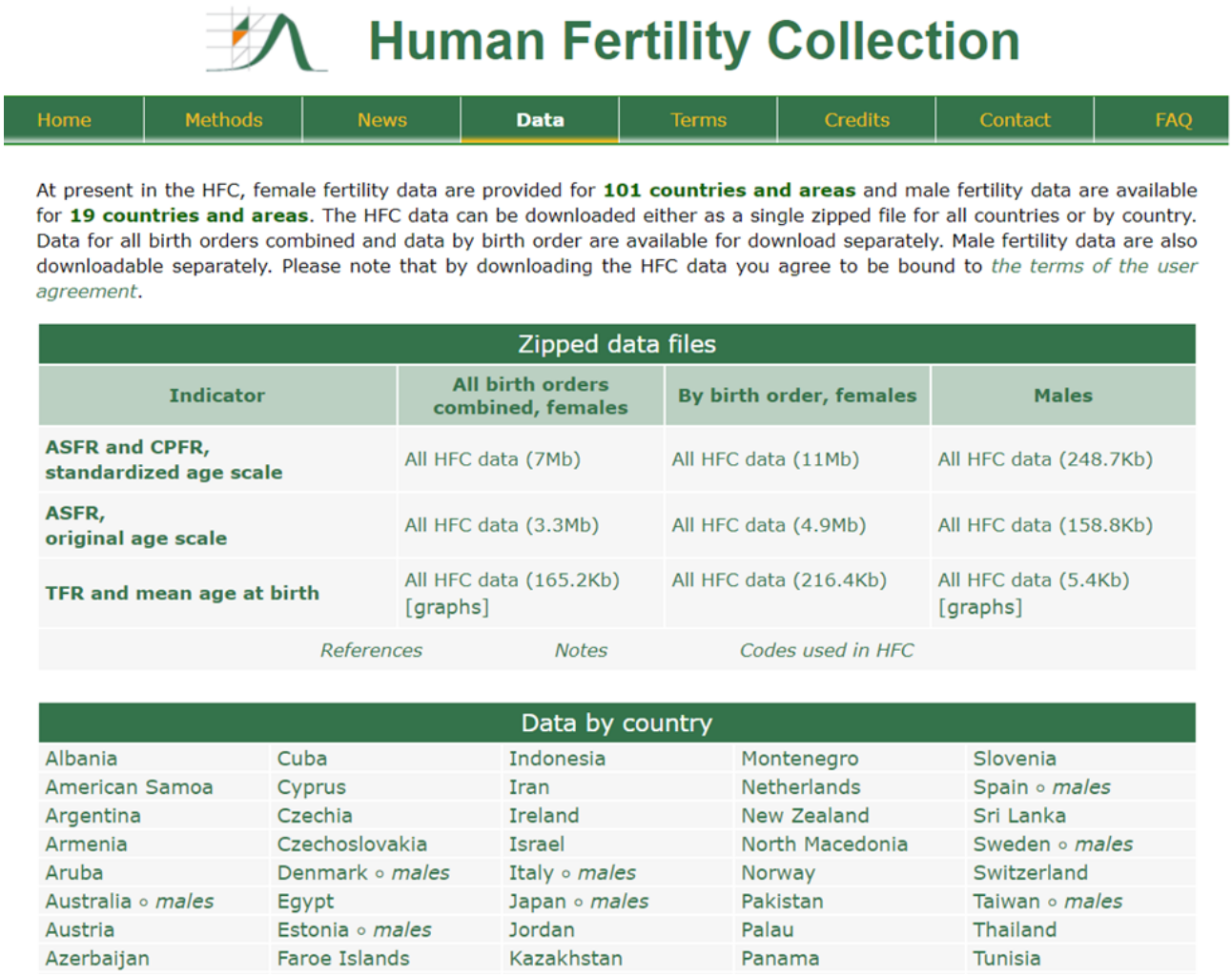
1.5.2 Single-source data files

Single-source data files contain raw data on the ASFRs. The data in these files are organized according to the standard HFC data file structure (see *Data formats* for details), but preserve the original age scale and the original birth order range (except that the birth orders higher than 5, when available, are aggregated in the birth order category 5+). The ASFRs for all of the birth orders combined and the ASFRs by birth order are provided in separate single-source data files.

1.5.3 Arrangements of the files on the web

The page "Data", which can be found on the horizontal menu of the HFC website, provides an overview of all the data available in the HFC. The *HFC Data page* is divided into two blocks: "Zipped data files" and "Data by country" (Figure 2). The three types of **pooled data files for the entire HFC** (see sub-section 1.5.1 for the description of these files) are available for download in the block *Zipped data files*. Users who wish to download large amounts of HFC data quickly may prefer to use these zipped files. Separate zipped files are created for data for all birth orders combined, for data by birth order, and for male fertility data.

Figure 2. HFC Data page



Human Fertility Collection

Home | Methods | News | **Data** | Terms | Credits | Contact | FAQ

At present in the HFC, female fertility data are provided for **101 countries and areas** and male fertility data are available for **19 countries and areas**. The HFC data can be downloaded either as a single zipped file for all countries or by country. Data for all birth orders combined and data by birth order are available for download separately. Male fertility data are also downloadable separately. Please note that by downloading the HFC data you agree to be bound to *the terms of the user agreement*.

Zipped data files

Indicator	All birth orders combined, females	By birth order, females	Males
ASFR and CPFR, standardized age scale	All HFC data (7Mb)	All HFC data (11Mb)	All HFC data (248.7Kb)
ASFR, original age scale	All HFC data (3.3Mb)	All HFC data (4.9Mb)	All HFC data (158.8Kb)
TFR and mean age at birth	All HFC data (165.2Kb) [graphs]	All HFC data (216.4Kb)	All HFC data (5.4Kb) [graphs]

References *Notes* *Codes used in HFC*

Data by country

Albania	Cuba	Indonesia	Montenegro	Slovenia
American Samoa	Cyprus	Iran	Netherlands	Spain ◦ *males*
Argentina	Czechia	Ireland	New Zealand	Sri Lanka
Armenia	Czechoslovakia	Israel	North Macedonia	Sweden ◦ *males*
Aruba	Denmark ◦ *males*	Italy ◦ *males*	Norway	Switzerland
Australia ◦ *males*	Egypt	Japan ◦ *males*	Pakistan	Taiwan ◦ *males*
Austria	Estonia ◦ *males*	Jordan	Palau	Thailand
Azerbaijan	Faroe Islands	Kazakhstan	Panama	Tunisia

Note: The screenshot shows the *HFC Data page* as of May 2020. It might not display all the countries actually included in the HFC.

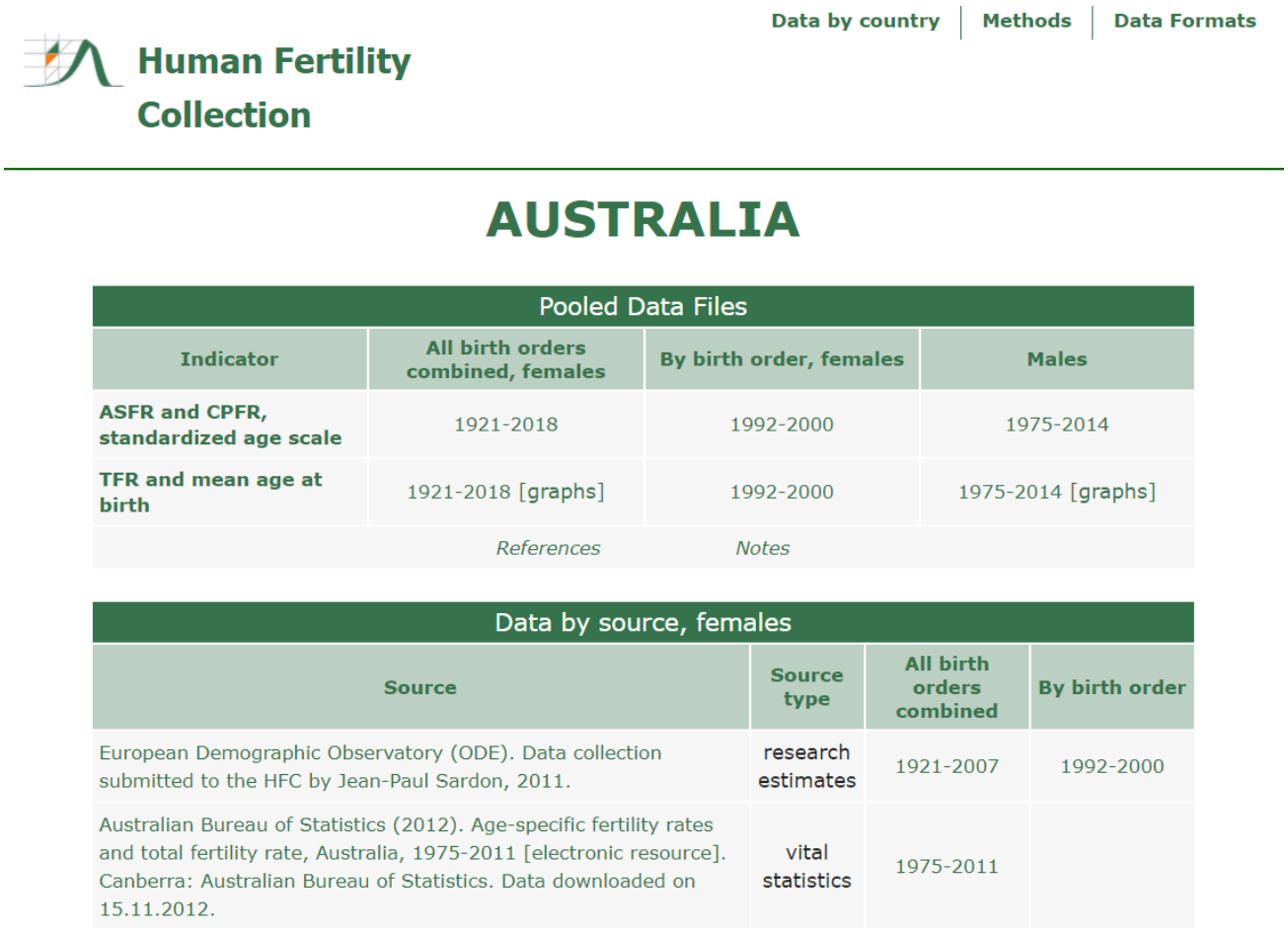
Country-specific data can be accessed in the block *Data by country*, which is right below the block *Zipped data files*, by clicking on country's name. The word "males" written next to country's name informs users that fertility data for males are also available for this country and can be downloaded on the county's page.

Data on the country page are presented in two or, if male fertility data are also available, three blocks (see Figure 3 showing the HFC country page for Australia for illustration). The block *Pooled data files*, which is on the top of the HFC country page, provides **country-specific pooled data files**, including files for the adjusted ASFRs and the CPFRs and files for the TFRs and the MABs (for an explanation, see also sub-section 1.5.1). The three types of pooled data files are created separately for data for all birth orders combined, for data by birth order and, if available, for male fertility data.

The other block, called *Data by source*, is designed for displaying **single-source data files** containing raw (original) data organized in the standard HFC format (see section 1.5.2). In case data on male fertility are also available, the block is split into two parts and the name of each part includes an extension specifying whether the data are on female or male fertility. As the ASFRs for the same years can originate from different sources, the files for each data source are shown separately. Users can download these files (either for data for all birth orders combined or by birth order) by clicking on the corresponding period (Figure 3).

Raw data files (with the ASFRs in their original format) are displayed next to the related single-source data files, and can be downloaded by clicking on the respective data source (Figure 3). The format of these files varies depending on the source: it can be an Excel, PDF, text file, etc. When the original ASFRs are obtained as electronic resources, the original URL is provided for the HFC data users, and can be found in the list of *References*.

Figure 3. The HFC country page for Australia



Note: The screenshot shows the country page for Australia as of May 2020. It might not display all the data available actually for this country in the HFC.

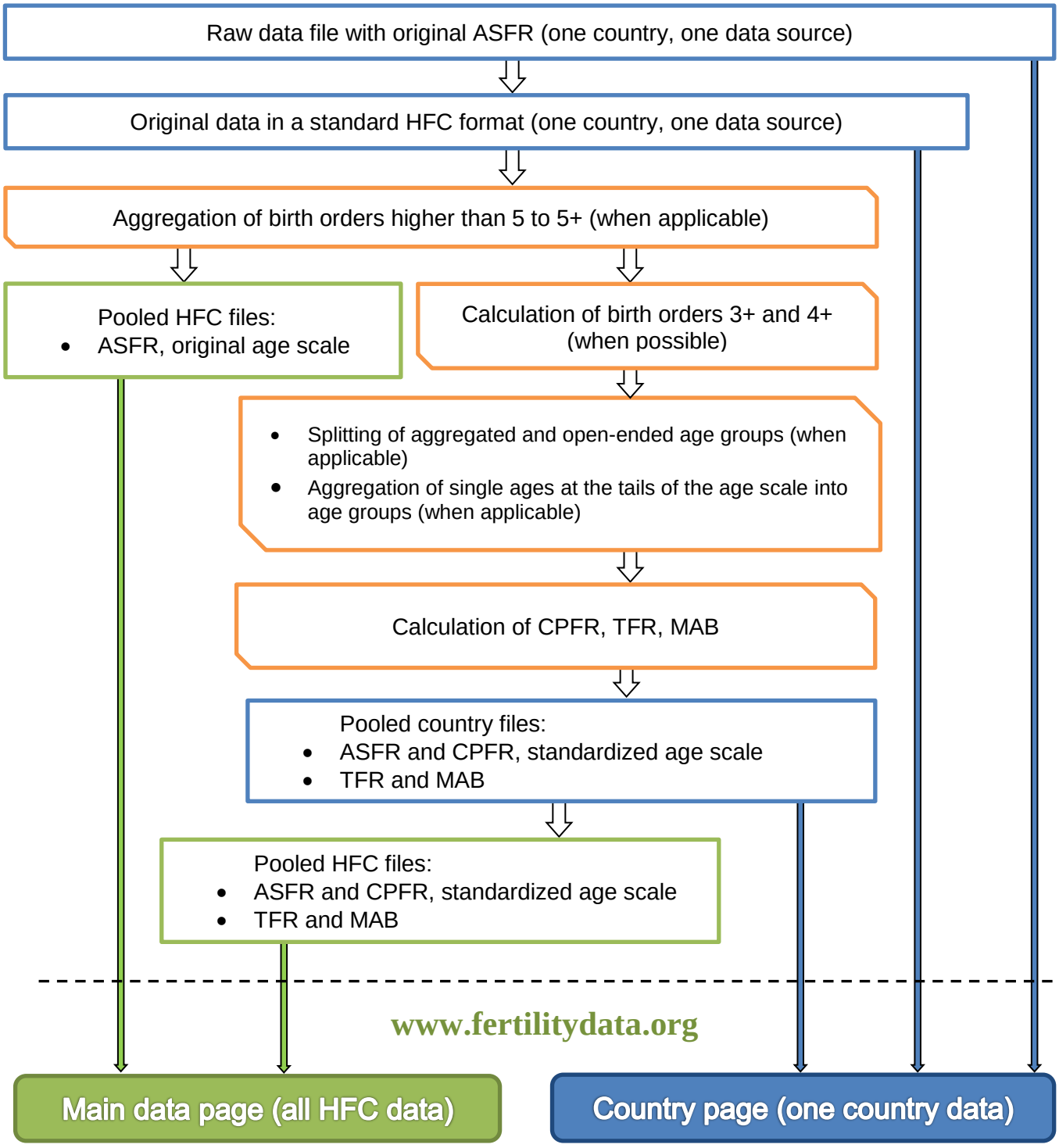
1.6 Data Flows in the HFC

The major steps of data processing in the HFC are illustrated in Figure 4 below.

It is important to note that fertility data for females and males are processed independently. The raw data on the ASFRs (for females and males) are collected by sources and countries separately, and are then transformed into uniform HFC format data files (input data files). In cases in which the birth order-specific data are available for the birth orders higher than 5+, the higher birth orders are summed up to 5+ before the subsequent data standardization steps are undertaken.

Depending on the original age scale, several data adjustment procedures (splitting or aggregation) are applied to the data. The computation of the aggregated fertility indicators (CPFR, TFR, and MAB) is then performed. The results of the calculations are merged into the country-specific pooled data files (separately for females and males), and then into the HFC pooled files (again, separately for females and males).

Figure 4 Data processing in the HFC



2. Common Adjustments to Input Data on the ASFRs

There is a significant degree of variability in the formats of the raw data in the HFC. For some countries and years, the ASFRs are available by single-year ages, while for the others they can be obtained only by aggregated age groups. There is also variation in the birth order ranges.

This section describes the methods used in the HFC data adjustment procedures, which have been performed to enable us to present the fertility data in a uniform format. The input data are the original ASFRs for all of the birth orders combined $f(x)$ and the birth order-specific ASFRs $f_i(x)$. Here and henceforth, the argument *x* denotes the age and the index *i* represents the birth order. Depending on the age definition used in the original estimates of the ASFRs, the newly produced estimates are also provided either by age in completed years (ACY) or by age reached during the year (ARDY). Regardless of the age definition, the identical formulae are being used.

The following data adjustment procedures, when needed, are applied to the original data for females in the HFC and, with some modifications (related to differences in the age scale defined as standard in the HFC), can be used to adjust the original ASFRs for males too:

1. Splitting of aggregated or open-ended age groups into one-year age groups:
   i) In cases in which the original ASFRs are given by aggregated age groups (e.g., 15–19, 20–24, 45–49), the *calibrated spline* estimator (Schmertmann, 2012; 2014) is applied for splitting (a more detailed description of the method is provided in section 2.1).
   ii) For the data by one-year age groups, but with an open-ended age interval at the beginning and/or at the end of the age scale (e.g., ≤15, 16,…, 48, 49+), *Hermite interpolation* is employed (see section 2.2 for more details).
2. Aggregating age groups and birth orders. When the ASFRs are available by single-year age groups, and the age scale begins with an age lower than 14 (e.g., 10 or 12) and/or ends with an age higher than 50 (e.g., 54), the rates are aggregated to obtain $f(\leq 14)$ and/or $f(50+)$, respectively (see section 2.3 for details). The original ASFRs for the other ages (15 to 49) remain unchanged.
3. In cases in which the birth order range in the original ASFRs does not correspond to the HFC standard, an aggregation of appropriate birth orders is performed (see section 2.3).

The adjusted ASFRs, together with the original ASFRs that required no adjustments, are then compiled into pooled data files, which are in turn used for computations of the CPFRs, the TFRs and the MABs (see section 3 for details). The adjustment procedure applied to the data, if any, can be seen from the value in the field 'Split': 0 – no adjustment; 1 – *calibrated spline* estimator,

2 – *Hermite interpolation*. All information on the structure of the files is available in the file *Data Formats*).

2.1 Splitting Aggregated Age Groups into One-Year Age Groups

The original ASFRs classified by aggregated age groups are split into one-year age groups in the HFC. In this section we describe the algorithm used to split data in closed (usually five-year) age intervals (e.g., 15–19, 20–24, …, 45–49). For cases in which the original rates are given by aggregated age groups with the first and/or the last open-age interval(s) (e.g., ≤14, 15–19, 20–24, …, 45–49, and 50+), the length of the open-age interval(s) is rated as the length of the second/next-to-last age groups (usually five years).

For the splitting of the aggregated age groups, the *calibrated spline* (CS) estimator proposed by Schmertmann (2012; 2014) is employed. The following description provides a very brief sketch of the method.

The CS estimator interpolates fertility rates by looking for a smooth curve, similar to that of the known fertility age patterns, and fitting it to the observed data. There are two criteria for the quality of approximation, named "fit" and "shape", for which the vectors of residuals should be close to zero (in the ideal case). In practice, this means that the fitting procedure must find an optimal balance between the shape and the fit. In the HFC, we follow the original approach, based on the assumption that the weights of the two criteria are of equal "importance". While the fit residuals can be easily defined as the difference between the quadratic B-spline basis function and the empirical values at the respective knots, the fit residuals have a complicated construction. They are estimated using the method of the principal component analysis. As a set of the known fertility age patterns (empirical basis), we use the same dataset as the one that was used in the original study by Schmertmann (2012): 304 single-year ASFR schedules from the Human Fertility Database (HFD) and 226 estimated schedules from the US Census International Database (IDB).

In the HFC, the uniform splitting procedure is applied to data for all birth orders combined, as well as to data by birth orders. In general, the CS estimator is heavily based on *a priori* information about the existing shapes of the ASFRs. Therefore, strictly speaking, the original algorithm by Schmertmann (2012) is not directly transferable to birth order-specific data. Nevertheless, due to the scarcity of detailed birth order-specific data, an identical empirical basis is applied to the birth order-specific rates in the HFC.

Because the CS estimator does not ensure the non-negativity of estimated rates and may change the resulting TFR, we use a two-step procedure for producing the adjusted ASFRs. In the first

step, the CS estimator is applied to calculate the single-year ASFRs varying from age 12 to age 54 from the original rates.

In the second step we apply the following adjustments:

1. The negative rates are replaced with zeros.
2. A proportional adjustment of the newly produced ASFRs is performed within all of the age groups to ensure that the TFR obtained from the new single-year ASFRs $f(x)$ is exactly equal to the TFR obtained from the original ASFRs[5]:

$$f(x) = \hat{f}(x) \frac{\sum_{j=1}^{K} (x_{j+1} - x_j) \cdot f(x_j; x_{j+1})}{\sum_{x=x_{\min}}^{x_{\max}} \hat{f}(x)} \qquad (2.1)$$

   Here x_j denotes knots of the original age scale, $f(x_j; x_{j+1})$ is the original rate at the aggregated age interval $[x_j; x_{j+1})$, *K* is the total number of age intervals in the original age scale, and $\hat{f}(x)$ is the interpolated rate at the age *x*.

   The procedure described above is applied separately for each birth order and for all of the birth orders combined. In the final stage, an iterative proportional fitting (IPF) procedure is applied to the birth order-specific data to keep the balance between the birth orders: at each age the sum of the birth rates by birth order should be equal to the birth rate for all of the birth orders combined. While the IPF does not change the ASFRs for all of the birth orders combined or the TFRs by birth order, it enables us to obtain a balance between the birth order-specific rates and the rates for all of the birth orders combined at each age. Further details on this procedure can be found in the HFD Methods Protocol (Jasilioniene et al., 2015). More details about the IPF technique are available in Fienberg (1970) and Bishop et al. (1975).

2.2 Splitting Open Age Intervals into One-Year Age Groups

For the original ASFRs that are presented by single-year age groups, but for which the first (e.g., ≤15) and/or the last (e.g., 49+) age interval is open, an additional splitting should be applied to obtain the standard HFC age scale ≤14, 15, …, 49, and 50+. The empirical calculations show that in cases in which the data between the open-ended age intervals are presented by one-year age groups,

[5] Note that in general the CS estimator does not guarantee the exact match between the newly created and the original five-year rates.

the CS method produces implausible ASFR estimates at the tails of the age distribution. To address this problem, we use the piecewise cubic *Hermite interpolation*. This method is identical to the method applied in the HFD (see the HFD Methods Protocol for the details: Jasilioniene et al., 2015)[6]. The data for each birth order are treated separately and independently from the data for the other birth orders. As in the HFD and in the splitting of the data with aggregated age intervals, we apply the IPF at the finale stage to ensure that balance between the birth orders is maintained.

2.3 Aggregating Age Groups and Birth Order Categories

For estimating the fertility rates for the age groups ≤14 and 50+, the following formulae are used in the HFC:

$$f(\leq 14) = \sum_{x=x_{\min}}^{14} f(x) \qquad f(\leq 15) = \sum_{x=x_{\min}}^{15} f(x) \tag{2.2}$$

$$f(50+) = \sum_{x=50}^{\max} f(x) \qquad f(59+) = \sum_{x=59}^{\max} f(x) \tag{2.3}$$

The same simple approach is employed both for all of the birth orders combined and for the order-specific data.

To compute the ASFRs for the birth order $k+$(where k=3, 4, 5), the rates for the birth orders k and higher are aggregated as follows:

$$f_{k+}(x) = \sum_{i=k}^{\max} f_i(x) \tag{2.4}$$

3 Computations of Aggregated Fertility Indicators

3.1 Cumulative Period Fertility Rates

When computed from the period age-specific fertility rates, the cumulative period fertility rate (CPFR) is a hypothetical construct that can be interpreted as the average number of children that would be born to a woman by age x if she experienced at all ages below x the set of age-specific fertility rates observed in a given year. In the HFC, the CPFRs are computed as follows:

---

6 The scripts for the application of the method can be found in the MPIDR technical report, "An 'R' package for the production of a Lexis database of fertility data" (Jdanov, Nash, 2011), available at: http://www.demogr.mpg.de/en/projects_publications/publications_1904/mpidr_technical_reports/an_r_package_for_the_production_of_a_lexis_database_of_fertility_data_4121.htm).

Cumulative period fertility rates by age *x* for all birth orders combined:

$$CPFR(x) = \sum_{z=x_{\min}}^{x-1} f(z) \qquad (3.1)$$

Cumulative period fertility rates by age *x* for birth order *i*:

$$CPFR_i(x) = \sum_{z=x_{\min}}^{x-1} f_i(z) \qquad (3.2)$$

where $f(z)$ is the ASFR for a specified age interval; *x* and *z* denote current age; and $x_{\min}$ corresponds to the lowest age at childbearing considered in the analysis.

For the open age intervals (i.e., ≤14 or 50+), the length of interval is assumed to be equal to one.

If the upper age limit of the summation is equal or very close to the maximum reproductive age (i.e., 50 years or higher), the cumulative period fertility rate equals the period total fertility rate (*TFR*).

3.2 Period Total Fertility Rates

The period total fertility rate represents the mean number of children a woman would have by the end of her reproductive life if she experienced at each age the age-specific fertility rates observed in a given year.

The TFR is calculated as a sum of the ASFRs pertaining to a specific period of time across all of the ages[7]:

$$TFR = \sum_{z=x_{\min}}^{x_{\max}} f(z) \qquad \text{(for all birth orders combined)} \qquad (3.3)$$

$$TFR_i = \sum_{z=x_{\min}}^{x_{\max}} f_i(z) \qquad \text{(by birth order } i\text{)} \qquad (3.4)$$

[7] In the HFC, we calculate the TFR and the MAB from the adjusted ASFRs by one-year age groups instead of using the original ASFRs.

3.3 Period Mean Ages at Birth

The period mean age at birth refers to the average age of the mother at childbearing, standardized for the age structure of the female population of reproductive ages. In the HFC, the mean age at birth is calculated on the basis of the schedule of the ASFRs.

The *mean age at birth* for all of the birth orders combined and by birth order *i* are:

$$MAB = \frac{\sum_{z=x_{\min}}^{x_{\max}} \bar{z} \cdot f(z)}{\sum_{z=x_{\min}}^{x_{\max}} f(z)} \tag{3.5}$$

$$MAB_i = \frac{\sum_{z=x_{\min}}^{x_{\max}} \bar{z} \cdot f_i(z)}{\sum_{z=x_{\min}}^{x_{\max}} f_i(z)} \tag{3.6}$$

Value $\bar{z}$ in formulae (3.5) and (3.6) is the mean age at birth within the elementary age interval [*z*, *z*+1):

$$\bar{z} = z + a(z), \tag{3.7}$$

where $a(z)$ is the average share of the age interval [*z*, *z*+1) lived before the birth of a child. We assume that all $a(z)$ values are equal to 0.5 if the age is defined as the age in completed years (ACY) and zero for the age reached during the year (ARDY).

In the HFC, we calculate the MAB from the adjusted ASFRs by one-year age groups instead of using the original data on the ASFRs. Respectively, if the original raw data are available only as aggregated age groups, the calculated MAB do not always match the corresponding officially reported estimates.

4 Male Fertility Data

In 2019 the HFC was expanded to incorporate fertility data for males. Although becoming a parent is a life event as important for males as it is for females, male fertility has been highly under-investigated by demographers. It is believed that the lack of demographic research on male fertility has been to a large extent caused by a shortage of data for males as well as various data quality issues encountered in these data (see Dudel and Klüsener, 2019 for an overview of related literature). The HFC intention in compiling male fertility data is therefore to improve their accessibility to researchers and other interested users and to contribute to promoting national and cross-national male fertility analyses.

HFC data users must be aware, however, that HFC fertility data for males, very much like for females, are assembled from different, and not necessarily official, data sources. Although the HFC team performs basic checking, the responsibility for the quality of data lies entirely with the original data producers and providers. For this reason, users are recommended to use these data with caution and to consult methodological documentation accompanying the data, whenever such descriptions are available.[8] As fertility data for males are more likely than those for females to suffer from data quality related problems, such as under-count of births or missing information on the age of the father, the knowledge of the methods possibly applied by the data producer to solve these problems might be of high significance for one's analysis.

Currently, the HFC provides only one type of data for males, which is the period age-specific fertility rates (ASFRs) for all birth orders combined. The processing of male fertility data essentially follows the same data standardization and computation steps as applied for female fertility data (see Figure 4). The few modifications that have been implemented in these procedures are mainly connected to differences in the HFC standard age scale set for data for males. Because of the differences, female and male fertility data are processed separately in the HFC.

Standardization of the age scale and the birth order range. As explained in section 2, all data compiled in the HFC go through a set of common adjustments standardizing the age scale and the birth order range of the original ASFRs. Differently than for females, for which the standard HFC age scale covers the ages ≤14, 15, …, 49, 50+, for males it spreads over a longer period of life and encompasses the ages ≤15, 16, …, 58, and 59+. The standard HFC birth order range includes the categories 1, 2, 3, 3+, 4, 4+, and 5+. However, since the birth order information is not available in

[8] If available, the methodological data descriptions are provided and can be downloaded along with the raw data files.

male fertility data assembled in the HFC so far, the procedures aimed at standardizing the birth order range have been applied only to fertility data for females (for the aggregation procedure applied to the birth order categories in the HFC, see sub-section 2.3).

When the original ASFRs are provided by single years of age but with the age scale surpassing the HFC standard age limits, i.e. going beyond the age ≤15 and/or the age 59+ in case of the ASFRs for males, the original single ages are aggregated to create these standard age groups. Again, if the original age scale includes only one-year age categories but is shorter than the HFC standard age scale for male data (e.g., the first age category is 17 or the last one is 65), the missing estimates at the age tail(s) are assigned the missing values (".") and are assumed to be zeroes in the further computations.

When the original ASFRs are organized by aggregated (e.g., five-year) age groups, the following adjustment procedures can be applied depending on the age structure of the original data:

1. *Splitting of aggregated age groups into one-year age groups.* The *calibrated spline* interpolation (Schmertmann, 2012; 2014) is applied in the HFC for this purpose (see sub-section 2.1).
2. *Splitting of open age intervals at the beginning and/or the end of the age scale into one-year age groups*. In the HFC, *Hermite interpolation* is used for that (see sub-section 2.2).

Computation of aggregated fertility indicators. The standardized ASFRs are compiled into pooled data files, which serve as input data files for the computation of the aggregated fertility indicators: the CPFRs, the TFRs, and the MABs. Apart from differences in the age scale, the computation of these indicators is the same as for females (as explained in section 3). The derived estimates are combined into the country-specific pooled data files (for males), and then into the HFC pooled files (for males).

Data files and their presentation on the HFD website. The raw as well as the output data files, including pooled and single-source data files, are displayed on the HFC website separately for females and males (see sub-subsection 1.5 for details). The structure of all of the HFC data files is explained in the file *Data formats*.

Acknowledgements

The HFD was largely modeled on the successful example of the Human Mortality Database (www.mortality.org), which was developed by the MPIDR and the University of California in Berkeley, and has become a key resource for high-quality mortality data. Similarly, the HFC followed the example of the Human Life Table Database (http://www.lifetable.de/), which was established by the MPIDR in collaboration with the University of California in Berkeley and the INED in Paris.

We are sincerely grateful to Jean-Paul Sardon for his support and advice, and for his invaluable contribution of data from the ODE collection. We also thank Carl P. Schmertmann for providing us with his original software for CS computations, and for taking the time to advise us.

Part of this project was funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement n° 284238 (EURREP).

For language editing we are grateful to Miriam Hils.

References

Bishop, Y., Fienberg, S., and Holland P. (1975). Discrete Multivariate Analysis: Theory and Practice. MIT University Press.

Caselli, G. and Vallin J. (2006). From Situating Events in Time to the Lexis Diagram and the Computing of Rates. In: Caselli, G., Vallin, J., and Wusch, G. (Eds.) *Demography and Synthesis*, Vol. 1. Amsterdam et al.: Elsevier, pp. 55–68.

Dudel, Ch. and S. Klüsener (2019). New opportunities for comparative male fertility research: Insights from a new data resource based on high-quality birth registers. *MPIDR Working Paper WP 2019-023*, November 2019, 58 p.

Fienberg, S. (1970). An Iterative Procedure for Estimation in Contingency Tables. *The Annals of mathematical Statistics*, vol. 41, #3, pp. 907–917.

Jasilioniene A., Jdanov D.A., Sobotka T., Andreev E.M., Zeman K., and Shkolnikov V.M. (2015). Methods Protocol for the Human Fertility Database. Last revision: 02.09.2015. Available at: www.humanfertility.org.

Schmertmann C.P. (2012). Calibrated Spline Estimation of Detailed Fertility Schedules from Abridged Data. *MPIDR Working Paper* WP 2012-022. Available at: http://www.demogr.mpg.de/en/projects_publications/publications_1904/mpidr_working_pa

pers/calibrated_spline_estimation_of_detailed_fertility_schedules_from_abridged_data_4645.htm

Schmertmann C.P. (2014). Calibrated Spline Estimation of Detailed Fertility Schedules from Abridged Data. *Revista Brasileira de Estudos de População* 31(2):291–307. Available at: http://www.scielo.br/pdf/rbepop/v31n2/a04v31n2.pdf

Appendix 1. Notations

	General
x	Age at childbearing
$x_{\min}$	Lowest age at childbearing considered in the analysis
$x_{\max}$	Highest age at childbearing considered in the analysis
	Empirical data
$f(x)$, $f_i(x)$	Unconditional age-specific fertility rates (ASFR) for all of the birth orders combined and by birth order
$\hat{f}(x)$, $\hat{f}_i(x)$	Interpolated age-specific fertility rates for all of the birth orders combined and by birth order produced by the CS estimator
$CPFR(x)$, $CPFR_i(x)$	Cumulative period fertility rate for all of the birth orders combined and by birth order by exact age *x*
TFR, TFR_i	Period total fertility rate based on unconditional age-specific fertility rates for all of the birth orders combined *f(x)* and by birth order $f_i(x)$
MAB, MAB_i	Period mean age at birth based on unconditional age-specific fertility rates for all of the birth orders combined *f(x)* and by birth order $f_i(x)$