Male fertility data for 17 high-income countries
Data documentation and methods

Christian Dudel*[1] and Sebastian Klüsener †[1,2,3]

[1]*Max Planck Institute for Demographic Research*
[2]*Federal Institute for Population Research*
[3]*Vytautas Magnus University*

Last revised: August 6, 2019

Contents

*Corresponding author; address: Max Planck Institute for Demographic Research, Konrad-Zuse-Str. 1, 18057 Rostock, Germany; email: dudel@demogr.mpg.de; phone: +49 381 2081221
†Email: sebastian.kluesener@bib.bund.de

1 General remarks

Fertility rates We provide age-specific fertility rates (ASFRs) for males in 17 high-income countries. In this paper, we describe the data and the methods used to calculate the ASFRs. Specifically, we discuss the register-based birth count data underlying the ASFRs and the methods used to handle the birth count data. For a summary in tabular form, see Table 2 at the end of this document.

Data access The ASFRs can be downloaded for free from the Human Fertility Collection (HFC):

`http://www.fertilitydata.org`

Population exposures In addition to birth counts, the ASFRs are based on population exposures. The population exposures are calculated from population counts, as supplied by and documented in the Human Mortality Database (`http://www.mortality.org`). For Germany, adjusted inter-censal estimates of population counts are used. These have been described elsewhere (Klüsener et al., 2018). The territorial coverage of the population exposures and the birth count data are consistent for all countries.

2 Data: Birth counts

Overview The ASFRs we provide are based on age-specific counts of live births for the following countries and years. The institutions that supplied these counts are given in brackets.

- Australia 1975-2014 (Australian Bureau of Statistics)
- Canada 1974-2016 (Statistics Canada)
- Denmark 1986-2015 (Statistics Denmark)
- England and Wales 1982-2016 (Office for National Statistics)
- Estonia 1989-2014 (Statistics Estonia)
- Finland 1987-2015 (Statistics Finland)
- France 1998-2013 (National Institute of Statistics and Economic Studies)
- Germany 1991-2013 (Federal Statistical Office/Dudel & Klüsener 2016; we also provide separate results for western Germany and eastern Germany)
- Hungary 1970-2014 (Hungarian Central Statistical Office)
- Italy 1999-2014 (Istat)
- Japan 2009-2016 (Statistics Japan)
- Poland 1986-2014 (Statistics Poland)
- Portugal 1980-2015 (Statistics Portugal)
- Spain 1975-2014 (National Statistics Institute)
- Sweden 1968-2015 (Statistics Sweden)
- Taiwan 1998-2014 (National Statistics)
- USA 1969-2015 (National Bureau of Economic Research)

Register data All birth counts are taken from birth registers. Most of the data is based on complete enumeration, except for some years for the United States. The U.S. data prior to 1972 consists of 50% random samples from the births registers of the U.S. states. In the following years, full birth register data gradually becomes available for an increasing number of U.S. states; and starting in 1985 complete birth register data is available for all of the states. For details, see the documentation available at `http://www.nber.org/data/vital-statistics-natality-data.html`; e.g., National Center for Health Statistics (n.d.) for information on the years 1972 to 1977.

Territorial coverage Most registers cover only births to the resident population. For this reason, some births to fathers might be missing if the mother was living abroad, especially if the birth occurred abroad. The opposite case is also possible: i.e., a birth to a mother might be covered by the register while the father is living abroad. It is, however, likely that there were very few such births. Moreover, the number of cases in which either the mother or the father was living abroad should roughly cancel each other out. For Australia and for England and Wales, the birth counts include all births in the corresponding national territory. The first years of the German data do not include the states of Mecklenburg-Vorpommern (missing in 1991 to 1994) and Saarland (missing in 1991; also see Dudel and Klüsener 2016). For Germany we also provide ASFRs separately for western Germany and eastern Germany. Eastern Germany covers all new states (Neue Bundesländer) including Berlin. The French and the U.S. data does not include oversea territories.

Undercoverage In Italy and Japan, not all live births are included in the data. In Italy, births are recorded at the municipality level. Municipalities can either use a long or a short form for registering births. Only the long form includes the age of the father. The data we have is only covering births recorded with the long form, which is used in most municipalities. When we compared this data with data from the Human Fertility Database (HFD), which covers all births, we found that depending on the year, between $6,126$ births and $22,805$ births (or 1% to 4%) are missing. For Japan, the data we have access to covers marital births only. While the proportion of births that are non-marital tends to be rather low in Japan, there are around $23,000$ fewer births (or around 2% of total births per year) in our data than in the HFD data. We corrected both the Italian and the Japanese data for these issues; the method is described in section 5.

Years We have chosen to use all of the years of data that are available to us for all countries, except for England and Wales. For England and Wales, we have data for 1980 and 1982 to 2016. The data for 1981 is incomplete because there was a registrars' strike in that year. We ultimately decided to use only data of the uninterrupted time series from 1982 to 2016.

Age range We provide the ASFRs for the age range 15 to 59. We chose to use 59 as the highest age because male fertility is extremely low after this age. Across all of the countries and years included in our data, the highest proportion of births that were to fathers aged 59+ was 0.2% in Italy in 1999. For most other countries and years, this figure was considerably lower. All of the relatively small number of births to fathers at ages 59+ were assigned to age 59. Note that for some countries, the raw data cover a narrower age

range (England and Wales, France, Germany, Japan, Portugal, Sweden). In these cases, we applied additional methods to distribute the birth counts across the whole age range (see below). For all other countries, data for the full age range of 15 to 59 is available.

Additional methods (ARDY) In the raw data of most of the countries, age is defined as the age at childbirth (in years); while in the Swedish data, age is defined as the age reached during the year (ARDY) in which the birth took place. To estimate the age at birth based on the ARDY data, we applied the method described in the methods protocol of the HFD (Jasilioniene et al., 2015).

Additional methods (missing values) In the data of all of the countries, the paternal age is missing for some births. For these cases, paternal age was imputed (see section 3).

Additional methods (age range) The raw data of some of the countries covers a narrower age range than 15 to 59. This is the case for the data from England and Wales (15 to 55), France (17 to 46), Germany (17 to 59), Japan (17 to 59), Portugal (15 to 49), and Sweden (15 to 50). For England and Wales, France, Portugal, and Sweden we applied the penalized composite link model (PCLM) proposed by Rizzi, Gampe, and Eilers (2015) to split the open-ended age interval for age 55, 46, 49, or 50, respectively (see section 4). For births to fathers under age 17 in France, Germany, and Japan, we applied a simple procedure that distributed births to ages 15, 16, and 17 based on the proportions of births in these ages in other countries (see section 4).

Additional methods (age intervals) In the Portuguese data and the Taiwanese data, age is only available by five-year age intervals. To split five-year intervals into one-year intervals we again used the PCLM approach (see section 4).

3 Methods: Imputation of missing values

Missing paternal age One of the biggest challenges researchers face when analyzing male fertility is dealing with missing values for the age of the father, because in many data sets the paternal age is not recorded for a sizable number of births (Dudel and Klüsener, 2018). For the countries and the years we study, the proportions of missing values range from below 1% (Sweden, 2002), to 47% (Denmark, 1994). For some countries, the proportions also vary considerably over time. For instance, for Germany, the highest proportion of missing values for a given year was 22% in 1999, while the lowest was 7% in 2013.

Conditional approach To deal with missing values, we have chosen to adopt the conditional approach, as discussed by Dudel and Klüsener (2018). Assuming that the maternal age is available for all births, this approach works as follows. $B^*(x,t)$ is the number of births for which the paternal age is observed to be x in year t. The asterisk is used to indicate that this number might be an undercount; i.e., some of the births with missing values might have fathers aged x. Let $B^*(x,y,t)$ denote the number of births to fathers aged x and mothers aged y, which again is only available for births with no missing information. $B(\mathrm{NA},y,t)$ represents the number of births for which the maternal age is known and equals

y, but the paternal age is missing and unknown. $P^*(x|y,t)$ is the paternal age distribution conditional on the maternal age, calculated as $B^*(x,y,t)/\sum_{i=\alpha}^{\beta} B^*(i,y,t)$, where α and β are, respectively, the first age and the last age of the reproductive phase of males; thus, the missing values are ignored. The ASFRs are then calculated as:

$$f(x,t) = \frac{B^*(x,t) + \sum_{j=\gamma}^{\delta} B(\text{NA}, j, t) P^*(x|j,t)}{E(x,t)}$$

where γ and δ denote the youngest and the oldest childbearing ages of women, and $E(x,t)$ is the exposure. For most countries, the age of the mother is always or almost always recorded, and the conditional approach can thus be easily applied.

Unconditional approach For Australia, there are also missing values, but only the marginal distributions of the paternal age and the maternal age are available, not the joint distribution. In this case, we decided to apply the unconditional imputation approach, which works as follows. Let $P^*(x|t)$ denote the proportion of births to fathers aged x, calculated by ignoring missing values; i.e., $P^*(x|t) = B^*(x,t)/\sum_{i=\alpha}^{\beta} B^*(i,t)$. The unconditional approach then calculates the ASFRs as:

$$f(x,t) = \frac{B^*(x,t) + B(\text{NA}, t) P^*(x|t)}{E(x,t)}$$

Since in the Australian data the proportions of missing values for the paternal age are consistently below 10%, the use of the unconditional approach should suffice to achieve reliable estimates (Dudel and Klüsener, 2018).

Imputations by statistical offices In England and Wales and in France, the statistical offices imputed missing age information. In England and Wales, the age of the father was taken from a birth record with otherwise similar characteristics (e.g., age of mother, marital status). This was only done for births for which some information on the father was available, but the paternal age was missing. As there were only few such cases, most of the missing values were not replaced with imputed values by the statistical office (for details see Office for National Statistics, 2017). For these births, we applied the conditional approach, as outlined above. In the French data, all of the missing values were replaced with imputed values by the statistical office. Imputation was done by dividing the births into three groups by age of the mother (24 or younger; 25 to 34; 35 or older) and assuming an average age difference of four years, three years, and two years, respectively. For example, if the age of the mother was 37, then the age of the father was imputed as 39 (Insee, personal communication, 2018).

Missing maternal age The age of the mother is almost always available. In most of the countries studied, the age of the mother is missing for fewer than three births per year. These births were dropped. While it is possible that in these countries the actual numbers of births with a missing maternal age are higher, and that the missing maternal ages were somehow imputed, we have no information that would allow us to determine whether this was the case. Moreover, the actual number of such cases would very likely still be small. But in Canada, Italy, and Portugal, the number of births for which the age of the mother is unknown is relatively high. In the Canadian data, the number of births with no maternal

age is high for the years prior to 1991 (between 2,405 and 12,620 births, or up to 4% of total births), is lower from 1992 to 2011 (between 10 and 500 births; or less than 1% of total births), and is zero from 2012 onward. For Italy, the number of births with missing maternal age ranges from 1,354 (1999; less than 1% of total births) to 12,217 (2002; 2% of total births), with the other years having values between these numbers. In Portugal, the number of births with missing maternal age is below 25 for all years except for 2008, when the number was 980 (less than 1% of total births). In these cases, the following procedure was used. First, for births for which the age of the mother is unknown but the paternal age is available, the conditional approach was applied; i.e., the maternal age was imputed based on the paternal age. Second, for births for which neither the maternal nor the paternal age was available, the unconditional approach was applied to assign the age of the mother. After this step, all births had a maternal age that was either observed or imputed. In a third step, the conditional approach was applied to impute the paternal age for births for which it was missing.

4 Methods: Splitting five-year age intervals and open-ended age intervals

Splitting five-year age intervals For Portugal and Taiwan, data is available in five-year age intervals only. In the literature, several methods for splitting five-year age intervals have been proposed. So far, however, these methods have not been applied to data on males. We applied the piecewise cubic Hermite (PCH) interpolation method described in the methods documentation of the Human Fertility Database (Jasilioniene et al., 2015), as well as a recently proposed approach based on quadratic optimization (QO; Grigoriev et al., 2018) and the penalized composite link model (PCLM) of Rizzi, Gampe, and Eilers (2015), which was implemented by Pascariu et al. (2018). When we compared these methods we found that the use of the PCLM approach produced the most reasonable results for splitting five-year age intervals (see below).

Results (five-year age intervals) To compare the three approaches for splitting five-year age intervals, we aggregated the Spanish single-year age interval data into five-year age intervals using the following intervals: 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50+. Exemplary results based on Spanish data for 2000 are shown in Figure 1 and 2. The results for the other years look similar. As we can see, the PCH interpolation produced some irregularities in splitted ASFRs that seem unlikely. The results of the QO approach appear to be more reasonable, but it creates a hump around age 45 that is not observable in the underlying data. The use of the PCLM method generated the most plausible results. Moreover, unlike the other approaches, this method can be used to seamlessly split the open-ended age interval 50+ into ages 50 to 59. For these reasons, we chose to use the PCLM method for splitting five-year age intervals.

Splitting open-ended age intervals (1) We also chose to use the PCLM approach to split open-ended age intervals, and we applied it to England and Wales (55+), France (46+), Portugal (49+), and Sweden (50+). In some cases, the application of the PCLM method produced schedules that appear to be artifacts. However, these artifacts can be expected to have no or only very minor effects on standard analyses of fertility. Still, at higher ages,

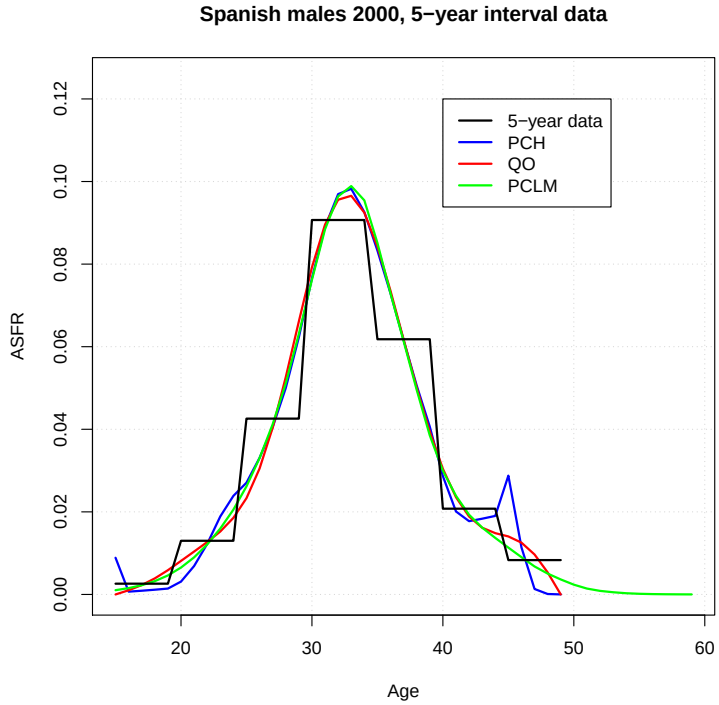

Figure 1: Example of splitted five-year age intervals (Spain 2000); binned five-year age interval data shown as a black line; split of data aggregated into five-year age intervals by piecewise cubic Hermite (PCH) interpolation in blue; split by quadratic optimization (QO) in red; penalized composite link model (PCLM) in green. Source: Own calculations.

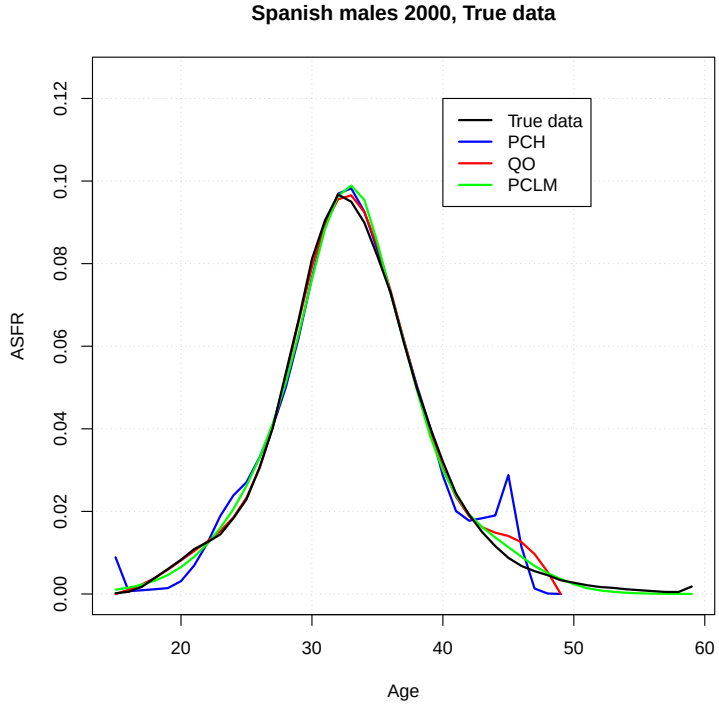

Figure 2: Example of splitted five-year age intervals (Spain 2000); original single-year age interval data shown as a black line; split of data aggregated into five-year age intervals in blue (PCH), red (QO), and green (PCLM). Own calculations.

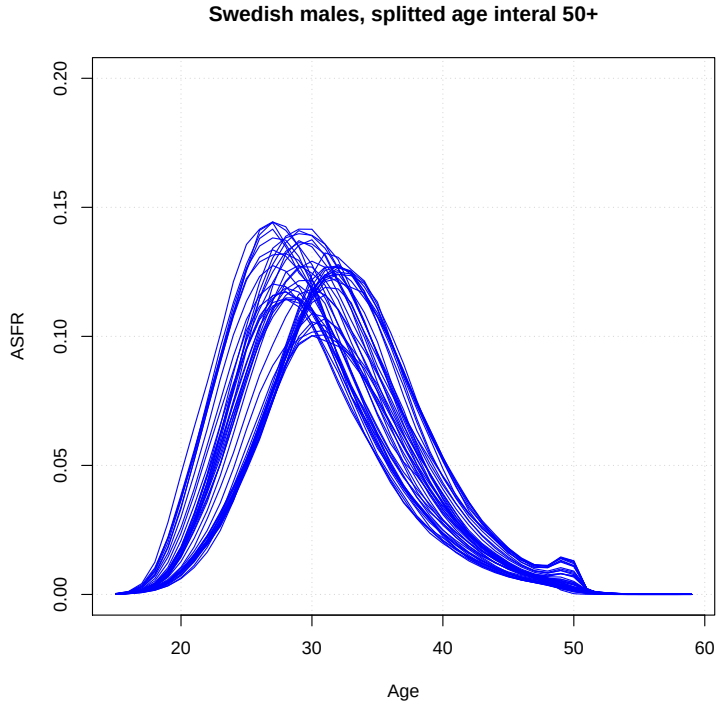

Figure 3: Example of age schedules with splitted open-ended age interval 50+ for Sweden 1968-2015 based on the PCLM approach. Source: Own calculations.

some minor irregularities can be found in our splitted data for England and Wales, France, Portugal, and Sweden. An example that shows the age schedules of Swedish males with a splitted open-ended age interval 50+ is given in Figure 3. Each line represents the age schedule of one of the years from 1968 to 2015. For the years 2005 to 2015, where the frequency of paternal births at advanced ages is relatively high, we can see a small hump around age 50 that is likely an artifact of the PCLM method. It might be attributable to the ceiling effect of the relatively low open-ended age interval (50+), which leads to an increase in the absolute number of births for ages 50+ relative to the absolute number of births for age 49. Thus, it is likely this finding is not a property of the true but unobserved data.

Splitting open-ended age intervals (2) To split births in the open-ended age interval "17 or younger" to ages 15, 16, and 17 for France, Germany, and Japan, we calculated the proportion of births to fathers aged 15, 16, or 17 relative to all births to fathers aged 17 or younger. This was done using the data for all of the countries and years for which we observed ages 15, 16, and 17 in one-year intervals. The proportions are 6% (age 15), 23% (age 16), and 71% (age 17). Births in the interval "17 or younger" were distributed according to these proportions. The numbers of births to fathers in this age group are very low. For example, in Germany in 2013, just 960 births, or less than 0.1% of the total number of births, were to fathers aged 17 or younger.

5 Methods: Correcting for undercoverage

The birth count data we use for Italy and Japan does not cover all births, as described earlier. In both cases we proceeded as follows. Age-specific birth counts for females taken from the HFD were used as a reference, $B^f_{HFD}(x,t)$. We calculated the age-specific birth counts for females for the data available to us, $B^f(x,t)$, and derived the difference to the HFD by age: $D^f(x,t) = B^f_{HFD}(x,t) - B^f(x,t)$. Each age-specific difference was added to the number of births for which the paternal age is unknown and the mother is aged x, $B_{new}(\text{NA},x,t) = B_{old}(\text{NA},x,t) + D^f(x,t)$. The age of the father was then imputed as described in section 3.

6 Consistency with HFD data

Indicators To assess how consistent our data is with the Human Fertility Database (HFD) data, we calculated two indicators. First, we derived the difference in the number of births for each country and year; i.e., $B(t) - B_{HFD}(t)$, where $B_{HFD}(t)$ is the number of births in the HFD and $B(t)$ is the number of births in our data. Second, we calculated the dissimilarity index for the distribution of the maternal age at childbirth for each country and year. If $P_{HFD}(y|t)$ is the age distribution of mothers in the HFD and $P(y|t)$ denotes the age distribution in our data, the dissimilarity index is defined as $D = 100\sum_{y=\alpha}^{\beta} 1/2|P(y|t) - P_{HFD}(y|t)|$. D can attain values between zero and 100. A value of zero means that the two distributions are completely identical. A value of x means that $x\%$ of one distribution would need to be shifted to obtain the other distribution.

Results Table 1 shows for each country the lowest and the highest value observed for both indicators. As Australia is not covered by the HFD, no results are available for that country. For Portugal and Taiwan, no results for the dissimilarity index are shown, as the data is in five-year intervals. For Denmark, England and Wales, Estonia, Finland, Germany, Hungary, Italy, Japan, and the U.S., the differences between the HFD data and our data are small or non-existent, both with respect to the number of births and the age distribution of mothers. In the cases of Italy and Japan, such a result was guaranteed by our treatment of undercoverage. For Canada before 1991 there are differences of up to $10,448$ births per year between the HFD data and our data. This is because the HFD data does not include information on Newfoundland and Labrador (Houle, Kubisch, and Jasilioniene, 2016), while our data does (the population exposures of the HMD also cover Newfoundland and Labrador). For Spain, the largest gap between our data and the HFD is 517 births, but for most years the difference is zero and the data is consistent. For Portugal and Taiwan, the raw data is consistent with the HFD data, but the birth counts implied by the ASFRs show minor differences after five-year intervals are split. The Swedish data also differs somewhat from the HFD data, but it is consistent with numbers published by Statistics Sweden. This may be because the HFD data was derived from a different data source (Historic Population Register). The only noteworthy differences between the HFD data and our data are found for the birth counts of France and Poland. There is no clear explanation for these discrepancies. The age distributions of mothers in our data and the HFD data are comparable for these two countries, though.

Table 1: Consistency with the Human Fertility Database.

Country	*Birth counts*		*Age distribution*	
	Difference		Dissimilarity index	
	min	max	min	max
Australia	—	—	—	—
Canada	-16	10,448	0.0%	0.1%
Denmark	-154	30	0.0%	0.7%
England and Wales	0	0	0.0%	0.0%
Estonia	-5	0	0.0%	0.0%
Finland	-40	29	0.0%	0.0%
France	0	9,377	0.2%	0.3%
Germany (total)	0	0	0.0%	0.0%
Hungary	-12	0	0.0%	0.0%
Italy	0	0	0.0%	0.0%
Japan	0	0	0.0%	0.0%
Poland	-8594	9,650	0.1%	0.1%
Portugal	(0)	(0)	—	—
Spain	0	571	0.1%	0.4%
Sweden	-1724	1,077	0.5%	1.2%
Taiwan	(0)	(0)	—	—
USA	0	0	0.0%	0.0%

Contributors and acknowledgements

Christian Dudel and Sebastian Klüsener created and prepared the data sets, and partly developed the methods needed to do so. They were financially supported by the Laboratory of Fertility and Well-Being and the Laboratory of Demographic Data at the Max Planck Institute for Demographic Research (MPIDR) in Rostock, Germany. The ASFRs for Germany (total/western/eastern) were previously published (Dudel and Klüsener, 2016), and are also available through the website of the journal Demographic Research: `https://www.demographic-research.org/volumes/vol35/53/default.htm`

Sigrid Gellers-Barkmann and Karolin Kubisch were responsible for data searches and for communications with national statistical offices. Angela Carollo, Angelo Lorenti, and Alice Goisis helped with the Italian data. Paul Samula, a student assistant at MPIDR, helped with data preparation. The data is made available through the HFC, which is a joint effort of the MPIDR and the Vienna Institute of Demography (VID) in Austria (`http://www.fertilitydata.org`). All errors in deriving the ASFRs are our own.

Bibliography

Dudel, C. and Klüsener, S. (2016). Estimating male fertility in eastern and western Germany since 1991: A new lowest low? *Demographic Research* 35: 1549–1560.

Dudel, C. and Klüsener, S. (2018). Estimating men's fertility from vital registration data with missing values. *Population Studies*: available online.

Grigoriev, P., Michalski, A.I., Gorlischev, V.P., Jdanov, D.A., and Shkolnikov, V.M. (2018). New methods for estimating detailed fertility schedules from abridged data. MPIDR Working Paper WP-2018-001.

Houle, R., Kubisch, K., and Jasilioniene, A. (2016). Human Fertility Database Documentation: Canada. Available online: https://www.humanfertility.org/Docs/CAN/CANcom.pdf.

Jasilioniene, A., Jdanov, D.A., Sobotka, T., Andreev, E.M., Zeman, K., Shkolnikov, V., Goldstein, J., Nash, E.J., Philipov, D., and Rodriguez, G. (2015). Methods Protocol for the Human Fertility Database. Available online: `https://www.humanfertility.org/Docs/methods.pdf`.

Klüsener, S., Grigoriev, P., Scholz, R.D., and Jdanov, D.A. (2018). Adjusting inter-censal population estimates for Germany 1987-2011: Approaches and impact on demographic indicators. *Comparative Population Studies* 43: 31–64.

National Center for Health Statistics (n.d.). 1972-1977 natality detail tape documentation. Available online: `http://www.nber.org/natality/1972/natl1972.pdf`.

Office for National Statistics (2017). User guide to birth statistics. Available online: `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/methodologies/userguidetobirthstatistics`.

Pascariu, M.D., Rizzi, S., Schoeley, J., and Danko, M. (2018). ungroup: Penalized composite link model for efficient estimation of smooth distributions from coarsely binned data. R package version 1.0.0.

Rizzi, S., Gampe, J., and Eilers, P.H.C. (2015). Efficient estimation of smooth distributions from coarsely grouped data. *American Journal of Epidemiology* 182: 138–147.

Table 2: Summary of the age-specific fertility rates we provide by country. For details of the imputation approaches, see section 3. For the "other adjustments" see sections 2, 4, and 5.

Country	Years	Age range (raw data)	Imputation approach	Maternal age missing	Other adjustments
Australia	1975-2014	15-59	Unconditional	–	No
Canada	1974-2011	15-59	Conditional	Yes (see section 3)	No
Denmark	1986-2015	15-59	Conditional	No	No
England and Wales	1982-2016	15-55	Conditional/other (see section 3)	No	See section 4
Estonia	1989-2014	15-59	Conditional	No	No
Finland	1987-2015	15-59	Conditional	No	No
France	1998-2013	17-46	Other (see section 3)	No	See section 4
Germany (total/western/eastern)	1991-2013	17-59	Conditional	No	See section 4
Hungary	1970-2014	15-59	Conditional	No	No
Italy	1999-2014	15-59	Conditional	Yes (see section 3)	See section 5
Japan	2009-2016	17-59	Conditional	No	See sections 4, 5
Poland	1986-2014	15-59	Conditional	No	No
Portugal	1980-2015	15-49	Conditional	Yes (see section 3)	See section 4
Spain	1975-2014	15-59	Conditional	No	No
Sweden	1968-2015	15-50	Conditional	No	See sections 2, 4
Taiwan	1998-2014	15-59	Conditional	No	See section 4
USA	1969-2015	15-59	Conditional	No	No