APPENDIX 3: Inclusion criteria and possible systematic differences between surveys

Selection of surveys for the analyses: criteria considered and applied

We included all relevant European multi-country surveys conducted since 1979 asking about ideal family size. In addition, we included country-specific surveys which give additional evidence for countries with lower coverage (e.g., Switzerland, Albania, Ukraine) or where additional empirical evidence was important for discussing the earlier findings on family size ideals (e.g., Austria). Our initial dataset covered 202 surveys conducted in 43 European countries and territories (considering Northern Ireland, eastern and western Germany as distinct units) over the period 1979-2012. An inspection of the data revealed that the surveys conducted in smaller countries often had low sample size (in five cases including fewer than 100 women aged 15-49), which implied higher uncertainty about the results and also more instability in the time trends of ideal family size for these countries. Consequently, we decided not to consider countries with population below 1 million (Cyprus, Luxembourg, Malta, Montenegro, Iceland) as well as the territory of Northern Ireland, where the sample size was low in all surveys (with fewer than 100 respondents in three out of seven available surveys). This resulted in excluding 21 surveys and reducing our dataset to 181 surveys in 37 countries.

Next we considered criteria for addressing high share of non-response, non-numerical responses and "do not know" answers concerning ideal family size in some datasets, which signalled potential problem with the quality of these surveys. We considered three alternative selection criteria, based on the total share of missing, uncertain and non-numerical responses—with the thresholds for excluding datasets set at 25%, 18% and 12%. An analysis of these alternative selection criteria led to the decision to exclude all surveys with more than 18% of missing or non-numerical responses. This implied excluding another 13 surveys conducted in 7 countries (these surveys are listed in Table A2 above); however, no additional country was excluded from our dataset: for each country with excluded surveys other surveys were available that met our selection criteria. This brought the number of analyzed surveys to 168 in 37 countries.

Our selection of an 18% threshold was based on the following main findings and considerations

- *Considering the 25% threshold*:

These cases were clear outliers and concerned only the Eurobarometer surveys (2001, 2006). We identified and excluded four surveys conducted in three countries (Austria 2006, Ireland 2001 and 2006, the Netherlands 2006) with 25-32% unknown or non-numerical responses.

- *Considering the 18% threshold:*

  This implies excluding additional nine surveys, in addition to those excluded on the basis of the 25% threshold. Eight of them were EB surveys (2001: Austria, Finland, western Germany, the Netherlands; 2006: Denmark, UK, 2011: Ireland, UK) and only one was a different survey (EVS 1981 in Denmark). Different considerations led to our decision to exclude these surveys. The surveys with such a high proportion of missing or non-numerical responses often showed indicators of ideal family size differing from the other surveys in individual countries (e.g. in Austria), which might be induced by non-response bias. Also the finding that for each country concerned there were other surveys (often including EB surveys conducted in different years) showing lower share of non-response and non-numerical responses suggested there might be data quality issues in some of these cases.

- *Considering the 12% threshold:*

  We would exclude additional 20 surveys conducted in 10 countries if we applied yet stricter selection criteria based on the 12% threshold. While some surveys within this range also show values that are not in line with the other surveys in a given country, signalling potential data quality problems (e.g., the 2001 EB survey for eastern Germany), non-response and "do not know" answers around 15% or lower were relatively common in other than Eurobarometer surveys. They were consistently reported in some countries (e.g., Belgium, Denmark, western Germany, the Netherlands, Sweden), providing an additional justification for our choice of a less-restrictive 18% threshold.

All in all, applying the 18% threshold had an important effect on the time series for Ireland, where all three EB surveys since 2001 have been excluded from our analysis. This implies that no data for Ireland after 1994 were included in our analyses (see also below).

Sensitivity analysis: are the presented results robust to different selection criteria?

Table A3 presents summary statistics for all considered surveys combined on the basis of different selection criteria applied. The table shows that the key results, when presented for all countries combined are not sensitive to the selection criteria applied. There is a slight upward trend in the (unadjusted) share of women with two-child ideals when stricter selection criteria are applied. This is to be expected: as surveys with high share of unknown and non-numerical answers are eliminated, the share of respondents

with specific ideals, including that of having two children, should increase. There is no similar trend in the series of data adjusted for non-response (last three columns in Table A3). More detailed analysis not presented here showed in addition that also the time trends of the key analyzed indicators were not sensitive to the selection criteria chosen. However, the exclusion of selected dataset did affect the results and trends for individual countries, especially Austria, Ireland, the Netherlands and the United Kingdom.

Table A3: Summary results for all surveys considered based on different selection criteria. European countries 1979-2012, women aged 15-49.

							Adjusted for non-response		
Selection criteria	# surveys	# countries	Average MIFS	Median MIFS	Std. dev. MIFS	Average share with 2-child ideal (%)	Average share with ideal 0+1 (%)	Average share with ideal 2 (%)	Average share with ideal 3+ (%)
All surveys	202	43	2.38	2.36	0.30	53	7	58	35
Countries with pop > 1 mill	181	37	2.36	2.36	0.28	54	7	59	34
Threshold 25%	177	37	2.36	2.36	0.28	55	7	59	34
Threshold 18%	168	37	2.36	2.35	0.28	55	7	59	34
Threshold 12%	148	36	2.38	2.36	0.28	56	7	59	35

Besides inspecting the aggregate indicators, we also studied how different exclusion criteria affected the observed maximum and minimum values. We found considerable influence of the exclusion of the EB datasets for Ireland, which reduced the observed maximum MIFS values in the most recent analyzed periods (1998-2012) by 0.04-0.16 in absolute terms. In the latest period considered, 2008-12, the highest MIFS value in our study, 2.54 in Albania, was well below the excluded value of 2.70 for Ireland. The observed minimum MIFS value was affected only in one period, 2005-7, by an exclusion of the Austrian EB 2006 survey with a MIFS of 1.73, well below the lowest MIFS in that period among the selected datasets (1.93 for Romania) and also below the lowest MIFS in the whole dataset (1.78 in eastern German 2001 EB survey). Finally, the exclusion of EB data for Ireland pushed the lowest observed shares of women with a two-child ideal to considerably higher levels. The exclusion of EB datasets for Ireland has thus clearly affected the reported "extreme" MIFS values. In the absence of other surveys in that country since the mid-1990s we cannot speculate to what extent the EB survey for Ireland might have been affected by the high share of non-response and "do not know" answers on ideal family size.

However, even if we did not exclude the listed datasets, the trends over time in maximum MIFS values and the lowest share of women with 2-child ideals, would remain similar. Only the trend of the lowest MIFS value would change following a U-shaped pattern with a notable increase in the latest observed period.

Possible systematic differences between surveys

Finally, we also analyzed whether there might be systematic differences between surveys with slightly different question wording on ideal family size, which might have affected our results. Appendix 1 shows that a distinction can be drawn between the EVS and WVS surveys on one side (where "if any" was added to "how many children" and the EB, ISSP, and country-specific surveys on the other side. We analyzed possible systematic differences by identifying different surveys for the same country that took place within a three-year interval or within one of the periods we focused on in our study (see Table 1) and comparing basic indicators of ideal family size for these consecutive surveys. We identified 30 such "pairs" of surveys in our sample. Overall, no systematic differences were found. The EVS/WVS surveys showed on average somewhat higher MIFS values (2.44) than the other surveys (2.35) and the share of women with two-child ideals was somewhat lower in the EVS/WVS (average of 54.7% vs. 56.4% in other surveys). Part of this difference was due to low ideal family size reported in 1979 EB surveys. When we excluded these EB 1979 datasets and analyzed only 23 remaining "pairs" of surveys, the difference between the compared surveys diminished. The average MIFS values were 2.38 in the WVS/EVS surveys and 2.33 in other surveys; whereas the average share of women with two-child ideals was identical (57%) in both types of surveys. This indicates that the wording of the question on general ideal family size had very little, if any, influence on the analyzed data.