STUDIA DEMOGRAFICZNE 1–2 (157–158) 2010

Krzysztof Tymicki
Institute of Statistics and Demography
Warsaw School of Economics

VALIDATION OF DATA QUALITY FROM POLISH FERTILITY SURVEY 2002 WITH USE OF COHORT FERTILITY RATES

INTRODUCTION

The majority of demographic research investigating reproductive behavior is nowadays based on data from retrospective surveys. Although surveys, *ex definitione*, are aimed at being representative for the whole population with respect to major characteristics such as place of residence, age, sex and education, they may sometimes provide biased fertility trends. Recently some authors have started to investigate validation of survey data with respect to their accurate representation of fertility trends on a population level (Kreyenfeld et al., 2010; Murphy, 2009). Their research clearly shows that there may be a serious bias of fertility patterns estimated with use of retrospective data in comparison to the population trends. This bias may result not only in under- or over- estimation of fertility (Murphy, 2009) but it may also indicate completely different patterns of fertility (Kreyenfeld et al., 2010). Therefore, since micro-level retrospective surveys are used for modeling reproductive behaviour of individuals, biased fertility patterns in the sample might lead to erroneous results and predictions of the models based on the survey data.

Taking into account the importance of data quality from retrospective surveys for proper inference based on micro-models of fertility, this paper aims to validate a data quality from the Polish Fertility Survey (FS 2002) conducted during the National Population Census (NC) of 2002. The data quality will be assessed through a comparison of cohort fertility rates reconstructed with use of FS data with the cohort fertility rates for the total population of females based on registered births data collected by the Central Statistical Office (CSO). Therefore, the main research question of the paper is whether the data from FS 2002 accurately represent cohort fertility trends in Poland. The answer to this question is crucial since the FS data could serve as a high quality source of information for building micro models of females' reproductive behaviour in Poland.

In general, deviation of fertility rates calculated with use of retrospective data from rates based on vital statistics (registration of births) can be framed within the conventional classification of sources of bias occurring in random samples (Weisberg, 2005):

– sampling error (due to the nature of random sampling);

– selection bias (some units have a different probability of selection than originally assumed);

– coverage bias (due to under- or over-coverage of individuals in the sample as compared to the population);

– non-response bias (missing information due to lack of answer);

– measurement bias (both due to respondents and to interviewers).

However, it has to be noted that in the case of surveys on a retrospective scheme, some sources of bias are more relevant than in the standard sample survey. Since the main aim of retrospective fertility surveys is to report precisely the reproductive histories of sampled females, a bias between reported fertility and fertility on the population level may be particularly sensitive to coverage, non-response and selection biases.

The coverage bias is mostly related to the fact that some units are more unlikely to appear in the sample due to their characteristics. An extreme example would be a telephone survey that does not take into account households without a telephone, which results in a biased sample. In a retrospective fertility survey one might expect under-coverage of childless individuals since usually the sampling unit is a household, and households with children are easier to reach than those of a single person or a couple without children (Kreyenfeld et al., 2010, Paradysz 1989). This, in turn, might result in higher observed fertility in a survey sample than in the population. However, a contrary example is given by Murphy (2009).

A second problem is related to the non-response bias which might result in underreporting of fertility in a retrospective survey. The non-response bias in retrospective questionnaires is mainly due to problems with recalling events. As reported in many studies (e.g. Beckett et al., 2001; Hayford and Morgan, 2008) failure to recall a distant event might constitute a major source of missing data. This is particularly true in the case of such events as cohabitation (Hayford and Morgan, 2008) or job histories (dates of entering and exiting the labor market). However, this does not seem to apply to recall of events related to births. Virtually all respondents (usually females) are able to recall their children's dates of birth (see information for German GGS: Kreyenfeld et al., 2010). However, the response rate can be significantly lower for males, who may have more problems remembering the dates of births of their offspring (Rendall et al., 1999).

A third issue leading to bias in the fertility patterns estimated on retrospective data is selection on survival, which might constitute a serious problem mainly for the reconstruction of older cohorts' fertility. For this reason, even in surveys that contain information about the fertility of females aged 85 and older it would be unreasonable to assume that reported fertility of survivors over 85 is representative for the whole cohort born 85 years ago. Probably only a small fraction of those females survived to the age of 85 and there still might be a differential survival with respect to fertility, resulting in upward or downward bias. Thus it is advisable to select a good cutting point for selecting cohorts with a high proportion of survivors.

The abovementioned sources of bias are difficult to remove from the survey sample. Therefore, it seems natural that fertility measures based on survey data will deviate from the population indicators. However, it is important to eliminate all other sources of possible bias (resulting from a sample structure) in order to minimize deviations and make sure that fertility trends calculated from survey data overlap with patterns on the population level.

In Poland these issues have been extensively studied by Paradysz (1989, 1992, 2000, 2002). However, due to the lack of a suitable benchmark (here: individual-level data from registration of births) it has not been possible to conduct a comparative analysis which could provide an answer to the question: to what extent is it reasonable to use retrospective data in order to reconstruct cohort and period fertility rates?

DATA AND SAMPLING

The original file from the Fertility Survey 2002 provided by the Central Statistical Office contained 264845 observations. The core questionnaire included questions about reproductive histories (up to the 20th birth) as well as union histories. It additionally comprised information about the woman's date of birth, level of education (at the moment of the survey), fertility intentions, size of the place of residence (at the moment of the survey) and voivodship (at the moment of survey). Although the number of variables is restricted, the main advantage of the database is its size and time coverage. Data from the FS 2002 cover females born between 1896 and 1986; however, in order to avoid selection bias, it is advisable to analyze data for cohorts born 1945-1986. The database obtained from the CSO has been cleaned, so there were no missing birth dates for either mothers or children. Additionally, potential missing cases with respect to the characteristics of the household and respondents were filled up using the main questionnaire of the census (both were run in parallel). In general, there was no information concerning birth histories for only 1.2% of females.

According to the information obtained from the CSO, the sample was drawn using a two-stage procedure. Since the survey took place during the National Population Census, the first sampling level contained census districts (27000 districts assumed for sampling), while in the second stage the sampling units were dwellings (280000 dwellings assumed). In the first stage, a stratified sampling scheme was applied. Within the strata, units were sampled with a probability proportional to their size (number of dwellings in a given district). In the second stage, units were drawn in a simple random sampling scheme.

This sampling scheme was applied in subpopulations selected from rural and urban parts of the voivodships, within which strata were defined as towns[1] in urban areas, and as counties (*powiat*) in rural areas.

The final sample prepared for fieldwork contained 276775 observations. When we compare this figure to the number of observations in the original database, it turns out that in the final file there are 11930 observations less than in the original sample (which is around 4% of the original sample). Presumably this difference is due to incomplete questionnaires, missing cases or simply problems with reaching the units of observation.

Additionally, for each observation the database contained weights which were equal to the inverse probability of being sampled in the respective stratum. The value of the inverse probability weight (*p-weight*) could be interpreted as equivalent to the numbers represented in the general population. The purpose of the inverse probability weights was making the sample representative for the population with respect to residence and age of the surveyed females.

In the analysis presented here we use the original sample of 264845 observations as well as the sample weighted using the inverse probability weights. Here, the weighting procedure results in augmenting of the original sample using inverse probability weights. Each case has been multiplied according to the value of its inverse probability weight. As a result the weighted file contained 15 992 004 observations. However, for the analysis of cohort fertility rates only the cohorts born between 1945 and 1986 were selected due to possible selection bias for older cohorts.

In order to assess the quality of the FS data, we use the National Population Census data with respect to major characteristics of females. Firstly, we compare distributions of such traits as: age, residence (both in terms of size and voivodship) and education in NC, and weighted and non-weighted data from FS. This comparison aims to ascertain the possible biases between the sample and the population. As a measure of discrepancy between the relevant distributions in the NC data and the FS a half sum of absolute differences between relative frequencies of two distributions is applied. This simple measure could be interpreted as the

[1] In the five biggest towns of Poland strata were defined as quarters of the town.

percentage of observations which have to be relocated between two distributions in order to make them identical.For every variable under analysis we compare the distributions obtained from weighted and non-weighted FS data with the respective distributions from the NC.

Secondly, we compare cohort fertility rates based on weighted and non-weighted data from the FS with population values calculated using vital statistics and registration of the births (see: Holzer-Żelażewska and Tymicki, 2009). Both comparisons aim to answer the main research question: whether cohort fertility patterns for females surveyed in the FS 2002 are consistent with population trends.

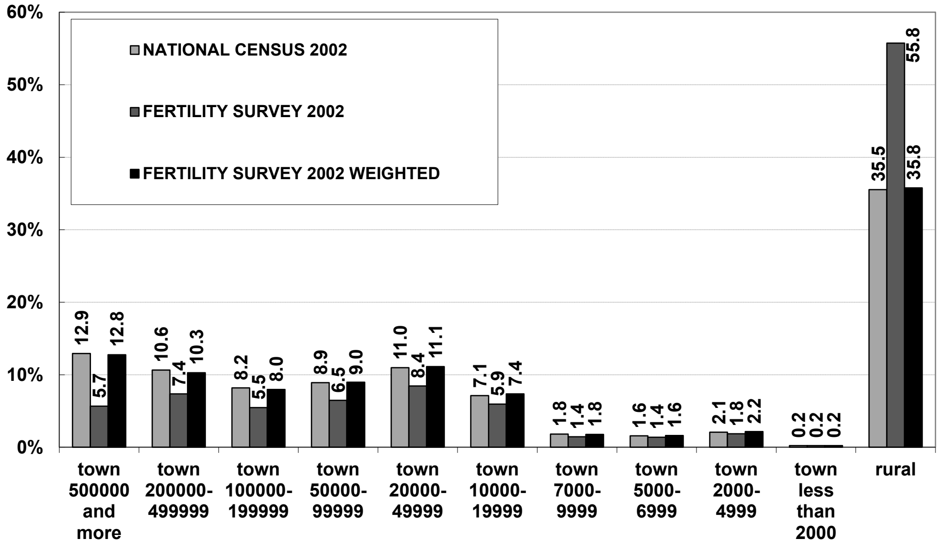
MAIN VARIABLES: THE FS AND NATIONAL CENSUS COMPARED

Fertility rates are sensitive to effects caused by the age, place residence and educational structure of the population or the sample. It is therefore essential to validate FS data with respect to those main characteristics. Any significant deviation in the sample from the population structure, which is not a random error, might obscure true fertility trends. Here, we compare the structure of the FS sample with data from the National Census with respect to age, residence and education. Both in the case of NC data and FS data, we compare distributions for females between 16 and 100 years of age at the moment of the survey or census.

PLACE OF RESIDENCE SIZE

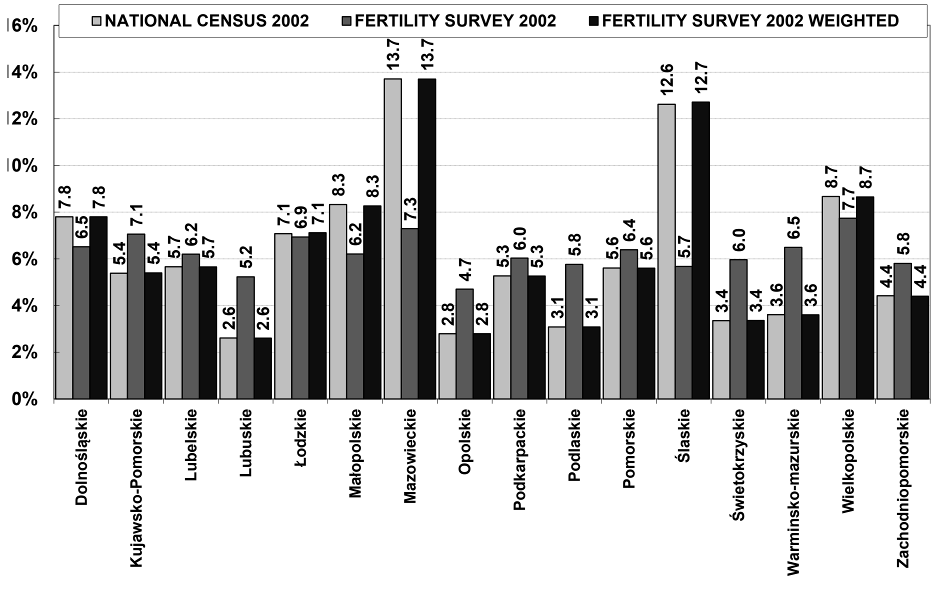
Place of residence can be analyzed in two ways: with respect to size (a variable with 11 categories) and as a geographic location represented by the voivodship. Figure 1 and Figure 2 provide comparisons for both variables in terms of values from the National Census, and non-weighted and weighted values from FS.

Figure 1. Distributions by place of residence size at the time of interview: NC vs FS data



Source: own calculations.

Figure 2. Distributions by voivodship of residence at the time of interview: NC vs FS data



Source: own calculations

With respect to size of place of residence, the non-weighted structure of data from the FS shows a significant deviation in comparison to the values from NC. This is particularly clear in rural areas. In the FS sample over 55% of women are from rural areas whereas the corresponding value from the NC is only 35%. In Figure 1, we can also see an underrepresentation of females from towns in almost every category. Such biases might lead to disturbances in fertility rates due to the fact that females living in rural areas usually show higher fertility than women residing in urban areas. However, if we compare the proportions for the weighted data, percentages for all categories are adjusted back to the level of the population. This is not a surprise since the inverse probability weights were built in order to account for deviations in place of residence of the surveyed females. In terms of the measure of difference, the deviation of non-weighted FS data from the population distribution is 0.2022 but only 0.0081 for weighted data (compare Table 1[2]). Thus, the weights applied are effective in reducing deviations between FS and NC data with respect to size of residence.

Table 1. Discrepancies between distributions based on National Census (NC), non-weighted Fertility Survey (FS) and weighted Fertility Survey

	Difference*	
	NC vs non-weighted FS	NC vs weighted FS
	One way distribution	
place of residence size	0.2022	0.0081
voivodship	0.1784	0.0016
age	0.0355	0.0249
education	0.0547	0.0258
	Two-way distribution	
age x place of residence size	0.2023	0.0288
age x voivodship	0.1812	0.0204
age x education	0.0764	0.0452

**Measured by half sum of absolute differences between frequencies in two distributions*

Source: own calculations.

In terms of regional distribution, the non-weighted FS sample shows a striking underrepresentation of females from Mazowieckie and Śląskie voivodships and an overrepresentation of females from Lubuskie, Opolskie, Świętokrzyskie and Warmińsko-Mazurskie. The overall measure of difference between NC data and non-weighted FS data is almost 0.18. The weights applied reduce the difference

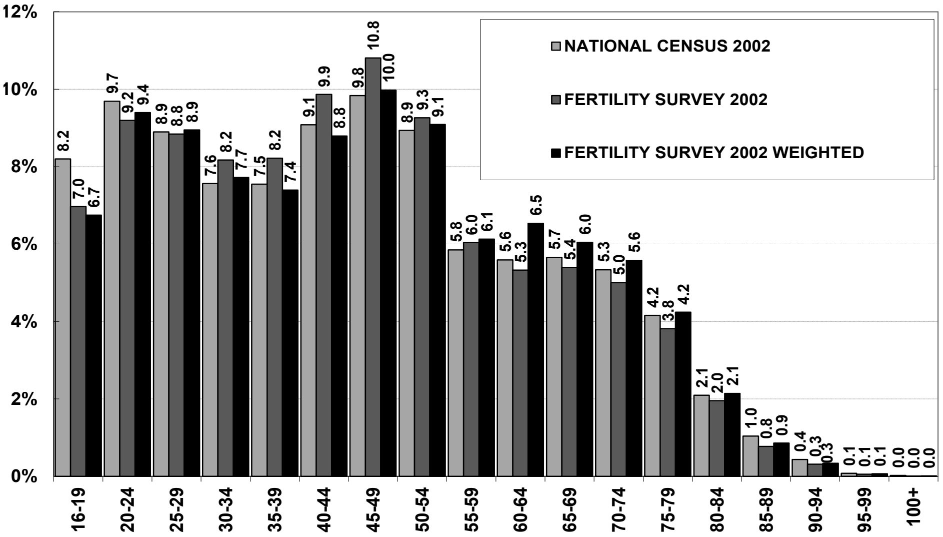
[2] This means that in case of non-weighted data we would have to displace around 20% of observations in the original database while in the weighted database the corresponding figure is only 0,01%.

almost entirely, to 0.0016. The effect of disturbances in the structure of the sample with respect to voivodship on fertility rates is difficult to assess. In 2002 underrepresented voivodships had a lower TFR in comparison with the country level[3]. Among the overrepresented voivodships the picture is less clear. Lubuskie and Opolskie had the lowest TFR rates nationwide whereas other overrepresented voivodships were above the level for the whole country. Therefore the effects of over- and under- representation in distribution of voivodships might cancel each other out for fertility rates on the country level.

AGE

The age structure of the sample shows much less deviation from NC data than the comparison by place of residence. The measure of difference between the age distribution in NC and non-weighted FS data presented in the Table 1 is 0.0355 whereas for weighted distribution of FS data it drops to 0.0249. Only a little change of the measure of difference could be due to the shift in the age pattern caused by weighting. The application of weights shifts the age distribution slightly to cover a higher percentage of females from older age groups (between 60 and 79). On the other hand, non-weighted data exhibit a slight overrepresentation of females between 35 and 49 years old. Thus, it might be concluded that despite weighting, a slight deviation from NC data with respect to age distribution is still present although its magnitude should not affect values of calculated fertility rates.

Figure 3. Distributions by age of respondents at the time of interview: NC vs FS data



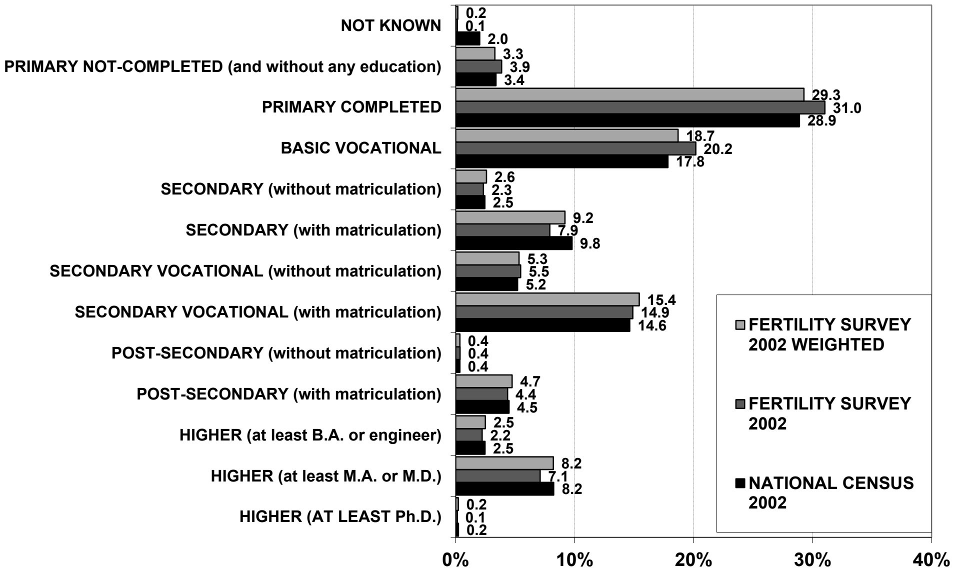
Source: own calculations

[3] It has to be noted that in 2002 and 2003 the lowest period TFR in the history of Poland was recorded.

EDUCATION

The level of education at the time of interview was measured in 12 categories (excluding „not known”). The overall deviation of the non-weighted FS distribution from the NC distribution was 0.0547, resulting mostly from an overrepresentation of females with completed primary and basic vocational schooling and a slight underrepresentation of females with secondary education and higher education. These differences were most likely caused by the biased place of residence structure. The application of weights to the FS data reduces these differences to the level of 0.0258 resulting in a distribution of educational groups not significantly different from that contained in the NC data.

Figure 4. Distributions by education level of respondents at the time of interview: NC vs and FS data



* secondarv school certificate
Source: own calculations.

TWO-WAY DISTRIBUTIONS

In addition to the analysis of distributions of populations under study by core variables, two-way distributions were compared: by age and residence, by age and voivodship and by age and education, in order to find out the accuracy of the applied weights in adjusting the FS sample distributions to the NC population ones. These differences are presented in the three bottom rows of Table 1 and clearly show that the weights significantly reduce the skewed structure of the original sample.

COHORT FERTILITY RATES

Taking into account the structure of the FS data presented above, it is crucial to test whether cohort fertility rates calculated with the non-weighted FS data and weighted FS data produce patterns which are consistent with those of cohort fertility based on registered births.

In order to validate cohort fertility rates obtained from the FS database, we use cohort fertility rates calculated for the whole population for birth cohorts 1945-1985 (total cohort fertility rate) and for cohorts 1955-1985 (by birth order). The population rates are based on aggregated data on vital statistics and individual data from registered births. These data were previously used to study cohort and period fertility rates in Poland by Holzer and Holzer-Żelażewska (1997) and Holzer-Żelażewska and Tymicki (2009).

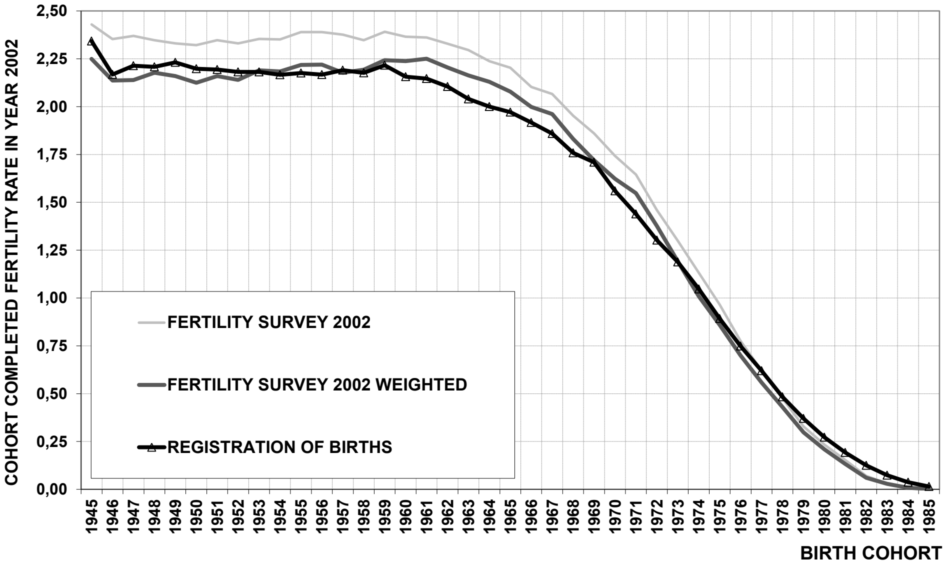
Another way of testing the validity of the FS data would be a comparison of cohort fertility rates clustered by place of residence (urban vs rural). Since the FS data are skewed with respect to the place of residence (and only to a limited extent with respect to other variables), we might expect that cohort fertility rates calculated separately for rural and urban areas using the FS and registration data should not differ significantly. However, this method of validation could not be applied since there are no data which would allow cohort fertility rates from registration data to be calculated separately for rural and urban areas.

The cohort fertility rates were calculated separately for weighted and non-weighted FS data: non-weighted data covered 264845 observations while the weighted database contains 15992004 observations. The calculations concerned cohorts born between 1945 and 1985 in order to avoid the selection bias for older cohorts and to achieve an overlap with the time period covered by the reconstruction of cohort fertility based on registered births. Another limitation was the fact that cohort fertility rates by parity could be only compared for cohorts born 1955 and after. This was caused by the fact that in the case of registration data there was no information about births by parity for earlier cohorts. For the FS data, cohort fertility rates were calculated using the distribution of women by birth cohort, the distribution of births by maternal age at birth (from 15 to 45) and the women's year of birth. The rates were calculated separately for parities: one, two, three, and four and higher.

TOTAL COHORT FERTILITY RATE

For cohorts born between 1945 and 1985, it was possible to compare total cohort fertility rates calculated for the FS data with rates based on registered births. As we can observe in Figure 5, the gap between non-weighted data from the FS and the data from registration of births seems quite wide especially for older cohorts born before 1970. This gap steadily decreases for subsequent birth cohorts and there are no differences to be observed among the youngest cohorts. The difference might be caused by an overrepresentation of females from rural areas in the non-weighted dataset. The difference among older cohorts in particular might be due to larger overall differences in fertility of females from rural and urban areas among older cohorts. These differences decrease for subsequent birth cohorts, which results in lower differences in cohort fertility based on the non-weighted FS data and registration data among younger cohorts. Thus, the convergence in the fertility levels of females from rural and from urban areas might be partially responsible for the decline in the difference between the two curves.

Figure 5. Total cohort fertility rate in 2002: comparison between values from birth registration and weighted and non-weighted FS data (cohorts 1945-1985)



Source: own calculations.

Application of the inverse probability weights to the FS data reduces the gap as compared with data from registration of births. The reduction is mostly apparent for older cohorts, born before 1970. Since the weights were aimed at adjusting the structure of the FS data to the population with respect to place of residence, it is clear that such a procedure should decrease observed level of cohort fertility where there is an overestimation due to the skewed sample structure. Among cohorts born after 1970, weights do not seem to have a large impact on the curve representing total cohort fertility, as the overall difference between rates based on the FS data and register data for first and second births is small. This might also result from the fact that, due to age at interview, younger cohorts were mainly at risk of first birth, and only to a limited extent at risk of higher-order births.

Although weighting significantly reduces the difference between the FS and registration data, it has to be noted that there are no significant deviations from the overall shape of the cohort fertility curves based on the two datasets. Despite the skewed structure of the original sample, the pattern of fertility derived from the non-weighted FS data seems to provide a fairly good approximation of the real fertility trends. Compared with other studies (e.g. Kreyenfeld et al., 2010), total cohort fertility rates calculated using the non-weighted and weighted FS databases do not significantly deviate from population trends observed in Poland.

COHORT FERTILITY RATES BY PARITY

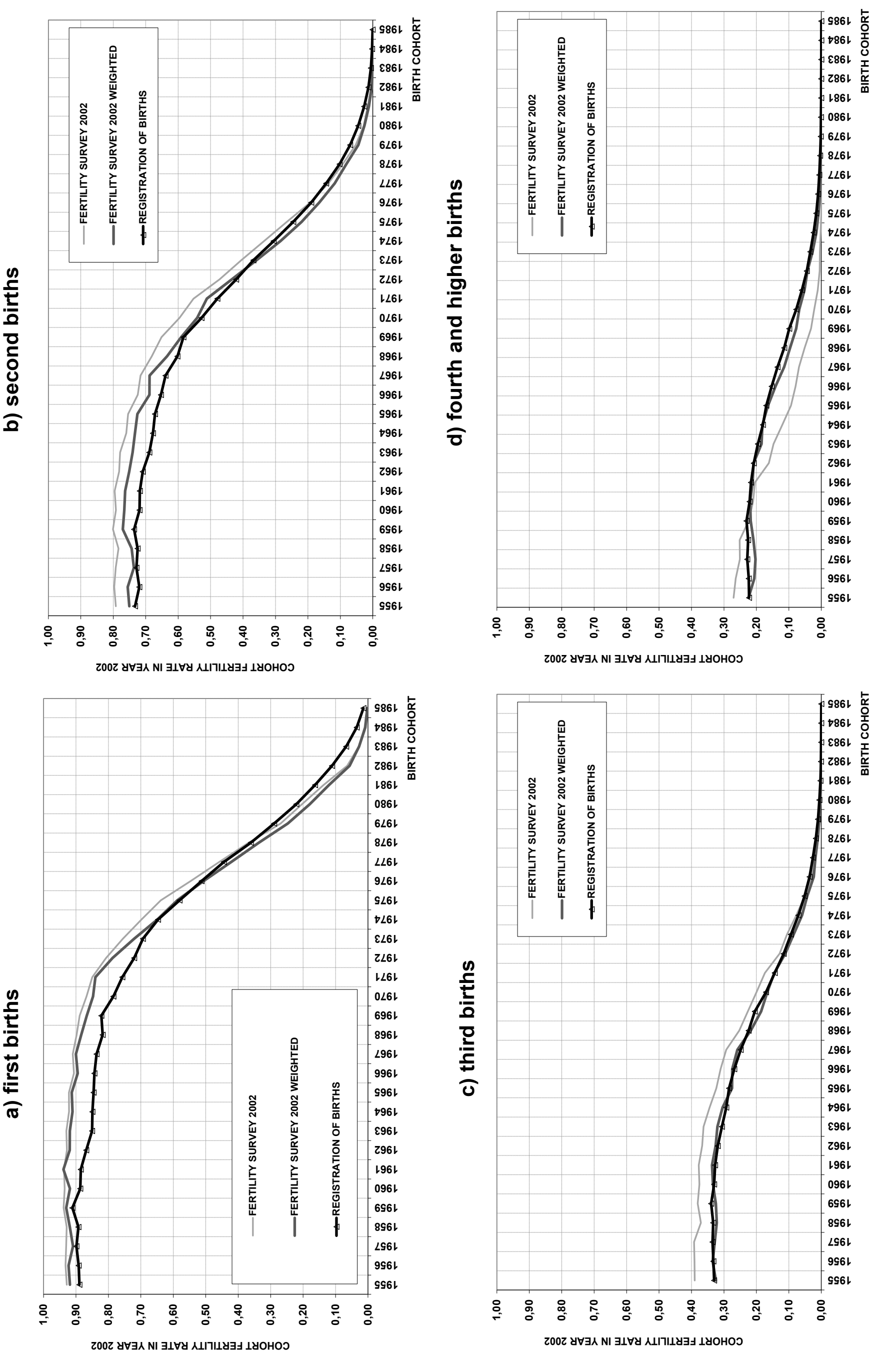
A more detailed insight into differences between cohort fertility rates estimated on the registered data and FS data can be obtained by analyzing cohort fertility rates by parity. The analysis takes into account cohort fertility rates by parity only for females born between 1955 and 1985 due to limitations of the registration data: it is not possible to break down the total fertility rates by parity for cohorts born before 1955.

Irrespective of parity, the non-weighted FS data overestimate cohort fertility, although the magnitude of the difference is moderate and can be noticed mostly among older cohorts. The overall trend for all parities demonstrates that application of weights to the FS data improves estimated fertility rates at higher parities (third, and fourth and higher-order births). Application of weights to the original FS data does not change the shape of the cohort fertility curves for first births.

Both weighted and non-weighted FS data produce higher cohort first birth rates than that based on the registration data. We can also observe that in the youngest cohorts, the FS data (both weighted and non-weighted) slightly underestimate actual cohort fertility. With respect to second births, applying weights shifts the curve down and decreases the difference between cohort fertilities obtained from the FS and the registration data.

Higher-order births (third, and fourth and higher) are significantly overrepresented if we use the non-weighted FS data. The application of weights shifts the fertility curves down to the level of rates calculated from the registration data. If we decompose the absolute difference in total cohort fertility rates (for cohorts 1945-1970) between the weighted and non-weighted FS data, almost 65% of the difference (ranging from 74% for cohort 1945 to 37% for cohort 1970) could be attributed to third and higher-order births. For younger cohorts (born after 1970), the share of the difference due to higher-order births decreases

Figure 6. Completed cohort fertility rate in 2002 by parity: comparison between values from registration of births and weighted and non-weighted FS data. (cohorts 1945-1985)



significantly. This is probably related to the overall decline in higher-order births in the population, which results in a lower proportion of females who reach parity three and higher. First and second order births contribute only a small fraction of the absolute difference between the weighted and non-weighted FS data.

Moreover, a close overlap of the parity-specific cohort rates for cohorts born after 1970 results from the fact that higher-order births occur later in life. Therefore individuals who were around 32 years of age or less on the interview date simply did not experience or were not at risk of higher-order births, and were not covered in the FS data.

DISCUSSION

The main aim of this paper was to validate data from the Fertility Survey 2002 through a comparison of cohort fertility rates based on these data with cohort fertility rates derived from registration data. This comparison aimed to answer the question whether data from the Fertility Survey 2002 accurately describe fertility trends among Polish women and therefore can be used for modeling on an individual level using various techniques such as the event history analysis.

The analysis has shown that due to the skewed sample structure (overrepresentation of females from rural areas), cohort fertility is higher in comparison to cohort fertility based on birth registration. It has to be noted, however, that within the scope of the overall trend in cohort fertility, there are no major discrepancies between the FS-based and registration-based rates. Both data sources provide quite similar patterns of cohort fertility.

Application of inverse probability weights to the FS data adjusts the sample distributions to the distributions of the National Census with respect to size of place of residence, age, education and voivodship. As a result, the total cohort fertility estimated on the FS data for females born 1945-1985 shifts down, however, there still are noticeable differences with respect to cohort fertility calculated from registration data. The remaining difference cannot be attributed to the biased sample structure. We may assume that it might be related to random errors or other sources of errors, such as a coverage bias which results in a lower representation of childless females or couples who might be difficult to reach in a standard survey.

The detailed analysis of parity-specific cohort fertility rates revealed that the difference between the weighted and non-weighted FS data is parity-sensitive. The non-weighted FS sample generates a higher proportion of females with more than three children, resulting in a higher cohort fertility. Application of weights removes the effect of the skewed sample (place of residence) and shifts the cohort fertility curves down. The weights accurately adjust high-parity births (3^{rd} and higher) and, to some extent, second births to the levels derived from the registration data. It has to be noted that weights do not affect the cohort rates for first births. Irrespective of using weighted or non-weighted data, cohort fertility rates for first births are significantly higher in comparison to values based on the registration data. Since the application of weights removes differences between distributions of the analyzed variables in the FS and the National Census, we observe a downward shift in the cohort fertility patterns for third, fourth and higher-order births. At the same time, only a moderate effect with respect to second birth rates is noticed and virtually no effect on first birth rates. This might enhance our conclusion that there are other sources of bias observed for the total cohort fertility rates, such as random sampling error or selection bias consisting in underrepresentation of childless females or couples and a slight overrepresentation of couples with one or two children.

On the basis of the presented analyses we may conclude that the non-weighted FS data overestimate cohort fertility for Polish females born 1945-1985 but the magnitude of the bias is far lower than the differences observed in other databases such as the German GGS survey (Kreyenfeld et al., 2010).

Application of inverse probability weights to the original FS dataset adjusts distributions by place of residence, age and education to the level obtained from the National Census. As a result, the weights significantly improve cohort fertility rates based on the FS data in comparison to fertility rates based on registration data; however, they do not entirely remove the observed difference between values derived from these data sets. The remaining difference results from random errors and most likely also from an underrepresentation of childless females or couples as well as an overrepresentation of females who have one or two children.

We also conclude that the use of non-weighted FS data for demographic modeling of processes on an individual level should not affect the final results or lead to erroneous conclusions, especially when modeling a transition to first birth. The extent to which reproductive patterns observed in the FS data deviate from cohort fertility based on registration is not as significant as in case of the German GGS data (Kreyenfeld et. al. 2010). However, when modeling transitions to first and higher-order births we have to be aware that the original FS sample deviates from the respective distribution in the National Census. Therefore it is advisable to consider using inverse probability weights in models on the micro level and to compare the results with the non-weighted ones.

REFERENCES

Beckett M., J. Da Vanzo, N., Sastry, C. Panis, Ch. Peterson, 2001, *The Quality of Retrospective Data: An Examination of Long-Term Recall in a Developing Country,* „The Journal of Human Resources", 36(3): 593-625.

Hayford S.R., P. Morgan, 2008, *The Quality of Retrospective Data on Cohabitation,* „Demography" 45(1): 129-141 .

Holzer J.Z., D. Holzer-Żelażewska, 1997, *Płodność kohortowa kobiet w Polsce w latach 1945-1994, (Cohort Fertility of Polish Women, 1945-1994),* „Studia Demograficzne" 2: 3-23.

Holzer-Żelażewska D., K. Tymicki, 2009, *Cohort and Period Fertility of Polish Women, 1945-2008,* "Studia Demograficzne", no. 1: 48-69

Kreyenfeld, Michaela, Anne Hornung, Karolin Kubisch, and Ina Jaschinski, 2010, *Fertility and Union Histories From German GGS Data: Some Critical Reflections.* MPIDR Working Paper 2010-023. Max Planck Institute for Demographic Research, Rostock; http://www.demogr.mpg.de.

Murphy M., 2009, *Where Have All the Children Gone? Women's Reports of More Childlessness at Older Ages Than When They Were Younger in a Large-Scale Continuous Household Survey in Britain,* "Population Studies" 63(2): 115-133 .

Paradysz J., 1989, *O błędach nielosowych w badaniu dzietności kobiet w ramach Narodowego Spisu Powszechnego 1970,* [w:] *Problemy badań statystycznych metodą reprezentacyjną*, Główny Urząd Statystyczny, Warszawa, Biblioteka Wiadomości Statystycznych t. 36: 154-159.

Paradysz J., 1992, *Dzietność kobiet w Polsce*, Główny Urząd Statystyczny, Warszawa.

Paradysz J., 2000, *Rekonstrukcja dzietności małżeńskiej kobiet w późniejszym wieku na podstawie ankiety retrospektywnej*, Studia Demograficzne, nr 1, 1999: 13 - 34.

Paradysz J., 2002, *Badanie małżeńskości i dzietności kobiet w narodowych spisach powszechnych*, Wiadomości Statystyczne nr 1, 2002: 77- 87.

Rendall M.S., L. Clarke, E.H. Peters, N. Ranjit, G. Verropoulou, 1999, *Incomplete Reporting of Men's Fertility in the United States and Britain: A Research Note,* „Demography" 36(1): 135-144.

Weisberg H. F., 2005, *The Total Survey Error Approach,* Chicago: University of Chicago Press.

ABSTRACT

The paper validates data from the Fertility Survey 2002 in order to establish their comparability with selected variables from the National Census and cohort fertility trends calculated with use of the birth register database. In the first part of the analysis, we compare distributions of residence, age and educational level from the Fertility Survey and the National Census in order to measure the sampling bias. The second part of the analysis compares cohort fertility rates based on the Fertility Survey and registration of births. In this part, both weighted and non-weighted Fertility Survey data were used in order to account for biases from the cohort fertility based on registration data. We conclude that the distribution of selected variables obtained from the Fertility Survey sample significantly deviates from the respective distribution from the National Census. This bias influences cohort fertility rates calculated using Fertility Survey data, which tend to overestimate cohort fertility rates observed in the population. Application of inverse probability weights removes the bias in the sample structure but does not entirely remove overestimation of cohort fertility rates. We also conclude that the remaining difference in cohort fertility results from other sources of bias which we cannot control for. However, the magnitude of bias should not have an impact on the results of statistical modeling with use of Fertility Survey data.

Key words: Fertility Survey 2002, data validation, cohort fertility, data quality, sampling bias