The fertility transition in South Africa: A retrospective panel data analysis

Data appendix

The NIDS data was collected by the South African Labour and Development Research Unit (SALDRU) during 2008. The survey asked females aged 15 and older to report the number of live births, their date of birth and – where applicable – the dates of child mortalities. As discussed in section 4, this type of retrospective birth data is expected to suffer from recall bias, and this bias is likely to increase with the recall period.

A.1 Missing birth year data

Our analysis of the data shows that, apart from failing to report all births, respondents also neglected to report birth years for 1,221 out of the total 19,683 recorded births. In order to explore the nature and extent of the resulting measurement error, we construct a variable that expresses the number of births with missing birth years as a share of total reported births for each woman, and regress this on a number of explanatory variables. The coefficients and standard errors (in brackets) of this regression are presented in Table A1 below. The results demonstrate that missing birth years are more likely to occur the older the woman is at the time the survey is taken, the more children she gave birth to, the lower her level of schooling and if she was African or Coloured rather than Indian or White (the reference group in this regression).

Table A1: OLS regression for unreported birth years as share of total live births

Birth year	-0.003*** *(0.0003)*
Number of reported live births	0.016*** *(0.0022)*
African	0.031*** *(0.0111)*
Coloured	0.039*** *(0.0127)*
Indian	0.014*** *(0.0141)*
Years of completed schooling	-0.009*** *(0.0003)*
Constant	0.266*** *(0.0240)*
R squared	0.2186
Observations	7126

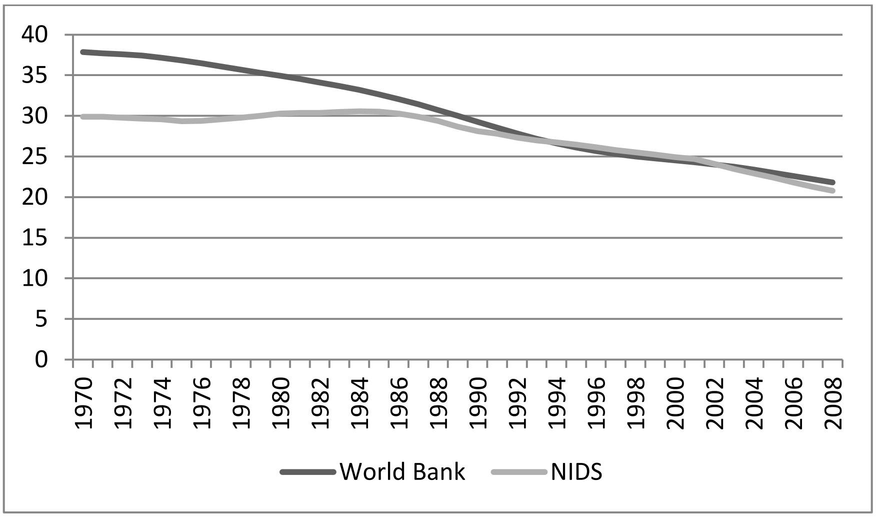
When constructing an annual panel data set, these undated births must be set to missing, which will exacerbate the under-capturing of fertility. In an attempt to reduce the effect of this bias we exclude women who fail to report the birth year of at least one of her children from the panel. The motivation for this restriction is that fertility is known to be under-captured for these women, but may not be for the rest of the sample. However, the fact that these women are disproportionately drawn from those with a large number of children means that this is unlikely to completely remove the downward bias. Since this is also shown to be more of a problem for women born longer ago – the same women who are likely to suffer from recall bias – the magnitude of fertility under-estimation can perhaps be addressed by restricting our sample period to the more recent past.

A.2 External validity checks

In order to investigate the magnitude of this bias, we compare a crude birth rate measure derived from the NIDS retrospective panel[1] with the estimates of the crude birth rate from the World Bank (2012). Figure A1 reveals that the NIDS data produce a crude birth rate that is very similar to the World Bank estimates for the 1985 to 2005 period, but that the under-capturing of births becomes a serious problem as soon as we look further back than 1985. This pattern is consistent with the bias that we would expect to arise from the above-mentioned sources, and suggests that the post-1985 sample is relatively reliable.

[1] The crude birth rate is calculated in the following way: for each calendar year we first calculate the proportion of women between the ages of 15 and 49 who reported giving birth in that year. This is done via a kernel weighted local polynomial smoother. We then multiply this number by 1000 and divide it by the share of the total population that consisted of females aged 15 to 49. This share is calculated from the gender and age group-specific population numbers from the ASSA 2008 lite model (for 1985-2008), from Udjo (1998) for 1970 and from linear interpolation for the years between 1970 and 1985.

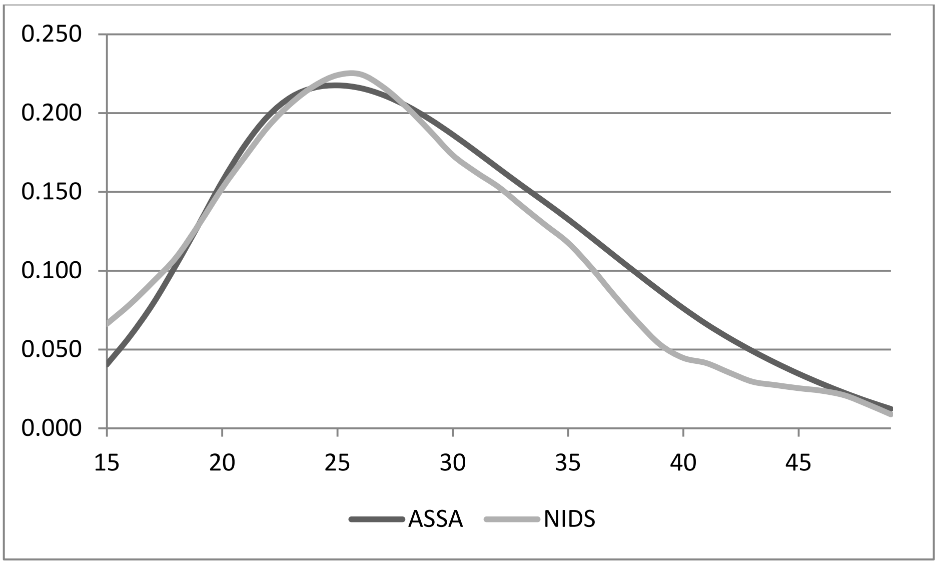
Figure A1: Comparison of crude birth rates (1970-2005): World Bank and NIDS data estimates



Source: World Bank (2012); NIDS, own calculations

Although Figure A2 demonstrates that the 1985 total birth rate was accurately captured by the retrospective element of the NIDS data, we may still be concerned that this overall rate obscures biases in the cross-sectional distribution of fertility. As an additional check of the external validity of the sample, we therefore compare the 1985 age-specific fertility rates as estimated from the NIDS data (again using a local polynomial smoother) and the Actuarial Society South Africa (ASSA) 2008 lite model (ASSA 2008). The NIDS data produce slightly lower estimates for women aged 30 to 45, but are generally very similar to the ASSA estimates.

Figure A2: Comparison of age-specific fertility rates (1985): ASSA and NIDS data estimates



Source: Actuarial Society South Africa; NIDS, own calculations

The analysis shows that the NIDS data are broadly consistent with the fertility trends documented elsewhere. It also demonstrates that the reliability of this retrospective panel decreases as the recall period lengthens and we find that the recall of events occurring before 1985 may be problematic. For this reason our empirical analysis will be restricted to the sample of women born after 1960: those who were in their highest birth probability years between 1985 and 2008.

A.3 Variables

The NIDS data also asked retrospective questions about schooling progress, the duration of relationships and migration that allow us to construct time-varying measures of years of education, marital status and province of residence. However, for the last two variables the questions asked are not informative enough to perfectly reconstruct the time variation in the variables of interest, so that we have to settle for proxy variables that provide noisy measures of the determinants that we want to control for.

With regards to schooling, the NIDS questionnaire asked respondents about the highest level of schooling completed, the first and last years in school and how many times each grade was repeated. This allows the construction of an accurate panel data measure for years of schooling completed at different points in time.

NIDS also asked respondents about their current relationship status – whether they were married, living with a partner, divorced, widowed or never married – as well as the duration of this relationship. The data therefore allow us to assign a relationship status to individuals for the duration of their current relationships, but provides no information about what happened before the start of this relationship. Values for these observations are inferred from the relationship patterns observed for other women, but will necessarily be a noisy measure of the actual relationship status. However, the results obtained in section 5 were found not to be sensitive to the omission of year-individual combinations for which these values had to be imputed.

A similar issue is encountered with the retrospective questions regarding area of residence. Individuals were asked where they were born, where they lived in 1994 and 2006, where they lived before moving to their current location, and when this most recent move occurred. This information can be combined with their current location to construct a relatively informative province of residence variable, although any migration between these dates (excepting the most recent one) will not be captured. Where individuals are known to have migrated between provinces and no date for this move is supplied (which is always the case unless this was the final move), individuals are assumed to have moved only once, and on the date that lies halfway between the two dates for which place of residence was reported.

The NIDS data do not contain any retrospective information on employment, wages or income, so we use estimates of real per capita GDP for different race groups in different periods, taken from Van der Berg *et al* (2006). The real GDP growth rate was taken from the South African Reserve Bank Quarterly Bulletins. In some specifications we also included a measure of the real discounted value of the child support grant, measured at period-specific grant values and eligibility ages. However, this variable was found not to significantly affect fertility outcomes.

We construct a variable capturing the number of children that women have had at a specific point in time based on the detailed retrospective birth records. To measure the impact of fertility as ‘replacement’ behaviour we also include an indicator of the number of children that have died (again measured at different points in time).We explore gender preferences by including dummy variables for whether a woman has had any boys or any girls. Again, this variable is created for the same individual at various time periods so that we can use it in our panel.