Sibling Death Clustering in India: Genuine Scarring *vs* Unobserved Heterogeneity*

Discussion Paper No. 03/552

Wiji Arulampalam[+] and Sonia Bhalotra[++]

June 2003

Keywords: Death clustering, infant mortality, state dependence, scarring, unobserved heterogeneity, dynamic random effects logit (probit), India.

JEL classification: J1, C1, I1, O1

*Acknowledgements: We are grateful to Marcel Fafchamps, Norman Ireland, Ian Preston, Arthur van Soest, Helene Turon and Mike Veall for helpful comments, and to Alfonso Miranda for help with the Stata programming We have benefited from presentations at the Royal Economic Society Conference (Warwick, 2003) and at the Universities of Bristol, Oxford, McMaster (Canada), Essex, Tilburg and Southampton. We are also grateful to ORC Macro International for providing us with the data. This organisation bears no responsibility for the analysis or interpretations presented in this paper.

[+] Department of Economics, University of Warwick, Coventry, CV4 7AL, U.K. Tel: +44 (0)24 76 523471 Fax: +44 (0)24 76 523032; Email: wiji.arulampalam@warwick.ac.uk. Homepage: http://www2.warwick.ac.uk/fac/soc/economics/staff /faculty/arulampalam. Also IZA, Bonn, Germany.

[++] Department of Economics, University of Bristol, 8 Woodland Road, Bristol BS8 1TN, U.K. Tel.: +44 (0)117 9288418; Email: s.bhalotra@bristol.ac.uk. Homepage: http://www.ecn.bris.ac.uk/www/ecsrb/bhalotra.htm Also Research Associate at CSAE (Oxford), QEH (Oxford), CHILD (Turino).

ABSTRACT

Data from a range of different environments indicate that the incidence of death is not randomly distributed across families but, rather, that there is a clustering of death amongst siblings. A natural explanation of this would be that there are (observed or unobserved) differences across families, for example in genetic frailty, education or living standards. Another hypothesis of considerable interest for both theory and policy is that there is a *causal* process whereby the death of a child influences the risk of death of the succeeding child in the family. Drawing language from the literature on the economics of unemployment, the causal effect is referred to here as scarring. This paper investigates the extent of scarring in India, distinguishing this from family-level risk factors common to siblings. It offers a number of methodological innovations upon previous research in the area. Estimates are obtained for each of three Indian states, which exhibit dramatic differences in socio-economic and demographic variables. The results suggest significant scarring in each of the three regions. Eliminating scarring, it is estimated, would reduce the infant mortality rate by 7% in the state of Uttar Pradesh, 3.1% in West Bengal and 2.9% in Kerala.

1. INTRODUCTION

It is well known that low-income countries have a high incidence of infant mortality. Recent research has revealed the less well-known and striking phenomenon of sibling death clustering. Data from a range of different environments indicate that the incidence of childhood death is not randomly distributed across families but, rather, that there is a positive association of sibling deaths.[1] A natural explanation of this would be that families in which child deaths are concentrated are poorer or share genetic or environmental risk factors that predispose all of their children to higher death risks. Another, more intriguing explanation is what, here, is called *scarring*: this is a process whereby the death of a child *causes* an increase in the risk of death of the subsequent child in the same family.

A causal process of particular interest operates by the death of a child shortening the time to the next birth. As it can take up to 24 months for the mother to recuperate physiologically from a birth[2], a short preceding birth interval for the index child elevates this child's mortality risk.[3] This process may operate in either of two ways. One possibility is that the death of an infant child results in the mother ceasing to breastfeed and, thereby, being able to conceive sooner than otherwise.[4] Henceforth, this is referred to as *the fecundity hypothesis*. An alternative possibility is that the death of a child leads parents to (intentionally) conceive sooner in a desire to "replace" their loss (e.g. Preston 1985). This is the *replacement*

---

1 See, for example, Zenger (1993), Guo (1993), Curtis et al (1993), Miller et al (1992), Das Gupta (1990), Bean *et al* (1988) and Hobcraft *et al* (1985).

2 A new pregnancy requires replenishment of vital nutrients like calcium and iron that are needed to support foetal development (e.g. daVanzo and Pebley 1993, Scrimshaw 1996). This problem is likely to be more acute in developing countries where bioavailability of these nutrients from staples like cereal is low and nutrient losses associated with infections challenge the capacity of women to produce healthy children.

3 The mortality-raising effect of short birth intervals is illustrated in, for example, Hobcraft *et al* (1983), Cleland and Sathar (1984), Koenig *et al* (1990), Gribble (1993) and Nath *et al* (1994).

4 See Bongaarts and Potter (1983), Cantrelle *et al* (1978), Chen *et al* (1974).

hypothesis. A further possibility, hitherto unrecognised in this literature, is that the event of a child death leaves the mother depressed, as a result of which her subsequent child's health is compromised, both in the womb and in early infancy (see, for example, Steer *et al. 19*92). This is referred to here as the *depression hypothesis.*[5] The depression argument is empirically distinguishable from the other two hypotheses because it does not work *via* the birth interval. Although it is of policy significance to establish which mechanism or mechanisms underlie scarring and there is little definitive research in this area, this paper does not attempt to offer any conclusive results in this direction. It is, instead, concerned primarily with the prior task of correctly identifying the extent of scarring.

In the last decade and a half, demographers have shown an active interest in death clustering. This research has focused on two aspects of the problem. First, it has demonstrated the significance of family-level unobservables in equations for childhood mortality. Second, it has shown that the correlation of risks amongst siblings needs to be taken into account in generating standard errors for the estimated parameters. This study contributes, first, by recasting the issue as one of distinguishing a causal effect flowing from the actual event of a sibling death (*scarring*) from a positive correlation of sibling death risks that arises because siblings are exposed to many common influences, some of which are unobservable (*unobserved heterogeneity*).[6] Second, it discusses and attempts to resolve a number of

---

[5] It is plausible that there are learning effects, which result in the mortality risk of the index child *falling* on account of the death of the preceding sibling. For instance, if the older sibling died of diarrhoea, the mother may rush to learn how to prevent diarrhoea-related infant death. Any positive degree of scarring that is identified is then net of learning effects.

[6] A few of these studies do include the previous child's survival status in a model along with controls for unobserved heterogeneity. However, the distinction between causality and common risks is not made. Instead, they focus on the role of unobserved heterogeneity, treating the previous child's survival status as just another exogenous regressor (e.g. Curtis *et al* 1993, Guo 1993 and Sastry 1997a & 1997b, Bolstad and Manda 2001).

specification issues that arise in models that incorporate both "dynamic effects" like scarring and unobserved heterogeneity. It argues that the statistical procedures used in previous research are inappropriate. The methodological issues raised can be summarised as follows.

It is common practise in the existing literature to discard information on children born before a certain date and, further, to discard the first child of each mother in the sample. In a causal model with unobserved family-level heterogeneity, this will tend to bias the estimator of the scarring parameter (see Section 4.2). This paper avoids this problem by using the complete birth history of each mother and specifying a separate reduced form model for first-born children. A test for the relevance of this problem is provided and, in order to assess the extent of the bias associated with procedures used in previous research, the estimates are compared with estimates obtained using the sample selections and specifications employed in earlier work. Another feature of most previous work is that it tends not to provide a clear estimate of the scarring effect because it conditions on the birth interval, the alteration of which, under certain hypotheses (discussed above), is what drives scarring. This paper reports results with and without the preceding birth interval included as a regressor. We argue that the model that suppresses the birth interval provides a more appropriate estimate of total scarring. Comparison with the model that includes the birth interval offers insight into the nature of the mechanism underlying scarring (see Section 4.1). Other specification issues raised in this paper relate to distributional assumptions, time-inconsistency and sensitivity to recall bias or measurement error in reporting of the age of death (see Section 4). Results are presented to show the percentage of observed persistence that can be explained by genuine scarring (i.e. by

---

Zenger (1993) estimates alternative models, including *either* the previous child's survival status *or* unobserved heterogeneity but not both.

the survival status of the preceding sibling), and the reduction in mortality that would be achieved if scarring were eliminated.

The analysis is conducted for three Indian states, in each of which we find evidence of significant scarring. This immediately raises the payoff to policy interventions that reduce mortality because it implies that preventing the death of a child also contributes to preventing the death of siblings of that child.

The next Section describes the data used, the overall incidence of infant death and the extent of sibling death clustering. The econometric model is set out in Section 3, where scarring is formally defined and distinguished from unobserved heterogeneity. Issues that arise in estimation of the model given the nature of the available data are discussed in Section 4, which also delineates the relation of this paper to previous research. Section 5 describes the empirical model and defines the variables. The results are set out in Section 6. The sensitivity of the estimated scarring effect to alternative specifications and procedures used in the existing literature is investigated in Section 7. Section 8 concludes with a discussion of the findings and limitations of this study.

2. THE DATA & DEATH CLUSTERING IN INDIA

This paper uses the second round of the National Family Health Survey of India (NFHS-II), which interviewed 90000 ever-married women aged 15-49 in 1998-99 and recorded complete fertility histories for the 73775 mothers amongst them, including the time and incidence of child deaths. There are 248785 children in the sample, the mean number per mother being 3.4 and the median number 3. NFHS-II was conducted in 26 Indian states and covered more than

99 percent of India's population. For details on sampling strategy and context, see IIPS and ORC Macro (2000). The data are in the public domain and can be downloaded from www.macrodhs.com. The analysis is performed for the three states of Uttar Pradesh (UP), West Bengal (WB) and Kerala, which exhibit remarkable differences in social, demographic, economic and political development (see Dreze and Sen 1997). UP is the largest Indian state with social and demographic indicators that put it below the Indian average. Kerala is an exceptional state that leads India in almost every index of human development. West Bengal lies between the two in social-demographic development while exhibiting better economic indicators (level of per capita income, poverty incidence) than the other two states. A profile of the three states is presented in Table 1. Of every 1000 births in India, 82 die before the age of 12 months. There is remarkable inter-state variation. The corresponding numbers are 116 in UP, 76 in WB and 36 in Kerala (see Table 1).[7]

The top panel of Table 2 shows the raw data probabilities of infant death conditional on the survival status of the preceding sibling.[8] This is a useful description since, in the formal analysis conducted in this study and also in some previous studies, a first-order Markov model is specified in which, conditional on the survival status of the preceding child, the survival status of earlier children does not influence the survival status of the index child (see Section 3). Consider, for illustration, the state of UP. The probability of infant death is higher by 0.15 (i.e. it is 0.24 rather than 0.09) if the preceding sibling died as an infant. An alternative

---

[7] These figures are averages over the data sample. As this contains complete retrospective fertility histories, it includes children born across almost four decades, 1961-1999. The average number of infant deaths per 1000 live births in India is estimated to have been 67 in the year 2001 (UNDP 2003).

[8] For each child within a family (except the first born), probabilities of death are calculated conditional first on the death of the previous child and, second, on the survival of the previous child. The *difference* between these two probabilities is averaged over the sample to obtain an estimate of the scarring. An alternative way of

expression of the relative risk is that an infant in UP is 2.6 times as likely to die if the preceding sibling died rather than survived. Overall, the Indian data exhibit a remarkable degree of death clustering. Without further analysis, however, it is impossible to say whether this reflects genuine scarring or whether it merely reflects risks common to siblings on account of shared family characteristics.

3. THE ECONOMETRIC MODEL

This Section sets out an econometric model that permits identification of scarring, taking care of the potentially confounding effects of unobserved inter-family heterogeneity.

Let there be n_i children in family i. For child j ($j=2,\ldots,n_i$) in family i ($i=1,2,\ldots, N$), the unobservable propensity to experience an infant death, y_{ij}^*, is specified as

$$y_{ij}^* = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \gamma y_{ij-1} + \alpha_i + u_{ij} \qquad (1)$$

where $\boldsymbol{x}$ is a vector of strictly exogenous observable child and family specific characteristics that influence y_{ij}^* and $\boldsymbol{\beta}$ is the vector of coefficients associated with $\boldsymbol{x}$. A child is observed to die when his or her propensity for death crosses a threshold; in this case, when $y_{ij}^* > 0$. It is assumed that this unobservable propensity is a function of the *observed* survival status of the previous child in the family, denoted y_{ij-1}, so that it is the actual *experience* of death of the previous child rather than his or her *propensity* to die that affects the survival status of the index child. The null of no scarring implies $\gamma=0$. This is consistent with the scarring hypotheses considered in Section 1.[9] The term α_i accounts for all time-invariant unobserved and, possibly, unobservable family characteristics which influence the index child's

---

representing the data is to look at the *ratio* of the two average conditional probabilities. Both the difference and the ratio statistics are presented.

[9] The estimated parameter γ should be interpreted as the 'average' effect of scarring over the time period considered. In work in progress we investigate whether scarring has declined over time.

propensity to die. This will include genetic characteristics and variables such as innate maternal ability. Note that, in this model, conditional on y_{ij-1}, $\boldsymbol{x_{ij}}$ and α_I, the survival status of older children other than the immediately preceding child is assumed to have no impact on y_{ij}^*.

Since (1) is a recursive model, some assumptions regarding the survival status of the first child in the family, y_{i1}, are required. A reduced form equation for the first child is specified as

$$y_{i1}^* = \boldsymbol{z_{i1}}'\boldsymbol{\lambda} + \eta_{i1} \qquad i=1,....,N \text{ and } j=1 \qquad (2)$$

where $\boldsymbol{z_{i1}}$ is a vector of exogenous covariates, $var(\eta_i)=\sigma^2_\eta$ and $corr(\alpha_i, \eta_i)=\rho$.[10] In principle, the vector of covariates in $\boldsymbol{x}$ and $\boldsymbol{z}$ need not be the same. The possibility of non-zero ρ is allowed for by the linear specification,

$$\eta_{i1} = \theta\, \alpha_i + u_{i1} \qquad (3)$$

where, by construction, α_i and u_{i1} are orthogonal to one another, $\theta = \rho\, \sigma_\eta/\sigma_\alpha$ and $var(u_{i1})= \sigma^2_\eta(1-\rho^2)$. Hence, it follows that,

$$y_{i1}^* = \boldsymbol{z_{i1}}'\boldsymbol{\lambda} + \theta\, \alpha_i + u_{i1} \qquad i=1,....,N \text{ and } j=1 \qquad (4)$$

$$y_{ij}^* = \boldsymbol{x_{ij}}'\boldsymbol{\beta} + \gamma y_{ij-1} + \alpha_i + u_{ij}, \qquad i=1,....,N \text{ and } j=2,..,n_i \qquad (1)$$

Equations (4) and (1) specify a complete model for the infant survival process, equation (4) describing the start of this process. We assume that the family-specific unobservables, α_i, are independent and identically distributed with density h and that the conditional distributions of both y_{ij}^* and y_{i1}^* are independently distributed with a distribution function F. Further discussion of h and F is in Section 4.5.

---

[10] In principle, there is no reason why the specification for the first-born should be the same as the rest of the model especially when the latter is a conditional model in the sense that it conditions on the survival status of

Family or mother-specific unobservables are thus captured by a random effect (α_i). The alternative of treating these as fixed effects (i.e. parameters) and estimating them along with the other parameters of the model gives rise to the incidental parameters problem (Neyman and Scott 1948). Although the maximum likelihood estimator is inconsistent in the fixed-effects model, in the model with $\beta=0$ and u_{ij} distributed as a logistic, it is possible to obtain consistent estimator of the scarring parameter, γ, by maximising the conditional likelihood function (CML) where the conditioning is carried out with respect to a set of sufficient statistics- this eliminates the family specific unobservables from the likelihood function (Chamberlain 1985). Although the CML estimator has the advantage of not requiring an assumption about the distribution of α_i, against it is the fact that it involves a considerable loss of information. For example, only the subset of families who have experienced at least one death contribute to the CMLE (Table 1 shows that this would involve losing 69.3%, 83.9% and 92.4% of observations in UP, WB and Kerala respectively).

Marginalising the likelihood with respect to α_i gives the likelihood function for family *i*

$$L_i = \int_{-\infty}^{\infty} \Big(\prod_{j=2}^{n_i} F[(\boldsymbol{x_{ij}}'\boldsymbol{\beta} + \gamma\, y_{ij\text{-}1} + \sigma_\alpha\, \tilde{\alpha})(2y_{ij} - 1)\Big]$$

$$F[(\mathbf{z}_i'\boldsymbol{\lambda} + \theta\ \sigma_\alpha\ \tilde{\alpha})\ (2y_{i1} - 1)]\ \Big)\ h(\tilde{\alpha})\, d\tilde{\alpha} \qquad (5)$$

where, $\tilde{\alpha} = \alpha/\sigma_\alpha$. We obtain parameter estimates by maximising this likelihood using Stata's maximum likelihood routines (Stata 7 2000).[11]

---

the previous sibling. This is an assumption we make in this paper. Heckman (1981c) shows that this approximation works quite well in applications.

[11] For an account of dynamic (causal) models with unobserved heterogeneity in the econometrics literature, see Wooldridge (2002). The distinction made in this paper between scarring and unobserved heterogeneity has been made in other contexts in both statistics and economics (see Heckman 1978, 1981a, 1981b, 1981c)

4. ISSUES OF MODEL SPECIFICATION AND TESTING

This Section describes potential problems that arise in an empirical specification of the model, indicating the nature of the parameter estimator biases in some previous studies and how this paper attempts to avoid such biases. In Section 4.1, it is argued that conditioning on the preceding birth interval will tend to lead to under-estimation of scarring, in addition to which are problems of endogeneity and measurement error associated with birth intervals.[12] Section 4.2 argues that the practice of left-truncation of the data common to most previous studies results in potential over-estimation of the extent of scarring. This can be avoided by using the complete birth history and estimating a reduced form model for first-born children jointly with the "dynamic" model for second and younger children. The longer time range of the data employed can, in principle, exacerbate problems of measurement error and time-inconsistent variables, which are therefore discussed in Sections 4.3 and 4.4 respectively. Measurement error in age of death may create an upward bias in the scarring coefficient. The use of time-varying covariates measured at the time of the survey is inappropriate when the infant deaths that are being analysed may have occurred decades before the survey. Discussion of distributional assumptions for y_{ij}^* and for unobserved heterogeneity, α_i, is in Section 4.5.

---

although its relevance to death clustering has not formerly been recognised. For example, in the literature on the economics of unemployment, scarring refers to the effect of a past episode of unemployment on the future probability of experiencing unemployment, after controlling for all observable (e.g. education) and unobservable (e.g. ability) individual characteristics. Scarring is alternatively referred to as state dependence. It is useful to clarify the language for the current context. The idea is that the event of death of a child *scars* or marks the survival prospects of the succeeding sibling. Alternatively, defining a state as a realisation of a stochastic process, one may think of *state dependence* in terms of the mortality risk facing a child being dependent upon the state (died in infancy or not) revealed for the previous child in the family. Since time is implicit in the sequencing of children, models that include the previous child's survival status are analogous to dynamic models.

[12] *Endogeneity* refers to the birth interval being a behavioural or choice variable. In particular, families that exhibit longer birth intervals may systematically be families that are better able to avert child deaths. The statistical implication is that the birth interval is potentially correlated with the error term in the model

4.1. Specification of Scarring Effects

As discussed in Section 1, previous studies are not specifically looking to identify scarring effects as distinct from unobserved heterogeneity across families. This is reflected in the specifications that they employ. Most demographic analyses of death clustering model unobserved heterogeneity alone, although a few include variables related to the scarring process. The associated statistical and interpretational problems were summarised in Section 1 and are discussed in more detail in Section 4.2. This section discusses the specification of variables. In some studies, the number of surviving older siblings is used instead of the survival status of the previous child (e.g. Bhargava 2003, Muhuri and Preston 1991). This is a compound indicator of fertility and mortality in the family. Moreover, it is insensitive to sequencing. For both reasons, it may not reflect true 'scarring' as defined in Section 3. Other studies include the survival status of the previous sibling (Curtis *et al* 1993, Guo 1993, Sastry 1997a & 1997b, Bolstad and Manda 2001). However, all of these studies also include the preceding birth interval. To the extent that the previous child's survival status, y_{ij-1}, impacts on the index child's death risk, y_{ij}^*, by altering the length of the birth interval, conditioning on the birth interval will tend to weaken the coefficient on y_{ij-1}. As a result, the degree of scarring will tend to be under-estimated. Another problem with this specification is that the birth interval is an endogenous variable and one for which valid instruments may be difficult to find.[13] There are also measurement problems with birth intervals as they may be shorter on account of

---

describing mortality risk for the index child. If this is not dealt with, the estimator of the coefficient on the birth interval will be biased.

[13] Endogeneity is defined in the previous footnote. Although *uptake* of contraception is a choice variable (endogenous), the *availability* of contraception is a potential instrument for birth interval. This does not appear to have been considered in the previous literature. Since information on contraception in the NFHS data is limited to recent births and using it would involve endogenous left truncation of the data (see Section 4.2), this exploration is left to future work.

premature birth (e.g., Gribble 1993) or longer on account of miscarriage (e.g. Madise and Diamond 1995). If these events are sufficiently common in the data, the coefficient on birth interval will reflect a compound of these effects.

In this paper, the scarring effect is captured entirely by the coefficient on previous sibling's survival status. To allow comparison with previous studies and, for the Indian data, to assess the impact on γ, results are presented, in Section 7, for a variant of the model in which preceding birth interval is included as an additional regressor. If the estimated scarring effect were diminished, the data would seem to be consistent with the fecundity or replacement hypotheses. If a positive degree of scarring persists, then there is room for the depression hypothesis or, indeed, other causal mechanisms that operate independently of the birth interval (see Section 1). Of course, in the absence of controls for the endogeneity of birth spacing, these results are only indicative.[14]

4.2 The Initial Conditions Problem in a Dynamic Model

A pervasive practise in previous research is to discard information on children born before an often arbitrarily selected date, such as five, ten or fifteen years before the date of the survey (e.g., Guo 1993, Curtis, *et al* 1993, Madise and Diamond 1995, Sastry 1997a & 1997b, Bolstad and Manda 2001, Bhargava 2003).[15] Studies that include the previous sibling's survival status or the preceding birth interval as a regressor also discard the first child in the truncated sample (since y_{ij-1} is undefined for these children). Left truncation of this sort,

[14] We are currently investigating a model that allows simultaneous determination of mortality risk and birth intervals in a paper in which the main objective is to identify the extent to which scarring can be explained by birth spacing. In this paper, the objective is to identify the extent of scarring as distinct from unobserved heterogeneity, rather than to identify the mechanism driving scarring.

[15] The fact this left truncation of the data by calendar time occurs at different points in the birth history of different households creates additional complications.

whether by calendar time or by birth-order of child, results in the problem that the start of the sample does not coincide with the start of the stochastic process under study. On account of the presence of family unobservable characteristics, α_i, in equation (1), the survival status of the previous child, y_{ij-1}, is endogenous and so discarding observations at the beginning of the sample results in an endogenously truncated sample.[16] This is the 'initial conditions problem' in dynamic models (see Heckman (1981c), for example).

In principle, consistent estimators can be obtained from an endogenously truncated sample if an equation for the first-child is specified and an appropriate identifying restriction can be found, that is, a variable that influences the first sample observation but does not appear in the equations for higher birth order children. In general, it may be difficult to find a valid identifying restriction although, in a non-linear model, identification may be achieved on the basis of covariates that change with j, the index child (Chamberlain 1984; Hyslop 1999). This study takes the alternative route of using all of the retrospective information available so that the first observation refers to the first-born child for each mother (which is the initial condition of the process).[17] As is clear from Section 3, a separate reduced form equation for the first-born child is specified and included in the model.[18] Identification of our model is

---

[16] Since α_i is time-invariant it will appear in the equation for every child in the family. In particular, it will appear in the equation for y_{ij}^* and also in the equation for y_{ij-1}^*. Therefore, in the equation for y_{ij}^*, the regressor, y_{ij-1}, is necessarily correlated with the error-component, α_i. This is what is meant by endogeneity of y_{ij-1} and, left unaddressed, it will tend to produce a (positive) bias on the coefficient of y_{ij-1}, which indicates scarring.

[17] There are applications in which data on the start of the process are unavailable. For example, in studying unemployment spells of individuals, researchers would ideally like to have data on school-leavers but must often make do with data that do not include the first spell of unemployment for each individual. On the other hand the series of Demographic and Health Surveys (DHS) available for developing countries do typically contain information on all children of a mother including the first-born. This paper argues the importance of using this information.

[18] If the first conception is a miscarriage then the first-born child may not adequately represent the initial condition of the process. This problem cannot be directly addressed since the DHS data (including the NFHS-II used in this paper) do not contain information on miscarriages. However, it is shown in Section 7 that, even

further aided by inclusion of a covariate that is different for each child in the family, namely, the age of mother at the birth of the index child.

Most previous studies neglect to recognise this problem.[19] Discarding initial observations creates an unnecessary[20] and often severe loss of information (see the number of observations recorded in rows 1-4 of Table 3). Moreover, it is an important issue in analyses of death clustering, as it will tend to bias the estimator of scarring. Given that the correlation between y_{ij-1} and α_i is positive, the bias will be positive. The direction and size of this bias is assessed in Section 7. Section 6 reports a test of the null hypothesis that $\theta=0$ in (3). This is a test of the hypothesis that the initial sample observation within a family (indexed *j=1*) can be treated as exogenous. Clearly, if $\theta=0$ then unobservables in the equation for the first observation are uncorrelated with unobservables in the [dynamic] equations for subsequent observations. In this case, the model described by (1) and (4) reduces to a simple random effects model; a separate specification of the equation for the initial sample observation is unnecessary. A further testable restriction that is investigated is $\theta=1$, which implies perfect correlation between the family specific unobservables in equations (1) and (4).

4.3 Measurement Error

---

when data on the first-*born* child are unavailable, consistent estimates of the scarring effect may be obtainable if a distinct equation for the first-*observed* child in the truncated sample is included in the estimation. For this reason, even if miscarriages do occur at the start of the process for some women, the resulting bias in γ is expected to be small.

[19] This is true of all of the relevant demographic research that we are aware of. The only study that reflects awareness of the endogeneity problem arising *via* the correlation of the survival status of previous children and family unobservables is Bhargava (2003). He uses samples of data restricted to 5 and 10 years before the date of the survey (NFHS-I for the state of UP alone 1991/92) and addresses the endogeneity problem by imposing the restriction that household possessions (indicators for whether bicycle, radio etc are owned at the time of the survey) and the number of boys and girls born before the index child influence the *number of surviving older children* but, conditional on this variable, have no influence on the mortality risk of the index child.

A reason sometimes offered in previous studies for left-truncation of the sample is that this is done to minimise recall error in the recorded date of child death, which is assumed to be larger the further away the mother is from the event (e.g. Sastry 1997a). It may seem implausible, *a priori*, that mothers ever forget the date of death of a child but the data do exhibit some age-heaping. In particular, the Indian data that are used in this study show heaping at six-month intervals (also see IIPS and ORC Macro 2000: Section 6.2). Since the model has infant death on both sides of the equation, with the index child's risk a function of the preceding child's survival status, positively correlated measurement error in these variables will tend to create an upward bias in the scarring coefficient.

The dependent variable and the survival status of the preceding child are both coded as binary variables that are unity if the child dies before the age of 12 months and zero otherwise. To investigate sensitivity of the estimates to age-heaping at 12 months the models were re-estimated with these variables defined to include deaths occurring at 12 months. The results were very similar (and so are not shown but available on request).

4.4. Time Inconsistency

Survey data used to study childhood mortality typically contain complete retrospective histories of births and child deaths experienced by ever-married women aged 15-49. The data we use for India are similar. A woman aged 49 in 1999 may have experienced a birth and an infant death as long ago as 1969. As a result, data on the current assets of her household or the facilities available in her village are unlikely to be informative in an analysis of childhood

[20] Sometimes left-truncation of the data is forced upon the researcher by the nature of the survey. Section 7 suggests how consistent estimates may be obtained in this case.

deaths. This is the time-inconsistency problem.[21] Several previous analyses use time-inconsistent information for variables such as household assets, toilet facility, electricity or access to piped water.[22] The results, which in principle are highly policy relevant, are unlikely to be robust. A further problem with some of these variables is that they are endogenous. For example, families will tend to simultaneously decide what resources to allocate to the purchase of a bicycle or a TV and what resources to spend on inputs into child health that will reduce child mortality risk (see Becker 1991, for example). Or access to facilities like piped water may be endogenous if families migrate to regions with piped water.

A few recent papers model community-level random effects (e.g. Bolstad and Manda 2000; Sastry 1997a). Bolstad and Manda show that omission of controls for community-level unobservables biases the standard errors of the estimates, while Sastry argues that neglecting to model community effects leads to over-estimation of the family effect.[23] However, community random effects run into the time inconsistency problem when the underlying

[21] There is plenty of evidence in the literature that both income mobility and geographical mobility in developing countries is considerable. The recent availability of household and individual-level longitudinal data for developing countries has made it possible to study income distribution dynamics. This research indicates considerable "churning" in the distribution with the identity of households classified as poor changing quite rapidly through time (see Baulch and Hoddinott 2000). There is also a non-negligible degree of geographical migration (see Williamson 1998). Community infrastructure tends to grow rapidly from a low base in the process of economic development. Social norms also often change rapidly with growth and migration. Together, these facts make implausible the assumption that current household assets or current community infrastructure are a good proxy for the socio-economic status of the household at the time that the children in question were exposed to the risk of infant death.

[22] Clearly the problem is mitigated by the fact that these studies left-truncate the data. Nevertheless, constancy of these variables for the time spans of 10-15 years covered in these studies remains questionable.

[23] In this paper, the emphasis is on obtaining an unbiased estimate of scarring, with the family component of the error term performing the job of mopping up any family-level variation in the data that is not explained by the exogenous covariates ($\boldsymbol{x}$ or $\boldsymbol{z}$). There is no problem with community-level variation becoming part of this term.

unobservables (ranging from community infrastructure to social norms) are subject to rapid change (see footnote 20).[24]

The left-truncation of the data referred to in Section 4.2 mitigates the time inconsistency problem by severing the retrospective information before it gets into the distant past. In the preferred specification in this paper, which involves using information on the entire history of births for every woman in the sample, time inconsistency is avoided by including in the model only those conditioning variables that are time-invariant or at least relatively sluggish (see Section 5). A covariate that varies with the index child is the age of the mother at birth of the index child. This, of course, does not pose any time-inconsistency problems and is useful for identification (see Section 4.2).

4.5. Distributional Assumptions

Conditional distribution of y_{ij}^* (F)

A popular assumption for *F*, the distribution of y_{ij}^* conditional on α_i, $\boldsymbol{x_{ij}}$ and y_{ij-1} is that it is logistic. In order to check for the sensitivity of the estimates to this assumption, the models were estimated with *F* specified, alternatively, as logistic, standard normal and extreme-value. Unlike the logistic and the standard normal, the extreme value distribution is not symmetric and, if tail behaviour is important in determining infant death probabilities, then results from the standard normal model might differ from the logistic. Since the results obtained were not very sensitive to the choice of F, we only report and discuss results where F is specified as a logit (other results available upon request).

Distribution of unobserved heterogeneity, α_i (h)

---

[24] Mother-specific (i.e. family-level) random effects included in this and previous studies are much more likely to be stable. We expect mother-specific unobservables in a model of child mortality to include genetic factors,

Following the literature, it is initially assumed that α_i, the component of the error term representing unobserved family-level heterogeneity, is independently and identically distributed as a normal variate. A weakness of this assumption is that it may not allow enough flexibility to model the fact that some families never experience any child deaths and that in some families all children die. This is the well-known mover-stayer problem in the statistical literature. Referring back to equations (1) and (4), a very large positive (negative) value for α_ι will give a very large (small) value for y_{ij}^* and hence a very large (small) probability of observing death of the index child. This can be accommodated by allowing for empirically determined masses at the two extremes, that is, at plus and minus infinity of the Normal mixing distribution.[25] This gives the following likelihood for family *i*,

$$L_i^* = \frac{\psi_0}{1+\psi_0+\psi_1}\left[\prod_{j=1}^{n_i}(1-y_{ij})\right] + \frac{\psi_1}{1+\psi_0+\psi_1}\left[\prod_{j=1}^{n_i} y_{ij}\right] + \frac{L_i}{1+\psi_0+\psi_1} \qquad (6)$$

where L_i is given by equation (5) and ψ_0 and ψ_1 are the unknown end-point parameters. The estimated proportion of families who will have a very large or a very small α_i is given by p_1 and p_0 respectively, where,

$$p_0 = \frac{\psi_0}{1+\psi_0+\psi_1} \quad \textbf{\textit{and}} \quad p_1 = \frac{\psi_1}{1+\psi_0+\psi_1}. \qquad (7)$$

In order to ensure the non-negativity of ψ, it was parameterised as $exp(\kappa)$ and κ was estimated. In practice, the data may not contain enough variation in order to allow us to identify ψ_1 and this is, indeed, what was found in this study (see Table 1 where the proportion of families that lose all or none of their children in infancy is reported).

---

attitudes or inherent maternal ability, all of which can plausibly be assumed stable over time.

Testing for the significance of inter-family heterogeneity

$$\text{Let } corr(\alpha_i+u_{ij}, \alpha_i+u_{ik}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2} = r \text{ say, for all } j \neq k \neq 1. \quad (8)$$

The correlation coefficient *r* gives the proportion of total error variance that is attributed to the unobservable family effect, α_i. For the estimates presented here, which assume a logistic distribution for u, $\sigma^2_u = \pi^2/3$. A test of H_0: $\sigma_\alpha^2=0$, which is a test that there are no unobservable family characteristics in the model and therefore that it collapses to a simple binary dependent variable model, is equivalent to a test of H_0: $r=0$ in equation (8). This can be tested as a likelihood ratio (LR) test but the test statistic will not be a standard χ^2 test since the parameter restriction is on the boundary of the parameter space. The standard LR test statistic has a probability mass of 0.5 at zero and $0.5\chi^2(1)$ for positive values. Thus a one-sided 5% significance level test requires the use of the 10% critical value (Lawless (1987)).

5. THE EMPIRCAL MODEL

The dependent variable, y_{ij}^*, is defined as unity if the child is observed to die before the age of 12 months and zero otherwise (infant death). The regressor of interest, y_{ij-1}, is similarly defined as the infant survival status of the preceding sibling. Children who have not had 12 months exposure (i.e. who are younger than 12 months) at the time of the survey are dropped from the sample. When the index child is not a singleton but, instead, a twin or triplet then care is taken to ensure that the preceding sibling is correctly identified and is the same for each twin. When the previous child is one of a multiple birth, then y_{ij-1} is defined as unity if all

[25] See Narendranathan and Elias (1982) for an application of this distributional assumption in the context of modelling individual unemployment.

children of that multiple birth died in infancy and as zero otherwise.[26] Sensitivity of the results to "heaping" in the reported age of death is investigated (see Section 4.3).

The rest of this Section describes the variables in the vector $\boldsymbol{x_{ij}}$, which are assumed to be identical to the variables $\boldsymbol{z_{ij}}$ in the first-child equation. Means and standard deviations of all variables in the model are in Appendix Table 1. Covariates often used in previous research that are time-inconsistent or endogenous are avoided.[27] The only time (i.e. sibling)-varying covariate used is age of the mother at birth of the index child. The only potentially endogenous variable in the model is y_{ij-1} and, as discussed in Sections 3 and 4.2, addressing this potential problem is an important part of the statistical approach taken here. Since this involves using retrospective histories that go back several years in time, cohort effects are introduced into the model although these are seldom included in previous demographic studies.

Child-specific regressors in the equation are child birth-order, gender and an indicator for whether the child is one of a multiple birth (twin, triplet, etc). Most previous studies allow a gender-specific intercept but virtually all discard data on multiple births. Another exogenous child-specific covariate that is included in the model and which assists in identifying the parameters of interest (see Section 4.2) is the age of the mother at birth of the index child. This is expected to reflect the physiological condition of the mother at a relevant time. Since several studies show child mortality risk to be U-shaped in mother's age, this is specified as a quadratic. Education of the mother is denoted by a set of dummy variables for level of

---

[26] This is the relevant assumption if the mechanism underlying scarring is the fecundity mechanism since the mother is only likely to stop breastfeeding if both twins or all three triplets die. We have confirmed that altering this definition so that y_{ij-1} is defined, as unity when *at least one* of the multiple births dies does not change the results. This is unsurprising since multiple births are uncommon (see Table 1).

[27] For examples of time-inconsistent variables used in previous research, see Section 4.4. For discussion of endogeneity, see footnotes 11 and 15. Examples of endogenous variables used as regressors in related studies are birth spacing, breast-feeding and contraceptive use.

education attained. This is relatively flexible, allowing for non-linear effects.[28] A similar set of indicators for educational level of the father is included. This is likely to be an important control for socio-economic status to the extent that fathers are the main earners (available data on household assets are not used because of the time inconsistency problem). Other family-level observable variables included in the model are religion and caste. These allow for "sociological" influences on child death risks.

Cohort effects are modelled by including dummy variables for year of birth of the mother. Mothers in the sample are born between 1948 and 1984. Roughly equal frequency groups are created by defining dummy variables for births during 1948-1959, 1960-1969 and 1970-1984. The cohort effects are expected to pick up any secular decline in death risks over time, other things equal. Previous research does not appear to have allowed cohort effects but this is possibly because, after left-truncation (see Section 4.2), the distribution of year of birth of mother is more peaked.

For observations with a large number of missing values, dummy variables are created to indicate missing values and these are included in the model estimation.

6. RESULTS

The main result is that we find evidence of scarring in each of the three Indian states after controlling for a number of exogenous child and family-specific characteristics and for all unobserved differences between families (see the bottom panel of Table 2).[29] Since coefficients are not directly interpretable in binary dependent variable models, the marginal

[28] For a review of the important effects of maternal education on childhood mortality, see Cleland and van Ginneken (1989) and Hobcraft (1993). Also see Rosenzweig and Schultz (1982).

[29] A full set of results for other covariates is available and can be requested from the authors. In this paper, attention is restricted to identification of the scarring parameter and associated methodological issues.

effect associated with γ is computed as the difference between the sample averages of the probability of death predicted by the estimated model when y_{ij-1}=0 and when y_{ij-1}=1.[30] The Table also shows the ratio of these conditional probabilities. Comparing these with the difference and ratio of the "raw data probabilities" discussed in Section 2 (and reported in the top panel of Table 2) affords an estimate of the percentage of raw persistence (or clustering) that is explained by scarring, using the model specified in Section 3. Scarring explains about 42% of the clustering observed in the data in UP; the corresponding proportions being 15% for WB and 28.9% for Kerala. As discussed, previous research has identified clustering with unobserved heterogeneity- these estimates show that in fact, almost half of observed clustering in UP is attributable to scarring, after holding constant unobserved heterogeneity. Comparing the averaged model predicted probability of death with that of the averaged predicted probability of death setting γ=0 offers an estimate of the reduction in mortality that would be achievable if scarring were eliminated- a useful alternative expression of its significance.[31] The estimates suggest that, in the absence of scarring, mortality rates would fall by 6.96%, 3.09% and 2.92% in UP, WB and Kerala respectively. The hypothesis that θ=0 is decisively rejected in the case of UP and WB. This confirms the importance of specifying a distinct reduced form equation for the first child that is estimated jointly with the dynamic equations for other children. Further, the estimate of θ is not significantly different from unity (in all three states), which implies that unobserved heterogeneity terms in the equations for the first

[30] This is approximately equivalent to the first partial derivative of the conditional probability of death of the index child (the conditional expectation of y_{ij}) with respect to the covariate.

[31] Since the model includes an end-point at minus infinity (p_0), which allows for a proportion of families to never experience deaths, we have adjusted the predicted probabilities by the estimated p_0.

child and for subsequent children are perfectly correlated.[32] Also, the restriction $\beta=\lambda$ cannot be rejected.

The proportion of the variance attributable to family-level unobservables (α_i) is estimated to be 12% in UP, 22% in WB and 7.2% in Kerala. For each state, the estimates decisively reject the null of no family-level unobservables (using the test statistic described in Section 4.5). This and the finding that exclusion of α_i from the model results in over-estimation of scarring (results not shown but available) underline the importance of controlling for α_i.

Many of the covariates in the vector $\mathbf{x}_{ij}$ are estimated to be important (results available upon request). The end-point of the α_i distribution at $-\infty$, p_0, is insignificant in all three states. There was insufficient variation in the data for p_1 to be determined (these terms are defined in Section 4.5). Thus, the specification of h as the normal distribution appears to perform adequately. Of course, the additional flexibility allowed by introducing mass points at the two extremes of h may turn out to be important in some other data sets.

7. SENSITIVITY OF ESTIMATED SCARRING EFFECT

7.1 Estimates obtained on a left-truncated sample

As discussed in Section 4.2, previous studies left-truncate the sample without seeming to recognise that, if the survival status of the preceding child is amongst the regressors, then this will result in a (positive) bias in its estimated coefficient. To confirm this prediction and to establish the extent of the bias, estimates of the model are obtained under these conditions

[32] In the case of Kerala, we cannot reject the null $\theta=0$. But, at the same time, we cannot reject $\theta=1$. Thus, the degree of variation in the data for Kerala does not permit any firm conclusions to be drawn as to the size of θ.

(Table 3) and compared with the preferred estimates reported in Table 2. Three specifications are investigated.

First, the first-*born* child in each family is discarded from the sample. This is relevant because previous studies do not model the survival status of first-borns. As expected, the resulting 'initial conditions' problem creates a positive bias (see Section 4.2). The scarring effect increases in all three states, the percentage increase in WB and Kerala being quite dramatic (see row 2, Table 3).

The next experiment follows the previous literature in discarding all children born before a certain calendar year. Most previous studies discard observations 5 or 10 years before the data of the survey. This again introduces the initial conditions problem and, therefore, a positive bias (Section 4.2). However, if scarring has been decreasing over time (which preliminary investigation by us suggests it has), then a *smaller* scarring effect may be observed in the sub-sample of children born 5 or 10 years before the survey date as compared with the full sample of children in the data, who were born over a span of 36 years. So as to focus on the initial conditions problem and minimise time effects, the left-truncation performed in this second experiment is pushed further back in time. Data are discarded for children born before 1971 so that information for 28 years is retained, with only the initial eight years of data, corresponding to 2.35% of children, being discarded. In this now truncated sample, y_{ij-1} is, of course, undefined for the first-*observed* child in each family. In line with previous research, these children are also excluded from the estimated model (results in row 3). The scarring parameter shows the expected upward bias, and it is of roughly similar

---

The imprecision with which θ is determined probably arises from Kerala having the smallest population and the lowest incidence of mortality amongst the three regions (see Table 1).

magnitude to that obtained in row 2. Rows 2 and 3 of Table 3 establish that in the few existing papers that implicitly contain estimates of scarring, these are over-estimates.

While the literature has revealed no recognition of the initial conditions problem and left-truncation is performed in studies where it seems unnecessary, there are contexts in which left-truncation may be necessary. Thus, for example, information on breastfeeding or antenatal care may be essential to the purpose of a study and these data are only available (in the Indian NFHS and also in several other DHS surveys) for the 3-5 years preceding the survey. What can be done to mitigate the truncation-induced endogeneity bias arising in estimation of dynamic models (e.g. that contain y_{ij-1} as a regressor) with unobserved heterogeneity (α_i)? As indicated in Section 4.2, consistent estimates may be obtainable upon an endogenously truncated sample if an equation for the first-*observed* child in the truncated sample can be specified and model parameters identified. Identification, in this experiment, relies upon mother's age at birth of the index child (which is unique to each child). Results are in Row-4. The scarring estimate is similar to the preferred estimate in row-1, indicating that this strategy goes a fair way towards redressing the initial conditions problem. Also, $\theta=0$ is rejected for UP and WB, which confirms the relevance of modelling the first-observed child[33]. This result is likely to be of practical importance in cases where the researcher is constrained to work with a left-truncated sample.

7.2 Introducing preceding birth interval as a regressor

Refer Section 4.1 where it was argued that the preferred model is one without the birth interval

[33] As was the case in row-1, the estimate of θ for Kerala is too imprecisely determined for any firm conclusions to be drawn.

but that a model including this variable both indicates the bias in the scarring parameter in previous research and offers insight into the mechanism underlying scarring. In this Section we present, for comparison, results obtained when birth interval is included as an additional explanatory variable in the model.

The preceding birth interval for the index child is defined as a set of dummy variables for 8-17, 18-23, 24-29 and more than 29 months; unsurprisingly, there are no observations with a value of less than 8 months (the average birth interval for each sample is reported in Table 1).[34] It is set to zero for first-born children. The data were coded to ensure that all children in a multiple birth have the same preceding birth interval. The birth interval dummies are positive and significant and their inclusion is seen to reduce the scarring effect in each of the three states (see row 5 Table 3). In UP, the scarring coefficient (γ) remains significant but, in WB and Kerala, it is rendered insignificant. The results suggest that a mechanism generating short birth intervals is one part of the scarring story but that, at least in UP, there is also some other scarring mechanism at work. As discussed in Section 4.1, these results are only tentative since the endogeneity of the birth interval has not been addressed in this experiment.

8. CONCLUSIONS

This paper has investigated the clustering of sibling infant deaths in India. In a departure from previous research in this area, the main aim of the paper was to identify the degree of scarring (defined in Section 1). Scarring is of considerable theoretical interest, contributing to understanding the inter-relations of family behaviour, fertility and mortality. It is also clearly

of interest to policy-making. As indicated in Section 1, evidence of scarring raises the payoff to interventions that reduce mortality. It can also be useful in targeting interventions at the most vulnerable households. More specific policy insight depends upon identifying the mechanism underlying scarring (see below). The paper offers some improvements on previous specifications in the literature that are argued to be important in obtaining an unbiased estimate of scarring after controlling for the confounding effects of observed and unobserved inter-family heterogeneity. The statistical issues raised in this paper are expected to be widely applicable in further demographic research. The Indian National Family Health Survey analysed here is one of about 69 Demographic and Health Surveys (DHS) available for low and middle income countries. The DHS data typically contain information on all children of a mother including the first-born. As discussed, data on first-born children have quite consistently been thrown away and it is argued here that this not only constitutes a considerable loss of information but is also a source of bias in dynamic models with unobserved heterogeneity. A set of testable restrictions on the model confirms the importance of some of the statistical innovations that are made. Estimation of some variants of the preferred model shows the extent of bias in the scarring parameter that would arise if some of the specification issues highlighted here were ignored.

The main result is that there is a significant degree of scarring in each of the three Indian states for which data were analysed. In order to assess the size of this effect, it is useful to consider the reduction in mortality that could be achieved if scarring were set to zero by a hypothetical policy intervention. This is estimated to be 7% in UP, 3.1% in WB and 2.9% in

---

[34] A set of four intervals is preferred to a quadratic in the birth interval because the distribution exhibits a long tail, which the quadratic form would exaggerate. The choice of intervals is guided by examination of the

Kerala. The fact that Kerala and West Bengal have smaller families (and a higher proportion of first-born children) probably limits the overall impact of scarring: the raw data also clearly indicate a greater degree of clustering in families with a larger number of children. It would be interesting to investigate, in future work, whether the degree of scarring is increasing in birth order and whether it varies with the gender of both the index child and the preceding sibling. Also, as indicated earlier, these estimates reflect average behaviour over the period under consideration. Further work investigating whether scarring has declined over time and comparing the rate of decline across states is merited.

Preliminary investigation of alternative mechanisms driving scarring suggests that shorter birth intervals following the death of a child in the family constitute an important part of the story, although birth spacing alone does not entirely account for scarring, particularly in the state of UP. To the extent that the birth spacing effect observed reflects the fecundity mechanism rather than the replacement mechanism (see Section 1), improving availability and uptake of contraception may be expected to reduce death clustering and overall mortality rates. However, if the replacement or depression mechanisms or, indeed, some other unidentified mechanism is relevant, then policy implications are less straightforward. Further research into the processes underlying scarring is merited.

---

distribution of the variable and by the demographic literature.

REFERENCES

Baulch, Robert and John Hoddinott (2000) (eds), *Economic Mobility And Poverty Dynamics In Developing Countries,* Frank Cass.

Bean, L. L., Geraldine, P. M. and Douglas L. A. (1988), Reproductive behaviour and child survival among Nineteenth Century Mormons, The thirteenth annual meeting of the Social Science History Association, Chicago, Nov.

Becker, Gary (1991), *A Treatise on the Family*, 2nd ed. Cambridge, Mass.: Harvard University Press.

Bhargava, Alok (2003), Family planning, gender differences and infant mortality: evidence from Uttar Pradesh, India, *Journal of Econometrics,* 112, 225-240.

Bolstad, W.M. and Manda, Sam O. (2001), Investigating child mortality in Malawi using family and community random effects: A Bayesian analysis, *Journal of the American Statistical Association,* Vol. 96, 12-19.

Bongaarts, John and Robert G. Potter (1983), *Fertility, Biology and Behaviour: An Analysis of the Proximate Determinants*, New York: Academic Press.

Cantrelle, P., B. Ferry, and J. Mondot (1978), Relationships between fertility and mortality in Tropical Africa, in *The Effects of Infant and Child Mortality on Fertility*, ed. S. Preston, 181-205, New York: Academic Press.

Chamberlain, G. (1984), Panel data, in S. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, North-Holland, Amsterdam, 1247-1318.

Chamberlain, G. (1985), Heterogeneity, omitted variable bias, and duration dependence,

chapter 1 in Heckman, J. J. and Singer, B. L. (eds) *Longitudinal Analysis of Labour Market Data*, Cambridge University Press.

Chen, L., S. Ahmed, M. Gesche, and W. Mosley (1974), A prospective study of birth interval dynamics in Rural Bangladesh, *Population Studies*, 28, 277-297.

Cleland, J. and Z. A. Sathar (1984), The effect of birth spacing on childhood mortality in Pakistan, *Population Studies*, 38, 401-418.

Cleland, John G. and Jerome K. van Ginneken (1989), Maternal education and child survival in developing countries: The search for pathways of influence, *Social Science and Medicine*, 27(12): 1357-1368.

Curtis, S. L., Diamond and I., and McDonald J. W. (1993), Birth interval and family effects on postneonatal mortality in Brazil , *Demography*, 33(1), 33-43.

DaVanzo, Julie and Anne R. Pebely (1993), Maternal depletion and child survival in Guatemala and Malaysia, Labor and Population Program Working Paper 93-18, RAND.

DasGupta, M. (1990), Death clustering, mothers' education and the determinants of child mortality in Rural Punjab, India , *Population Studies*, 44, 489-505.

Dreze, Jean and Amartya Sen (1997), *Indian Development: Selected Regional Perspectives*, Oxford: Clarendon Press.

Gribble, J.N. (1993), Birth intervals, gestational age and low birth weight: are the relationships confounded?, *Population Studies*, 47, 133-146.

Guo, Guang (1993), Use of sibling data to estimate family mortality effects in Guatemala, *Demography*, Vol. 30, No. 1, February, 15-32.

Heckman, J. J. (1978), Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence , *Annales de L'INSEE*, no. 30-31, 227-269.

Heckman, J. J. (1981a), Statistical models for discrete panel data, in *Structural Analysis of Discrete Data with Econometric Applications*, ed. C. F. Manski and D. McFadden, 114-178, Cambridge: MIT Press.

Heckman, J. J. (1981b), Heterogeneity and state dependence, in *Studies in Labor Markets*, ed. S. Rosen, Chicago, Chicago Press.

Heckman, J. J. (1981c), The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process, in *Structural Analysis of Discrete Data with Econometric Applications*, eds. C. F. Manski and D. McFadden, 114-178, Cambridge: MIT Press.

Hobcraft, John (1993), Women's education, child welfare and child survival: A review of the evidence, *Health Transition Review*, 3(2): 159-175.

Hobcraft, J.N., J.W. McDonald, S. Rutstein (1983), Child spacing effects on infant and early child mortality , *Population Index*, 49, 585-618.

Hobcraft, J., J.W. McDonald, S. Rutstein (1985), Demographic determinants of infant and early child mortality: A comparative analysis, *Population Studies*, 39, 363-385.

Hyslop, Dean R. (1999), State dependence, serial correlation and heterogeneity in intertemporal labour force participation of married women, *Econometrica*, 67(6), 1255-94.

IIPS and ORC Macro (2000), *National Family Health Survey (NFHS-2) 1998-9*: India. Mumbai: International Institute for Population Sciences (IIPS).

Koenig, M.A., J.F. Phillips, O.M. Campbell, S. D'Souza (1990), Birth interval and childhood mortality in rural Bangladesh, *Demography,* 27, 251-265.

Lawless, J. F. (1987), Negative binomial and mixed Poisson regression, *Canadian Journal of Statistics*, 15, 209-25.

Madise, N.J. and Diamond, I. (1995), Determinants of infant mortality in Malawi: an analysis to control for death clustering within families, *Journal of Biosocial Science*, vol. 27(1), 95-106.

Miller, J. E., James T., Anne R. P., and Barbara V. (1992), Birth spacing and child mortality in Bangladesh and the Philippines *Demography*, 29, 305-318.

Muhuri, P. and Preston, S. (1991), Effects of family composition on mortality differentials by sex among children in Matlab Bangladesh, *Population and Development Review*, 17, 415-434.

Narendranathan, W. and Elias, P. (1993), Influences of past history on the incidence of youth unemployment: Empirical findings for the UK, *Oxford Bulletin of Economics and Statistics*, 55 (2), 161-185.

Nath, D.C., K.C. Land, K.K. Singh (1994), Birth spacing, breastfeeding and early child mortality in a traditional Indian society: a hazards model analysis, *Social Biology*, 41, 168-180.

Neyman, J. and Scott, E. (1948), Consistent estimates based on partially consistent

observations , *Econometrica*, 16, pp. 1-32.

Preston, Samuel H. (1985), Mortality in childhood: lessons from WFS, in John G. Cleland and John Hobcraft (eds.), *Reproductive Change in Developing Countries*, Oxford: Oxford University Press, pp. 46-59.

Rosenzweig, Mark and T. Paul Schultz (1982), Child mortality in Colombia: individual and community effects, *Health Policy and Education,* 2(2), 305-348.

Sastry, Narayan 1997a, Family-level clustering of childhood mortality risk in Northeast Brazil, *Population Studies*, Vol. 51, Issue 3, November, 245-261.

Sastry, Narayan 1997b, A nested frailty model for survival data, with an application to the study of child survival in Northeast Brazil, *Journal of the American Statistical Association*, Vol. 92, Issue 438, June, 426-435.

Scrimshaw, S. (1996), Nutrition and health from womb to tomb, *Nutrition Today*, 31, 55-67.

Stata 7 (2000), *Stata Statistical Software,* Stata Corportation.

Steer R.A., T.O. Scholl, M.L. Hediger, R.L. Fischer (1992), Self-reported depression and negative pregnancy outcomes, *Journal of Clincial Epidemiology*, October, 45(10), 1093-9.

UNDP (2003), Human Development Report.

Williamson, J. (1988), Migration and Urbanization , in H. Chenery and T.N. Srinivasan (eds), *Handbook of Development Economics*, Volume 1.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge, Massachusetts.

World Bank (2000), *India- Reducing Poverty, Accelerating Development: A World Bank Country Study*, Oxford: Oxford University Press.

Zenger, E. (1993), Siblings' neonatal mortality risks and birth spacing in Bangladesh. *Demography*, 30(3), 477-488.

Table 1: Descriptive Statistics

	Uttar Pradesh	West Bengal	Kerala
Demographic variables			
Probability of infant death [all live births]	0.116	0.076	0.036
Age of mother in 1998/9	35.2	35.3	37.3
Age of mother at first marriage	15.7	16.2	18.9
Age of mother at first birth	18.0	18.1	20.3
% women that have never used any method of contraception	54.4	16.4	15.5
% women who can read and write	10.5	25.7	52.5
Total children ever born per mother	5.5	4.2	3.3
% women with 1-2 children	28.6	50.8	59.0
% women with 3-4 children	33.1	32.2	33.4
% women with 5 or more children	38.3	17.0	7.6
Mean (median) birth interval in months[(v)]	30.4 (26)	33.0 (28)	35.2 (29)
% families with no infant deaths	69.3	83.9	92.4
% families in which all births die in infancy	1.32	0.58	0.47
% multiple births	0.68	0.75	0.76
% first-born children	24.4	33.9	39.3
Probability of infant death amongst first-borns	0.133	0.082	0.039
Economic & infrastructure variables			
Rank in per capita income	12	6	8
Growth rate	2.2	3.2	3
Poverty incidence	40.2	26	29.2
Toilet facility	26.7	45.1	85.2
Electricity	36.6	36.7	71.8
Population and sample size			
Population share	17.1	7.91	3.2
Population in millions	171.5	79.3	32.4
Number of mothers in sample	7297	3606	2340
Number of live births in sample	29937	10627	5950

Notes:

(i) The demographic variables and the sample sizes are authors' calculations from the Indian NFHS-II and refer to the period spanned by the entire fertility history of women aged 15-49 in 1998-9. Unless otherwise indicated, figures are sample averages.

(ii) The economic variables are from World Bank (2000). Poverty incidence is for 1994, the growth rate of economy is for the period 1991-2 to 1996-7 and the ranking of states by per capita income is for 1996-97. The growth rate and rankings use the 1980/81-based GDP series.

(iii) The toilet and electricity data are from the NFHS-II Fact Sheets in the NFHS-II final report (2000).

(iv) Population is as recorded by the Registrar-General's Office of the 2000 Census on 1 July 2000.

(v) This is the average preceding birth interval and so it is calculated on a sample excluding first-born children.

Table 2: Clustering and Scarring in Sibling Infant Deaths

	Uttar Pradesh	West Bengal	Kerala
[1] ***Incidence of infant death***	0.1164	0.0759	0.0356
Panel 1: Raw Data Probabilities			
[2] *Prob*(y_{ij}=1|y_{ij-1}=1)	0.241	0.194	0.125
[3] *Prob*(y_{ij}=1|y_{ij-1}=0)	0.092	0.060	0.029
[4] Persistence due to y_{ij-1} (difference measure)([2]-[3])	0.150	0.134	0.096
[5] Persistence due to y_{ij-1} (ratio measure) ([2]/[3])	2.63	3.25	4.31
Panel 2: Estimated Conditional Probabilities			
[6] *Prob*(y_{ij}=1|y_{ij-1}=1, .)	0.1472	0.0675	0.0601
[7] *Prob*(y_{ij}=1|y_{ij-1}=0, .)	0.0838	0.0475	0.0324
[8] Persistence due to y_{ij-1} (diff measure) ([6]-[7])	0.0635	0.0201	0.0277
[9] Persistence due to y_{ij-1} (ratio measure) ([6]/[7])	1.76	1.42	1.86
[10] % Raw persistence explained ([8]/[4])	42.3	15.0	28.9
[11] Predicted probability of infant death, adjusting for (p_0)	0.1006	0.0550	0. 0290
[12] % reduction in mortality if γ=0 (with respect to [11])	6.96	3.09	2.92
[13] θ [z: θ=0]	0.76 [4.4]	0.88 [3.6]	1.80 [0.6]
[14] [z: θ=1]	[1.4]	[0.5]	[0.3]
[15] % variance explained by α_i	12.0	22.0	7.2
[16] LR test: σ_α^2 =0 (p-value)	74.5 (0.0)	40.6 (0.0)	3.7 (0.03)
[17] Probability mass at -∞ = p_0 (standard error)	$2.3\text{x}10^{-5}$ (0.00048)	$0.48\text{x}10^{-5}$ (0.028)	0.15 (0.25)
[18] Maximised value of log likelihood	-10083.41	-2599.72	-818.73
[19] Number of women in sample	*7297*	*3606*	*2340*
[20] Number of children	*29937*	*10627*	*5950*

Notes:

(i) The estimates in Panel 2 are obtained by maximising the likelihood function in equation (5), corresponding to joint estimation of equations (1) and (4) in the text. F is logit, h is normal with mass point at -∞ (see Section 4.5). In addition to the previous child's survival status the equations include child gender, mother's education, father's education, an indicator for whether the child is one of a multiple birth, dummy variables denoting the birth order of the index child, indicators of ethnicity and religion, a quadratic in the age of the mother at the birth of the index child and cohort dummies. The effects of the covariates were allowed to be different between the two equations as specified in (1) and (4). A full set of results is available from the authors. The dependent variable y_{ij} is1 if child j in family i died before the age of 12 months and zero otherwise.

(ii) [6] is obtained by using the estimated parameters to predict y_{ij} for each observation under the condition that y_{ij-1}=1, and then averaging over all observations. [7] is similarly obtained by setting y_{ij-1}=0.

(iii) [11] is the probability of infant death predicted by the model when the scarring coefficient (γ) is set to zero. Note that this prediction is for the sample of all children, including first-borns since the model includes a (reduced-form) equation for first-borns. It is adjusted for p_0, the estimated proportion of families (p_0) who never experience an infant death.

(iv) For [13], [14], see Section 4.2. For [16] see Section 4.5. For [17] see equation (7).

Table 3

Scarring Estimates Under Alternative Sample Selections and Specifications)

Specification	**Uttar Pradesh**	**West Bengal**	**Kerala**
1. Preferred model (Table 2, [8])	0.0635* *(29937)*	0.0201* *(10627)*	0.0277* *(5950)*
2. Drop first-borns	0.074* *(22640)*	0.049* *(7021)*	0.046* *(3610)*
3. Left truncate & drop first observation	0.0733* *(22026)*	0.0553* *(6709)*	0.0379* *(3466)*
4. Left truncate but model first observation	0.0671* *(29316)*	0.0246* *(10302)*	0.0253* *(5801)*
5. Add birth interval	0.048* *(29937)*	0.015 *(10627)*	0.021 *(5950)*

Notes: Refer discussion in Section 7 of the text. Reported figures are marginal effects of scarring computed by the difference measure (see Notes to Table 2). An asterisk indicates significance of the estimated coefficient, γ, at the 5% level. Figures in parentheses are the number of observations used in the estimation.

Appendix: Table 1 –

Means and Standard Deviations of Variables Used in the Analysis

	INDIA		*UP*		*WB*		*Kerala*	
	mean	***s.d..***	**mean**	***s.d..***	**mean**	***s.d..***	**mean**	***s.d..***
Infant mortality	0.08	*0.27*	0.12	*0.32*	0.08	*0.26*	0.04	*0.19*
Infant mortality (sibling)	0.07	*0.25*	0.10	*0.30*	0.07	*0.25*	0.03	*0.16*
Female	0.48	*0.50*	0.47	*0.50*	0.49	*0.50*	0.48	*0.50*
Multiple birth	0.01	*0.11*	0.01	*0.12*	0.02	*0.12*	0.02	*0.12*
Birth order 1	0.30	*0.46*	0.24	*0.43*	0.34	*0.47*	0.39	*0.49*
Birth order 2	0.25	*0.43*	0.21	*0.41*	0.26	*0.44*	0.32	*0.47*
Birth order 3	0.18	*0.39*	0.17	*0.38*	0.17	*0.37*	0.16	*0.37*
Birth order 4	0.12	*0.32*	0.13	*0.34*	0.10	*0.30*	0.07	*0.25*
Birth order 5	0.07	*0.26*	0.09	*0.29*	0.06	*0.23*	0.03	*0.17*
Birth order >5	0.08	*0.27*	0.13	*0.34*	0.07	*0.25*	0.03	*0.16*
Hindu	0.76	*0.43*	0.82	*0.38*	0.73	*0.45*	0.47	*0.50*
Muslim	0.14	*0.34*	0.17	*0.37*	0.25	*0.43*	0.38	*0.48*
Other religion	0.10	*0.30*	0.01	*0.09*	0.02	*0.15*	0.15	*0.36*
Scheduled caste	0.18	*0.39*	0.20	*0.40*	0.23	*0.42*	0.09	*0.28*
Scheduled tribe	0.13	*0.34*	0.02	*0.14*	0.06	*0.23*	0.01	*0.10*
Caste data missing	0.01	*0.09*	0.05	*0.22*	0.00	*0.07*	0.00	*0.00*
Ma educ missing	0.00	*0.02*	0.00	*0.02*	0.00	*0.04*	0.00	*0.00*
Ma no education	0.60	*0.49*	0.75	*0.43*	0.50	*0.50*	0.11	*0.32*
Ma incomp primary ed	0.10	*0.30*	0.05	*0.21*	0.20	*0.40*	0.20	*0.40*
Ma complete prim ed	0.07	*0.26*	0.08	*0.27*	0.05	*0.22*	0.09	*0.28*
Ma incomp sec ed	0.13	*0.33*	0.06	*0.24*	0.16	*0.36*	0.32	*0.47*
Ma secondary, higher	0.10	*0.30*	0.06	*0.24*	0.09	*0.28*	0.28	*0.45*
Pa educ missing	0.00	*0.05*	0.00	*0.06*	0.01	*0.09*	0.00	*0.06*
Pa no education	0.32	*0.47*	0.33	*0.47*	0.30	*0.46*	0.07	*0.26*
Pa incomp primary ed	0.12	*0.32*	0.07	*0.25*	0.22	*0.41*	0.20	*0.40*
Pa complete prim ed	0.09	*0.28*	0.11	*0.31*	0.06	*0.23*	0.11	*0.32*
Pa incomp sec ed	0.22	*0.41*	0.19	*0.40*	0.22	*0.41*	0.33	*0.47*
Pa secondary ed	0.12	*0.32*	0.12	*0.33*	0.07	*0.26*	0.17	*0.37*
Pa higher ed	0.14	*0.34*	0.17	*0.37*	0.12	*0.32*	0.10	*0.30*
Age ma at birth of index child	22.92	*5.29*	23.20	*5.52*	21.97	*4.96*	23.28	*4.46*

Authors' calculations based on NFHS-2 (1998-99).