WORKING PAPERS

EAST-WEST CENTER WORKING PAPERS

Population and Health Series

No. 122, November 2010 (revised May 2011)

Further Development of Methodology for Multivariate Analysis of the Total Fertility Rate and Its Components Based on Birth History Data

Robert D. Retherford, Hassan Eini-Zinab, Minja Kim Choe, Naohiro Ogawa, and Rikiya Matsukura

EAST-WEST CENTER
COLLABORATION • EXPERTISE • LEADERSHIP
SINCE 1960

The East-West Center promotes better relations and understanding among the people and nations of the United States, Asia, and the Pacific through cooperative study, research, and dialogue. Established by the U.S. Congress in 1960, the Center serves as a resource for information and analysis on critical issues of common concern, bringing people together to exchange views, build expertise, and develop policy options.

The Center's 21-acre Honolulu campus, adjacent to the University of Hawai'i at Mānoa, is located midway between Asia and the U.S. mainland and features research, residential, and international conference facilities. The Center's Washington, D.C., office focuses on preparing the United States for an era of growing Asia Pacific prominence.

The Center is an independent, public, nonprofit organization with funding from the U.S. government, and additional support provided by private agencies, individuals, foundations, corporations, and governments in the region.

East-West Center Working Papers are circulated for comment and to inform interested colleagues about work in progress at the Center.

For more information about the Center or to order publications, contact:

Publication Sales Office
East-West Center
1601 East-West Road
Honolulu, Hawai'i 96848-1601

Telephone: 808.944.7145
Facsimile: 808.944.7376
Email: EWCBooks@EastWestCenter.org
Website: EastWestCenter.org

EAST-WEST CENTER WORKING PAPERS

Population and Health Series

No. 122, November 2010 (revised May 2011)

Further Development of Methodology for Multivariate Analysis of the Total Fertility Rate and Its Components Based on Birth History Data

Robert D. Retherford, Hassan Eini-Zinab, Minja Kim Choe, Naohiro Ogawa, and Rikiya Matsukura

Robert D. Retherford is Senior Fellow and Coordinator, Population and Health Studies, East-West Center.

Hassan Eini-Zinab is Project Assistant, Population and Health Studies, East-West Center.

Minja Kim Choe is Senior Fellow, Population and Health Studies, East-West Center.

Naohiro Ogawa is Director, Nihon University Population Research Institute, Tokyo.

Rikiya Matsukura is Staff Researcher, Nihon University Population Research Institute, Tokyo.

This working paper is one of the outputs for NIH grant 1R01HD057038 (Multivariate Analysis of the Total Fertility Rate and Its Components. Principal Investigator: Robert D. Retherford).

Table of Contents

Abstract

A discrete-time survival model (the complementary log-log model) is used to model parity progression from woman's own birth to first marriage, from first marriage to first birth, from first birth to second birth, and so on, with one model for each parity transition. Predictor variables in each model include woman's age and duration in parity as well as socioeconomic variables. The models are applied to birth history data. Collectively the models yield estimates of marriage and birth probabilities by age, parity, and duration in parity, denoted P_{ait}, by socioeconomic characteristics. The probabilities P_{ait} are multivariate in the sense that they can be tabulated by categories or values of one socioeconomic variable while holding other socioeconomic variables constant. The probabilities P_{ait} allow construction of a multidimensional life table that follows women by age, parity, and duration in parity one year at a time from age 10 to age 50. Because the probabilities P_{ait} are multivariate, the multidimensional life table is also multivariate, as are all measures derived from it. The derived measures considered here include both period and cohort estimates of parity progression ratios (PPRs), age-specific fertility rates (ASFRs), mean and median ages at first marriage, mean and median closed birth intervals, mean and median ages at childbearing (both overall and by child's birth order), total fertility rate (TFR), and total marital fertility rate (TMFR). The methodology is tested on data from the 2003 Demographic and Health Survey of the Philippines.

Building on earlier papers (Retherford et al. 2010a, 2010b), this working paper further develops methodology for multivariate analysis of the total fertility rate (TFR) and related measures, based on birth history data from a single survey. In the earlier paper, the related measures included parity progression ratios (PPRs), mean and median ages at first marriage, mean and median closed birth intervals by child's birth order, and the total marital fertility rate (TMFR). In the present paper, the related measures additionally include age-specific fertility rates (ASFRs) and mean and median ages at childbearing by child's birth order and for all birth orders combined. For ease of exposition, these related measures are referred to collectively as "components of the TFR." The estimates of PPRs, ASFRs, TFR, and TMFR are measures of the quantum of marriage and fertility, and the estimates of mean and median ages at first marriage, mean and median closed birth intervals, and mean and median ages at childbearing are measures of the tempo or timing of marriage and fertility. The improved methodology is tested and illustrated by applying it to data from the 2003 Philippines Demographic and Health Survey (DHS).

Overview of methodology

In both the earlier and improved versions of the methodology, a discrete-time survival model — the complementary log-log (CLL) model[1] — is used to model parity progression from a woman's own birth to her first marriage (B-M), from first marriage to first birth (M-1), from first birth to second birth (1-2), and so on, with a separate model for each parity transition. In this context, a woman's parity is defined in the usual way, as the number of children that she has ever borne, except that parity 0 is subdivided into two parity states: never-married with no children and ever-married with no children. If so desired, the B-M and M-1 transitions can be replaced by a single 0-1 transition, but this is not done in this paper.

Earlier methodology.

In the earlier version of the methodology, the basic predictor variable in the CLL model for each parity transition is duration in parity. The set of predictor variables can also include additional socioeconomic characteristics of interest. (The term "socioeconomic" is used broadly here to include not only conventional socioeconomic characteristics, such as urban/rural residence and education, but also health-related and environment-related characteristics.)

Each CLL model for a particular parity transition yields an estimate of the discrete-time hazard function P_{it}, where *i* denotes starting parity and *t* denotes years of duration in parity. In this context "discrete" means one-year duration intervals. The multivariate hazard function for the particular parity transition yields a multivariate life table from which a PPR and mean and median failure times for that parity transition are calculated. In the case of the B-M transition, a

[1] The complementary log-log (CLL) link function is preferred over the more commonly used logit link function, because e^b is a relative risk in the former case and an odds ratio in the latter case, and because a relative risk is easier to interpret than an odds ratio. As a practical matter, however, parity progression probabilities by single years of duration in parity are small, implying that duration-specific odds ratios closely approximate duration-specific relative risks, so that model-predicted values of TFR and its components are very close to the same, regardless of which link function is used (Retherford et al. 2010a, 2010b).

"failure" is a first marriage, so that mean and median failure times are mean and median ages at first marriage. In the case of higher-order transitions, a "failure" is a next birth, so that mean and median failure times are mean and median closed birth intervals by child's birth order. A PPR or mean or median failure time is multivariate in the sense that it can be tabulated by categories of one socioeconomic predictor variable while holding constant the other socioeconomic predictor variables that are included in the model.[2]

The multivariate PPRs for the various parity transitions are combined into a multivariate TFR (births per woman over her reproductive lifetime) by means of the formula

$$\text{TFR} = p_B\, p_M + p_B\, p_M\, p_1 + p_B\, p_M\, p_1\, p_2 + \ldots$$
$$+ p_B\, p_M\, p_1\, p_2\, p_3\, p_4\, p_5\, p_6\, p_7\, p_8\, p_9\, p_{10}\, p_{11}\, p_{12}\, p_{13}\, p_{14} \qquad (1)$$

where p_B denotes the PPR for the B-M transition, p_M denotes the PPR for the M-1 transition, p_1 denotes the PPR for the 1-2 transition, and so on. In the case of our Philippines illustrative example, births of order 16 or higher are ignored, so that the 14-15 transition is the last transition that is considered. Each PPR is calculated as the unconditional probability of failure between the start and end of the corresponding multivariate life table for the specified parity transition. A multivariate total marital fertility rate (TMFR) is obtained by setting p_B equal to 1 in the above equation for TFR. TMFR is actually a total ever-marital fertility rate, but for simplicity we refer to it here simply as a total marital fertility rate. Because age is not included in the set of predictor variables in the underlying CLL models, the earlier methodology does not yield estimates of ASFRs and mean and median ages at childbearing.

Henceforth we shall refer to this earlier version of the methodology as the "P_{it} method". The P_{it} method is a multivariate generalization of Feeney's PPR-based method, which is not multivariate, for estimating the TFR and related measures from birth histories (Feeney 1987). For a more detailed explanation of the P_{it} method, see Retherford et al. (2010a, 2010b).

Improved methodology.

Essential features of the improved methodology, which includes age as well as duration in parity and socioeconomic variables in the set of predictor variables, are the following: In the CLL model for any particular parity transition, basic predictor variables are woman's age at starting parity, denoted by A, and woman's duration in parity, again denoted by t. Additional predictor variables are socioeconomic characteristics of interest, as in the P_{it} method. Again there is a separate model for each parity transition. Collectively, the models for the various parity transitions yield model-predicted estimates of probabilities of failure (first marriage or next birth) by age, parity, and duration in parity, and by socioeconomic characteristics. The basic failure probabilities estimated from the CLL models are denoted P_{Ait}. The probabilities P_{Ait} are then transformed (actually re-labeled, as will be explained later) into the probabilities P_{ait}, where a is woman's age at duration t. The P_{ait} are conditional probabilities of failure between t and t+1 and

[2] In the case of a continuous socioeconomic variable, one tabulates by selected values of that variable instead of categories.

simultaneously between a and a+1; i.e., they are conditional on "survival" to age a, parity i, and duration t.

The probabilities P_{ait} for a particular set of values of the socioeconomic variables allow calculation of a multidimensional life table for that set of values of the socioeconomic variables. This multidimensional life table, which spans all parity transitions, replaces the separate life tables for the various parity transitions in the P_{it} method. The three dimensions of the multidimensional life table are age, parity, and duration in parity. "Failures" in the multidimensional life table are no longer restricted to a particular type of failure but instead include all types, including first marriages, first births, second births, and so on. Starting at age 10 (an appropriate age for the Philippines, since a considerable number of first marriages and births occur below age 15), women are survived through this life table one year at a time by age, parity, and duration in parity until they reach age 50. From the multidimensional life table one can calculate not only a TFR but also all of the various components of the TFR mentioned above. All measures calculated from the multidimensional life table are consistent with each other. This means, for example, that TFR always has the same value, whether calculated from multidimensional life table ASFRs, from multidimensional life table PPRs, or more simply as total multidimensional life table births per woman. Because the P_{ait} are multivariate, the multidimensional life table is itself multivariate, as are all of the measures derived from it. Henceforth we shall refer to this elaborated version of the methodology as the "P_{ait} method".

Common features of the P_{it} and P_{ait} methods

As in the usual calculation of the TFR as the sum of directly observed ASFRs, mortality of women is ignored when calculating model-predicted values of the TFR and its components using either the P_{it} method or the P_{ait} method.

In both the P_{it} method and the P_{ait} method, the CLL model for a particular parity transition is applied not to the original "person sample" but instead to an "expanded sample" of person-year observations created from the original person observations (Allison 1982; 1995). A separate expanded data set is created for each parity transition. For each woman in the original sample who at some time in the past made it to the parity transition's starting parity, a person-year observation is created for each single-year value of duration in parity t up to the year of failure or censoring. A person-year observation is created for the year in which failure occurred (if a failure did occur), but person-year observations are not created for censored years.

Cases where both first marriage and first birth occur in the same year, cases where two births from two different pregnancies occur in the same year, and cases where multiple births from the same pregnancy occur during the year are included in the expanded data sets. This is possible because each parity transition is modeled separately, so that two or more successive parity transitions (i.e., two or more successive failures) for a particular woman can occur in the same year of age (but not in the same year of duration in parity t, which reverts back to zero immediately after the first in a set of multiple events because of the increase in the woman's parity). In the case of multiple births from the same pregnancy, births are assumed to occur sequentially, and birth orders are randomly assigned. This way of constructing the expanded data sets guarantees that the estimates of P_{it} or P_{ait} calculated from the expanded data sets incorporate all events that occurred, even when more than one event occurred in the same year of age. As will be explained in more detail later, an assumption underlying this way of modeling multiple events in a one-year age interval is that all events, whether first marriages or births, occur at the start of the interval.

The expanded person-year data set for a particular parity transition makes it easy to include time-varying socioeconomic predictor variables in the CLL model for that transition. For example, if a person moves from rural to urban, some of the person-year observations created for that person are coded as rural and some are coded as urban (which can be done if the migration data are sufficiently detailed). The CLL model can also handle time-varying effects of predictor variables, by interacting predictor variables with time or some function of time (e.g., a quadratic function of t and t^2).

In both the P_{it} method and the P_{ait} method, multivariate estimates of TFR and its components can be derived not only from cohort data but also from period data. Because the CLL model handles left-censoring as well as right-censoring, one simply bases period estimates of TFR and its components on expanded data sets that treat person-year observations before and after the period of interest as censored. Otherwise the methodology is the same for period data (pertaining to synthetic cohorts) and cohort data (pertaining to real cohorts). The only difference is how the expanded person-year data set is constructed. In our illustrative application to Philippines DHS data, “period” is defined as the 5-year period before survey, and “cohort” is defined as the earlier lifetime experience of women age 45-49 at time of survey.

Both the P_{it} method and the P_{ait} method allow calculation of model-predicted estimates of TFR and its components by categories of one socioeconomic predictor variable while controlling for (i.e., holding constant) the other socioeconomic predictor variables included in the underlying CLL models. Thus, regardless of the complexity of the underlying models, results of the estimation can be presented in simple bivariate tables or graphs, the essential meaning of which is readily understood by non-statisticians. This is a useful feature of the methodology, especially when trying to communicate results to policymakers.

In both the P_{it} method and the P_{ait} method, if socioeconomic variables are omitted from the CLL model equation for each parity transition, the model-predicted estimates of TFR and its components pertain to the total population and are then comparable to TFR and its components calculated directly from birth histories using conventional methods. In the case of cohort data, the three sets of estimates should agree closely. The agreement will not be perfect, however, because the P_{it} and P_{ait} methods impose functional forms on the data (in the underlying parity-specific CLL models) that only approximate reality. In the case of period data, one again expects estimates of TFR and its components derived by the P_{it} method to agree closely with estimates derived by conventional birth history methods. By contrast, we expect period estimates derived by the P_{ait} method to differ from estimates derived by conventional birth history methods. The reason for difference is that, in any one-year age interval, the distribution of women by parity and duration in parity will usually differ between the synthetic multidimensional life table population and the observed population.

The end products of both the P_{it} method and the P_{ait} method are model-predicted values of TFR and its components. Coefficients of predictor variables in the underlying CLL models are estimated but are of less interest, because the coefficients are both numerous and complex, involving many interactions and other non-linearities that make them difficult to interpret. TFR and its components are, moreover, usually the quantities in which the analyst or policymaker is most interested. Of course, model-predicted values of TFR and its components are only a useful supplement to, not a substitution for, an analysis of coefficients. This paper, however, does not include an analysis of coefficients, estimates of which are not shown.

Standard errors of the model-predicted estimates of TFR and its components can be calculated by the jackknife method, following the approach used in DHS surveys. This was done and significance tests were conducted in the earlier papers using the P_{it} method (Retherford et al. 2010a, 2010b). Application of the jackknife method, which is a brute-force method, was found to be very computer-intensive, however, requiring weeks of two fast desktop computers operating around the clock. Application of the jackknife method promises to be even more computer-intensive in the case of the P_{ait} method. It is not applied in this paper.

The P_{ait} method in more detail

In our illustrative application to 2003 Philippines DHS data, the calendar month of survey, being an incomplete month for most women, is omitted from the person-year data sets, in conformity with usual DHS practice. The 5-year period before survey then includes the 60 previous calendar months. The person-year data sets for cohorts also omit the incomplete calendar month of survey. Thus, for each respondent, “time of survey” refers to the end of the first complete calendar month before interview.

The illustrative application includes two socioeconomic predictor variables: urban/rural residence (specified by a dummy variable *U*) and woman’s education (specified by dummy variables E_M and E_H, representing medium and high education with low education as the reference category), as assessed at time of survey. These socioeconomic variables are treated as time-invariant over a woman’s birth history, due to lack of information about their values in each earlier year before the survey.

In the overview in the previous section, duration in parity was denoted by *t*. In the underlying CLL models, however, duration in parity is actually specified in two different ways within the same model. One specification is the counter variable *t* ($t = 1, 2, ...$), as in the previous section. In the models, *t* is treated as a continuous variable. The second specification is a set of dummy variables representing single years of duration in parity. In order to achieve a better fit to the data, the dummy variable specification is used for the basic duration-in-parity variable. The continuous specification *t* is used when duration in parity is interacted with other predictor variables in order to model time-varying effects of predictor variables.

Although ages at first marriage and births in the birth histories in the 2003 Philippines DHS (as in all DHS surveys) are specified by both year and month, we aggregate months into years, yielding person-year data sets instead of person-month data sets. This is done because monthly data sometimes result in empty cells in the dummy-variable specification of duration in parity (e.g., there are no births in the month following a previous birth), in which case the maximum likelihood estimation procedure for fitting a discrete-time survival model does not converge to a solution (Allison 1995).

The basic form of the CLL model is the same for each single-parity transition, except that, in the case of transition from woman's own birth to first marriage (B-M), duration t ranges from 1 to 30 (corresponding to ages 10 to 39, since the B-M multivariate life table starts at age 10), whereas in the case of higher-order transitions, t ranges from 1 to 10. The ranges start with one instead of zero, for reasons having to do with the way that the CLL models are estimated. After the models are fitted and the failure probabilities P_{ait} are estimated, the range 1-30 is translated to 0-29, and the range 1-10 is translated to 0-9, in conformity with usual life table notation, as explained in more detail below. First marriages after age 40 and next births after 10 years of duration in parity are rare and are ignored.

In our illustrative application to the 2003 Philippines DHS, the general form of the CLL model for the B-M transition is

$$P = 1 - \exp\{-\exp[b_0 + b_1T_1 + b_2T_2 + ... + b_{29}T_{29} + U(e_0 + e_1 t + e_2 t^2) + E_M(f_0 + f_1t + f_2 t^2) + E_H(g_0 + g_1t + g_2 t^2) + mUE_M + nUE_H]\} \quad (2)$$

where P is the predicted probability of a first marriage (also called the discrete hazard of first marriage) in a one-year duration interval; $T_1, ..., T_{29}$ are 29 dummy variables representing the first 29 of 30 duration intervals (the 30th interval being the reference category); t is a counter variable (equal to 1, 2, ..., 30) that also denotes duration interval; b_0 is an intercept term (implying that $P = 1-\exp[-\exp(b_0)]$ for the 30th duration interval when t is set to 30 and $T_1, ..., T_{29}$, U, E_M, and E_H are all set to zero), and $b_1, ..., b_{29}, e_0, e_1, e_2, f_0, f_1, f_2, g_0, g_1, g_2, m,$ and n are coefficients to be fitted, along with the intercept b_0, to the data. Starting age A does not appear in this equation, because starting age is constant at age 10 rather than variable in the sample, so that starting age is incorporated in the intercept b_0. This equation for the B-M transition is the same as in the earlier P_{it} version of the methodology (Retherford et al. 2010a, 2010b), except that the interaction terms UE_M and UE_H have been added to the set of predictor variables in order to allow a more flexible fit to the data. (The CLL models in this paper, whether in the P_{it} method or the P_{ait} method, now include all two-way interactions among the predictor variables.)

The model in equation (2) is fitted by maximum likelihood. In the model that the computer sees and fits, the term $U(e_0+e_1t+e_2t^2)$ appears as $e_0U+e_1Z_1+e_2Z_2$, where, following procedures recommended by Allison (1995), Z_1 and Z_2 are defined as $Z_1=Ut$ and $Z_2=Ut^2$. Similarly, the terms $E_M(f_0+f_1t+f_2t^2)$ and $E_H(g_0+g_1t+g_2t^2)$ appear as $f_0E_M+f_1Z_3+f_2Z_4$) and $g_0E_H+g_1Z_5+g_2Z_6$), and the sum mUE_M+nUE_H appears as mZ_7+nZ_8. The model then appears to the computer as a discrete-time proportional hazards model, which is fitted in the usual way. (In effect, the variables $Z_1, ..., Z_8$ trick the model into handling non-proportionality, which appears here in the form of time-varying effects of predictor variables.)

As just noted, effects of socioeconomic predictor variables in equation (2) are specified as time-varying. For example, the effect of a one-unit increase in U (from 0 to 1) — i.e., the effect of urban relative to rural — is to multiply the underlying continuous-time hazard of a first marriage for a rural person with specified values of E_M and E_H (indicating the person's education) by $\exp(e_0+e_1t+e_2t^2+mE_M+nE_H)$, which is the relative risk. A time-varying specification of the effect of *U* on the probability of first marriage is necessary because the effect of urban residence, relative to rural residence, is to lower the probability of first marriage at younger ages and increase it at older ages, inasmuch as urban marriages tend to be postponed to later ages, relative to rural marriages. Thus the effect of urban residence on the risk of first marriage is not constant over duration in parity; i.e., the effect is not proportional. Similarly, the effect of education is modeled as time-varying, because the effect of more education is also to lower the probability of first marriage at younger ages and raise it at older ages. At higher-order parity transitions, for similar reasons as well as other reasons (Retherford et al. 2010a, 2010b), the effects of *U*, E_M, and E_H on the probability of next birth are also modeled as time-varying, again with a quadratic specification of duration in parity, as discussed shortly.

In the case of the B-M transition, values of P_t for particular values of *U*, E_M, and E_H are calculated from the fitted model in equation (2) as follows: P_1 is calculated by setting $T_1 = 1$, $T_2 = T_3 = ... = T_{29} = 0$, and $t = 1$ on the right side of the equation. P_2 is calculated by setting $T_1 = 0$, $T_2 = 1$, $T_3 = T_4 = ... = T_{29} = 0$, and $t = 2$ on the right side of the equation. And so on, up to and including P_{29}. In the case of the last duration-in-parity interval, which is the reference category for the dummy variables representing duration in parity, P_{30} is calculated by setting $T_1 = T_2 = ... = T_{29} = 0$, and $t = 30$ on the right side of the equation.

For single-parity transitions higher than birth to first marriage (B-M), the underlying model is

$$P = 1-\exp\{-\exp[b_0 + b_1T_1 + b_2T_2 + ... + b_9T_9 + A(c_0 + c_1 t + c_2 t^2) + A^2(d_0 + d_1 t + d_2 t^2) + U(e_0 + e_1 t + e_2 t^2) + E_M(f_0 + f_1t + f_2 t^2) + E_H(g_0 + g_1t + g_2 t^2) + U(h_1A + h_2 A^2) + E_M(j_1A + j_2A^2) + E_H(k_1A + k_2A^2) + mUE_M + nUE_H]\} \quad (3)$$

Age at starting parity is now included in the model, because age at starting parity is now variable instead of fixed at 10. An A^2 term is included as well as an *A* term in the set of predictor variables, because the rise and fall of fecundability as age increases suggest that the effect of starting age on parity progression will be non-linear, and that a quadratic specification of starting age may adequately capture this non-linearity. The effects of both *A* and A^2 are specified as time-varying (i.e, *t*-varying) because the effects of duration in parity on parity progression change as starting age increases, due not only to biological influences (changing fecundability) but also to behavioral influences. An example of a behavioral influence is that couples are more likely to settle into a life style with few or no children the longer they delay marriage and childbearing. Our methodology and data do not allow separate measurement of biological and behavioral influences, however.

Values of P_{Ait} for higher-order transitions are calculated from equation (3) in a manner similar to that used to calculate P_{Ait} for the B-M transition from equation (2). The main differences are that, in equation (3), the range of *t* is 10 years instead of 30 years, and *A* is variable instead of fixed at 10.

Equations (2) and (3) generate values of P_{Ait} for specified values of the socioeconomic variables. After these probabilities are generated, then, in order to simplify the multidimensional life table equations that come later, parities B, M, 1, 2, ..., 15 are re-labeled 0 to 16, without any change in the numerical values of the probabilities themselves. The variables *A* and *t* are also re-labeled by replacing *A* with $a = (A-10)+(t-1) = A+t-11$, and by replacing *t* with $t-1$. Duration in parity *t* then starts at 0 instead of 1 (consistent with conventional demographic notation), and age in the multidimensional life table starts at 0 instead of 10. As an example of the re-labeling of *A* and *t* (assuming that parity has already been relabeled), the probability $P_{Ait} = P_{27,4,3}$ is re-labeled as $P_{ait} = P_{19,4,2}$. (The formula $a = A+t-11$ also works for the B-M transition, with *A* fixed at 10.)

Because of smaller numbers of women and births at higher parities, the estimation algorithm for the CLL model in equation (3) does not always converge to a solution at higher parity transitions. Non-convergence occurs when one or more of the four cells in the 2x2 cross-classification of the dichotomous dependent variable FAILURE against any of the dichotomous predictor variables is empty (Allison, 1995). (Note that, at the level of a person-year observation, the response variable is not a probability of failure, which is unobservable, but is instead the dummy variable FAILURE (1 if failure occurred during the year, 0 otherwise).) In such cases the problem of non-convergence can sometimes be circumvented by using a quadratic specification of basic life table time; i.e., by replacing the dummy variables $T_1, T_2, ..., T_{29}$ with t and t^2. The computer program automatically tries to do this when non-convergence first occurs. Equation (3) then becomes

$$P = 1-\exp\{-\exp[b_0 + b_1 t + b_2 t^2 + A(c_0 + c_1 t + c_2 t^2) + A^2(d_0 + d_1 t + d_2 t^2) + U(e_0 + e_1 t + e_2 t^2) + E_M(f_0 + f_1 t + f_2 t^2) + E_H(g_0 + g_1 t + g_2 t^2) + U(h_1 A + h_2 A^2) + E_M(j_1 A + j_2 A^2) + E_H(k_1 A + k_2 A^2) + mUE_M + nUE_H]\} \quad (4)$$

The earlier model specification in equation (3) is preferred over the model specification in equation (4), because the dummy variable specification of basic life table time allows a more flexible fit. The variables t and t^2 are always used instead of dummy variables in the interaction terms in both model specifications, however, in order to avoid convergence problems and to keep the model from becoming unduly complicated.

Eventually, as parity increases, even equation (4) may not converge, necessitating the use of an open-parity interval. Suppose, for example, that this interval is 9+, so that the last transition is 9+ to 10+ (from ninth or higher-order birth to next birth). For this open-ended transition the computer program first tries the model

$$P = 1-\exp\{-\exp[b_0 + b_1 T_1 + b_2 T_2 + ... + b_9 T_9 + A(c_0 + c_1 t + c_2 t^2) + A^2(d_0 + d_1 t + d_2 t^2) + U(e_0 + e_1 t + e_2 t^2) + E_M(f_0 + f_1 t + f_2 t^2) + E_H(g_0 + g_1 t + g_2 t^2) + U(h_1 A + h_2 A^2)$$

$$+E_M(j_1A + j_2A^2) + E_H(k_1A + k_2A^2) + mUE_M + nUE_H + r_1 I + r_2 I^2]\} \quad (5)$$

where, for each person-year observation of parity 9 or higher within the 9+ open parity interval, I denotes the woman's parity at the start of that year (before any birth that may have occurred within that person-year observation), and A denotes starting age (i.e., age at attainment of parity 9, 10, 11, 12, 13, or 14, depending on the particular parity at the start of the person-year observation). Parity transitions beyond 14-15 are ignored. If, because of convergence problems, this equation cannot be estimated, the computer program automatically replaces the dummy variables $T_1, ..., T_9$ with t and t^2 in equation (5) and tries to fit that model. If this does not work, one must back up and try the same model for the 8+ to 9+ transition. Assuming that the model can be fitted to the 9+ to 10+ transition, the fitted model equation then generates values of P_{ait} for each starting parity between 9 and 14 by single years of age and duration in parity.

The reason why equation (5) is not used for all parity intervals is that the effects of many socioeconomic variables vary substantially by parity transition. For example, in the Philippines the level of a woman's education makes almost no difference in her probability of progression from parity 1 to parity 2, but makes a great deal of difference in her probability of progression from parity 3 to parity 4. If a single model were to be used for all parity transitions, it would be imperative to interact parity with the other predictor variables, so that a large number of three-way interactions would have to be added to the set of predictor variables. The model would then become unduly complicated and prone to convergence problems. We therefore use equation (5) only when absolutely necessary, namely to model behavior in the open parity interval where three-way interactions are probably not very important. Moreover, there are very few births in the open parity interval, so errors in model specification pertaining to that interval tend not to have much impact on the model-predicted fertility measures.

The creation of the expanded data set for the open parity interval requires further explanation. In the case of a 9+ to 10+ interval, the approach is to create separate expanded data sets for transitions 9-10, 10-11, ..., 14-15. The handful of births of order 16 and higher are ignored. The separate expanded data sets are then pooled to form the expanded data set for the transition 9+ to 10+. A woman can contribute person-year observations to more than one of the individual data sets that are pooled. For example, if the open parity interval is 9+, a woman who was parity 11 at time of survey contributes person-year observations to the expanded data sets for transitions 9-10, 10-11, and 11-12.

The next step is to construct the multidimensional life table from the failure probabilities P_{ait}. Because the multidimensional life table is easier to understand if the radix is a number greater than one, we set the radix to 1,000. The choice of radix is arbitrary, however, and has no effect on the final results.[3] Because births of order 16 and higher are ignored, the parity dimension of the multidimensional life table is truncated at parity 15 (the last parity transition being 14-15).

In the underlying CLL models for parity transitions, duration in parity t ranges from 0 to 29 years for the B-M transition and from 0 to 9 years for higher-order transitions, but in the

[3] The computer programs set the radix to 1.

multidimensional life table, duration in parity can extend to higher numbers of years. In the case of the B-M transition, for example, duration in parity can be as high as 40 years (the difference between ages 10 and 50). This is handled by setting $P_{a,0,t} = P_{t,0,t} = P_{a,0,a} = 0$ when $t > 29$ or, equivalently, $a > 29$. In the case of higher-order transitions, it is handled by setting $P_{ait} = 0$ when $i > 0$ and $t > 9$.

In sum, the ranges of *a*, *i*, and *t* in the multidimensional life table are as follows: Age *a* ranges from 0 to 39 (prior to re-labeling, ages 10 to 49). Parity *i* ranges from 0 to 16 (prior to re-labeling, B, M, 1, 2, ...,15). Duration in parity *t* ranges from 0 to 39 in the B-M transition (but from 1 to 30 in the underlying CLL model for the B-M transition), and from 0 to 39 in higher-order transitions (but from 1 to 10 in the underlying CLL model for each higher-order transition).

Failures in the multidimensional life table are denoted as f_{ait}, and the number of women by age, parity, and duration in parity at any point in the life table is denoted as S_{ait}. Persons reaching parity 16 (previously parity 15) before age 40 (previously age 50) in the multidimensional life table are assumed to remain at parity 16 until they reach age 40 at the end of the life table; i.e., it is assumed that $P_{ait} = 0$ for $i > 15$.

Multidimensional life table calculation formulae for the 0-1 transition (previously the B-M transition) are then

$$S_{0,0,0} = 1{,}000 \tag{6}$$

$$S_{a,0,t} = S_{a,0,a} = S_{a-1,0,a-1}(1-P_{a-1,0,a-1}) \quad \text{for } a > 0 \tag{7}$$

$$f_{a,0,t} = f_{a,0,a} = S_{a,0,a}\, P_{a,0,a} \tag{8}$$

where $S_{a,0,t}$ denotes the number of women age *a* who have not yet had a first marriage by duration *t* (which equals age *a* in the case of the B-M transition), and $f_{a,0,t}$ denotes the number of first marriages between durations *t* and *t*+1 (equivalently, between ages *a* and *a*+1).

For higher-order parity transitions (re-labeled transitions 1-2, 2-3, and so on), basic formulae are:

$$S_{a,i,0} = \sum(S_{a,i-1,t}\, P_{a,i-1,t}) \quad \text{for } i > 0 \text{ and summation over } t \tag{9}$$

$$S_{ait} = S_{a-1,i,t-1}(1-P_{a-1,i,t-1}) \quad \text{for } a > 0 \text{ and } t > 0 \tag{10}$$

$$f_{ait} = S_{ait}\, P_{ait} \tag{11}$$

where f_{ait} now denotes the number of $(i+1)^{th}$ births to women of parity *i* between ages *a* to *a*+1 and durations *t* and *t*+1. Note that $S_{a,i,0}$ can also be written as $\Sigma f_{a,i-1,t}$ for $i>0$ and summation over *t*.

Equations (9) - (11) make the simplifying assumption, mentioned earlier, that all failures occur at the start of a one-year age or duration interval. The equations then allow that more than one event (failure) can occur at the same age *a* in the multidimensional life table, inasmuch as, in equation (9), a woman can advance from parity i-1to parity i with no change in age. This is consistent with the way that the probabilities P_{ait} are estimated from the CLL models for the various parity transitions, which, as mentioned earlier, makes it possible for a woman to experience multiple events in a one-year age interval (but not in a one-year duration interval, since t immediately changes to zero after a failure.)

A property of the multidimensional life table is that if, at any given age a, one sums S_{ait} over i and t, the sum exceeds the initial number of 1,000 women at the start of the multidimensional life table. This seemingly illogical result occurs because, as is evident from equation (9), each time a woman age a experiences an event, she is duplicated; i.e., she is counted again at the start of the age interval a to a+1. For example, application of equation (9) at the start of the 0-1 transition (previously the birth-to-first-marriage transition) yields

$$S_{0,1,0} = S_{0,0,0}\, P_{0,0,0} = f_{0,0,0} \qquad (12)$$

implying that S_a for $a = 0$, calculated by summing $S_{0,i,t}$ over i and t, will be greater than the radix of 1,000, because all women who had a first marriage at age 0 (previously age 10) get counted twice at age 0. Women who both got married and had a first birth at age 0 get counted three times. The phenomenon of women getting counted more than once occurs not only at age 0 but also at every higher age, because every time there is a failure at age a, the woman who experienced the failure is, in effect, duplicated at age a. This does not mean that there is anything wrong with the multidimensional life table formulae. It just means that summing S_{ait} over i and t does not yield the total number of women at age a. The total number of women at age a is always 1,000, because no one dies in the multidimensional life table.

Once S_{ait} and f_{ait} are calculated from the P_{ait} using the calculation formulae in equations (6) - (11), TFR is calculated as

$$\text{TFR} = (\textstyle\sum f_{ait})/1{,}000 \qquad (13)$$

where the summation is over a, i (except $i = 0$, which is omitted because failures are first marriages instead of next births), and t.

If the summation in equation (13) is over a and t only, what remains is a sum of parity-specific terms, which can also be written as a sum of birth-order-specific terms (parity being a characteristic of the woman and birth order being a characteristic of the newly born child):

$$\text{TFR} = (B_1 + B_2 + \ldots + B_{15})/1{,}000 \qquad (14)$$

where B_1 pertains to first births, B_2 pertains to second births, and so on.

The total number of first marriages in the multidimensional life table is calculated as

$$B_0 = \textstyle\sum f_{a,0,t} = \sum f_{a,0,a} \qquad (15)$$

where the summation is over a.

PPRs are then calculated as

$$p_B = B_0/1{,}000 \tag{16}$$

$$p_M = B_1/B_0 \tag{17}$$

$$p_i = B_{i+1}/B_i \;, \qquad i = 1, 2, ..., 14 \tag{18}$$

where p_B is the PPR for woman's own birth to her first marriage, p_M is the PPR for first marriage to first birth, p_1 is the PPR from first to second birth, and so on. (In the case of the left sides of equations (16) - (18), we revert to our original notation for parity.)

The total marital fertility rate is calculated as

$$\text{TMFR} = \text{TFR}/p_B \tag{19}$$

where it is assumed that all births occur subsequent to a formal first marriage or informal first union. (If this assumption is problematic, one collapses the B-M and M-1 transitions into a single 0-1 transition to start with.)

Failure rates by age, parity (i.e., re-numbered parity), and duration in parity, denoted as F_{ait}, are obtained by dividing f_{ait} by 1,000. ASFRs for single-year age groups, F_a, are calculated by summing the F_{ait} over i (excluding parity 0, in which case $F_{a,0,t}$ denotes an age-specific first marriage rate) and t. If desired, ASFRs for 5-year age groups can be obtained by summing the single-year F_a within a five-year age group and dividing the sum by five.

As mentioned earlier, the multidimensional life table is internally consistent, regarding estimates of TFR and its components. For example, when ASFRs and PPRs are derived from the multidimensional life table, TFR calculated from ASFRs (TFR_{asfr}) and TFR calculated from PPRs (TFR_{ppr}) have the same value.

Once one has values of f_{it} (obtained by aggregating f_{ait} over a), one computes mean and median ages at first marriage using values of $f_{it} = f_{0,t}$. The formula is

$$\text{Mean age at first marriage} = \sum[(f_{0,t}/B_0)(t)] + 10 \tag{20}$$

where the summation is over t. (Recall that in the case of transition to first marriage, age 10 is translated to $a = 0$, so that $t = a$.) In effect, equation (20) is a weighted average of untranslated ages between 10 and 39, where the weights are the proportion of first marriages at each age.

Similarly,

$$\text{Mean closed birth interval for the M-1 transition} = \sum[(f_{1,t}/B_1)(t)] \tag{21}$$

where the summation is over t. Formulae for mean closed intervals for higher-order transitions have a form similar to that of equation (21).

Once one has the single-year ASFRs F_a, one calculates a mean age at childbearing over all birth orders of children as

$$\text{Mean age at childbearing} = \sum[(F_a/\text{TFR})(a)] + 10 \qquad (22)$$

where the summation is over single years of age a. The formula for mean age at childbearing by child's birth order (which is i+1 if the mother's parity at the start of a one-year interval is i and i>0) is the same as equation (20), except that F_{ai} (obtained by summing the F_{ait} over t) and TFR_{i+1} (total number of births of order i+1 to women of parity i in the multidimensional life table replace F_a and TFR.[4]

The formula for median age at first marriage, based on values of $f_{it} = f_{0,t}$, is

$$\text{Median age at first marriage} = t, \text{ such that } (\sum f_{0,t})/B_0 = 0.5 \qquad (23)$$

where the summation ranges from 0 to t. The formulae for median closed birth intervals and median ages at childbearing, both overall and by child's birth order, are similar. Note that these medians are true medians. Because of truncation problems, medians in DHS survey reports are defined differently, as the age or duration by which half of the initial cohort experiences failure.

It might seem that the formula for mean age at marriage in equation (20) should be $\Sigma[(f_{0,t}/B_0)(t+0.5))]+10$ instead of $\Sigma[(f_{0,t}/B_0)(t)]+10$, and similarly for the formulae for mean closed birth intervals and mean age at childbearing in equations (21) and (22), since we are dealing with discrete one-year time intervals. It turns out, however, that t should be used instead of t+0.5, consistent with the assumption in equations (6) - (11) that all failures occur at the beginning of a one-year age or duration interval. This can be seen as follows:

Suppose that age at first marriage for a particular woman is calculated as the date of her first marriage minus the date of her own birth. This calculation can be done in at least two different ways using DHS data. The first way (which at first blush is the better of the two ways) is to calculate age at first marriage as the difference between century-month of first marriage and century-month of woman's own birth, divided by 12, without any truncation of the result to an integer value. This calculation is done for each woman and the results averaged.

The second way, which is what we actually do in this paper (we do it this way in order to achieve internal consistency in the way that century-month codes (CMC) are converted into years in the overall set of calculations), is first to convert century-month dates of a woman's own birth and her first marriage into calendar-year dates of woman's own birth and first marriage. This involves, for each date separately, a division by 12 and truncation to an integer number, such as calendar year 1960. (The conversion formula is Year = int((CMC–1)/12)+1900.) One

[4] Note that one cannot calculate a mean closed birth interval as the difference between two mean ages at childbearing by child's birth order. The reason is that not all the women who had an i^{th} birth had an $(i+1)^{th}$ birth. The same point applies to median closed birth intervals and median ages at childbearing.

then calculates a woman's age at first marriage as the difference between the calendar year of her first marriage and the calendar year of her own birth. For a particular woman, this difference may be off by one year in either direction, because of variation in the number of truncated months when calculating calendar year of the woman's own birth and calendar year of her first marriage. For example, if the woman was born 1 January 1960 and got married 31 December 1980, we compute her age at first marriage as 20 years (because all we know is the calendar years 1960 and 1980), but the true value is 21 years. If, by contrast, she was born 31 December 1960 and got married 1 January 1980, we still compute her age at first marriage as 20 years, but this time the true value is 19 years. Because a true value of 19 appears to be just as likely as a true value of 21, our best estimate is 20.0 years, not 20.5 years. When one computes mean age at first marriage over all women, these kinds of errors in the estimate of mean age at first marriage (sometimes too high and sometimes too low), which occur more or less randomly, mostly cancel out. The end result, which we have verified using Philippines DHS data, is that mean age at first marriage is very close to the same, regardless of which of the two ways of calculating mean age at first marriage is used. The same argument applies to the calculation of mean age at childbearing (both overall and by birth order).

Method for calculating unadjusted and adjusted estimates of TFR and its components

Earlier it was mentioned that the methodology allows tabulation of TFR or one of its components by categories of one socioeconomic predictor variable while controlling for (i.e., holding constant) other predictor variables. The estimates tabulated in this way are referred to here as "adjusted estimates". Adjusted estimates of TFR or one of its components are calculated using the logic of what is sometimes referred to as multiple classification analysis (MCA) (Andrews, Morgan, and Sonquist 1969; Retherford and Choe 1993). In MCA, "unadjusted" means "without controls", and "adjusted" means "with controls".

For a particular parity transition such as the B-M transition, unadjusted values of P_{ait} by, for example, urban/rural residence are calculated from a CLL model that includes U as the sole socioeconomic predictor variable. Thus, in the case of equation (2), one drops the terms containing E_M and E_H. Values of P_{ait} for urban are then calculated by setting $U = 1$ in the fitted equation, and values of P_{ait} for rural are calculated by setting $U = 0$ in the fitted equation.

Adjusted values of P_{ait} by urban/rural residence for the B-M transition are calculated from equation (2) with all of the predictor variables U, E_M, and E_H included. Education, represented by E_M and E_H, is viewed as the control variable. To obtain adjusted values of P_{ait} for urban, one sets $U = 1$ and E_M and E_H equal to their duration-in-parity-specific (i.e., t-specific) mean values in the expanded data set to which the CLL model for parity transition i to i+1 is fitted.[5] To obtain adjusted values of P_{ait} for rural, one sets $U = 0$ and E_M and E_H equal to the same t-specific mean values used to calculate the adjusted P_{ait} for urban. In this way E_M and E_H are held constant when U is varied from 0 to 1. Each parity transition has its own set of t-specific

[5] Duration-specific means of E_M and E_H are used instead of age-duration-specific means, because, for any given parity transition, some combinations of a, t, and education (E_M or E_H) may be empty (no person-year observations), so that mean values of E_M and E_H cannot be calculated. Duration-specific means are also used in the P_{it} method. Alternatively, one can hold E_M and E_H constant at their observed values for each individual person-year observation, but this approach does not work for a multivariate analysis of trends in the TFR and its components, in which the expanded samples of person-year observations are pooled over two or more surveys (Retherford et al. 2010a, 2010b).

mean values of E_M and E_H derived from the person-year data set for that parity transition. A t-specific mean value of E_M or E_H is calculated by averaging the values of E_M or E_H over all person-year observations that have the specified value of t. Duration-specific means are used instead of overall means, because duration-specific means often vary systematically over duration in parity. This is especially true in period data in countries like the Philippines, where older women tend to have lower levels of education (as well as higher fertility) than younger women.

Unadjusted and adjusted multidimensional life tables by urban/rural residence are then calculated from the unadjusted and adjusted values of P_{ait} by urban/rural residence. Unadjusted and adjusted values of TFR and its components by urban/rural residence are then calculated from the unadjusted and adjusted multidimensional life tables by urban/rural residence.

As is clear from the above discussion, the multidimensional life table part of the analysis treats age, parity, and duration in parity differently from residence and education. Age, parity, and duration in parity are multidimensional life table variables pertaining to the three basic dimensions of the multidimensional life table. Residence and education are socioeconomic variables, which are not treated as basic dimensions of the multidimensional life table. In effect, the multidimensional life table approach treats the multidimensional life table as the response variable and the socioeconomic variables as the predictor variables. A multidimensional life table is estimated for each category of one socioeconomic variable while holding constant the other socioeconomic variable (or more than one other socioeconomic variable in a more elaborate analysis). When calculating unadjusted and adjusted values of TFR from a multidimensional life table, the basic multidimensional life table variables (age, parity, and duration in parity) are never held constant.

Illustrative application to 2003 Philippines Demographic and Health Survey data

By way of illustration, the methodology is applied to 2003 Philippines DHS data. Regarding marriage, the Philippines DHS treats informal unions, of which there are many, the same way as formalized unions. Both are treated as marriages. Despite this coding, there are still some births reported by ever-married women as having occurred before first marriage (i.e., before first formalized marriage or first non-formalized union), and there are also some births reported by never-married women, from whom birth histories were also collected. We refer to these births simply as premarital births. We do not exclude any premarital births or any women who had a premarital birth. Instead, we treat all such women as newly married at the time of their first birth, by coding or re-coding date of first marriage back to the date of first premarital birth. This coding and re-coding introduce small biases in the estimates of mean and median age at first marriage and related measures that are discussed in the earlier papers (Retherford et al. 2010a, 2010b).

In what follows, the cohort analysis of the 2003 Philippines DHS is based on the previous experience of women age 45-49 at time of survey, and the period analysis, pertaining to the five-year period before survey, is based on women age 10-49 at time of survey. Predictor variables, as already mentioned, are residence (urban, rural) and education (low, medium, high). The distribution of these two groups of women by residence and education is shown in Table 1.

Table 1. Percentaged distribution of sample women by residence and education

Predictor		Women age 45-49 Unweighted	Women age 45-49 Weighted	Women age 10-49 Unweighted	Women age 10-49 Weighted
Residence					
	Urban	53	56	52	56
	Rural	47	44	48	44
Education					
	Low	45	44	37	35
	Medium	31	31	40	41
	High	24	25	24	24
Total N (women)		1,343	1,343	17,519	17,519

Notes: Women age 45-49 are the basis for the cohort analysis. Women age 10-49 are the basis for the period analysis. "Low" education means less than secondary, "medium" means some or completed secondary, and "high" means more than secondary. Percentages do not include missing cases; only four women have missing information, pertaining to education. This table and all subsequent tables and figures are based on the 2003 Philippines Demographic and Health Survey. All subsequent tables and figures incorporate weights.

Expanded samples of person-year observations for the cohort analysis and the period analysis, shown in Table 2, are created from the two groups of women in Table 1. The sample sizes in Table 2 indicate number of person-year observations in the data sets to which CLL models are fitted. Two separate data sets, one for the cohort analysis and one for the period analysis, are created for each of 16 parity transitions (B-M, M-1, 1-2, …, 14-15), for a total of 32 data sets. When fitting CLL models for open parity intervals, some of these person-year data sets are combined, as explained earlier. Further details of how the expanded sample is constructed are found in Retherford et al. (2010a).

To test the methodology, Table 3 compares selected fertility measures derived by three different methods: (1) the birth history method (not multivariate), (2) the P_{it} method, and (3) the P_{ait} method.

The fertility measures in the "birth history method" column of the Table 3 are calculated in the following way: In the case of cohort estimates of these measures in the upper half of the table, PPRs are calculated directly from the marriage and birth histories as the fraction of women at each parity who ultimately continue onward to the next parity. (Note that mortality does not enter the calculation, because of the retrospective nature of the data.) These PPRs are then used to calculate TMFR and TFR using equation (1) with single-parity transitions all the way to 14-15 (no open-parity interval). Each birth to each woman in the cohort has an age of mother at childbirth attached to it, calculated in this case as the difference between century-month of childbirth minus century-month of mother's own birth, divided by 12. Mean and median ages at childbirth (mean A_c and median A_c) are calculated directly from these ages at childbirth. When calculating these cohort measures directly from the birth histories, first marriages and births occurring in the incomplete calendar month of survey are ignored, as are first marriages after age 40 and next births at 10 or more years of duration in parity, in order improve comparability with results derived by the P_{it} and P_{ait} methods.

Table 2. Expanded sample sizes

Parity transition	Cohort analysis	Period analysis
B-M	18,307	30,418
M-1	3,392	5,209
1-2	4,907	8,589
2-3	5,423	7,995
3-4	5,261	6,524
4-5	4,524	4,922
5-6	3,288	3,376
6-7	2,422	2,353
7-8	1,713	1,692
8-9	1,101	1,058
9-10	671	636
10-11	442	471
11-12	257	250
12-13	116	96
13-14	54	37
14-15	42	27

Notes: Expanded sample sizes are numbers of person-year observations. Each cell in the table corresponds to a separate data set, for which the sample size (number of person-year observations) is shown. There are 32 data sets. For each data set, weighted and unweighted sample sizes are the same. B-M denotes the transition from a woman's own birth to first marriage, and M-1 denotes the transition from first marriage to first birth. In the period analysis, periods are the five-year period before the survey. In the cohort analysis, cohorts are defined as women age 45-49 at the time of the survey.

In the case of the period estimates in the lower half of the table, the birth history estimates of TMFR and TFR are derived by Feeney's (1987) PPR-based method. Fertility estimates derived by Feeney's method, which is not multivariate, conform more closely to the data than do corresponding estimates derived by the P_{it} method, inasmuch as Feeney's method does not impose a mathematical functional form on the data. (The P_{it} method imposes a double-exponential functional form.) As already explained earlier, the P_{it} method is a multivariate version of Feeney's method. Results from the two methods are comparable when socioeconomic variables are omitted from the CLL models that underlie the P_{it} method. The period estimates of mean A_c and median A_c in the lower half of the table are calculated from estimates of single-year ASFRs derived by the conventional birth history method.[6] When applying Feeney's method and

[6] The procedure for calculating mean A_c and median A_c is first to multiply each ASFR derived by the conventional birth history method by 100,000, so that there is a greatly inflated number of births in each single-year age group of women. (A number larger than 100,000 could be used, but 100,000 is sufficient in this case.) This inflated number of births is considerably larger than the actual number of births in the age group. The inflated births in each single-year age group are then spread out evenly over the age group, so that no two births are assigned the same age of mother at childbirth. The births from all the age groups between 10 and 49 are then ordered from low to high age of mother at childbirth. It is then a simple matter to calculate mean A_c and median A_c. A similar approach is used for calculating mean and median ages at childbirth by child's birth order derived from ASFRs by child's birth order. A similar approach is also used for calculating median age at first marriage and median closed birth intervals by the P_{it}

the conventional birth history method, first marriages and births occurring in the incomplete calendar month of survey are once again ignored, as are first marriages after age 40 and next births at 10 or more years of duration in parity, in order improve comparability with results derived by the P_{it} and P_{ait} methods.

Table 3. Comparison of selected fertility measures derived by the birth history, P_{it}, and P_{ait} methods

Fertility measure	Birth history method	Underlying CLL models omit residence and education: P_{it} method	Underlying CLL models omit residence and education: P_{ait} method	Underlying CLL models include residence and education set to their interval-specific mean values: P_{it} method	Underlying CLL models include residence and education set to their interval-specific mean values: P_{ait} method
COHORT ESTIMATES					
TMFR	4.70	4.59	4.63	4.39	4.38
TFR	4.50	4.38	4.41	4.16	4.15
Mean A_c	28.6	na	28.4	na	28.7
Median A_c	28.5	na	28.3	na	28.6
PERIOD ESTIMATES					
TMFR	3.69	3.65	3.38	3.51	3.16
TFR	3.48	3.45	3.18	3.28	2.94
Mean A_c	27.9	na	28.0	na	28.3
Median A_c	27.7	na	27.9	na	28.2

Notes: A_c denotes age at childbearing. Mean and median A_c refer to all births regardless of birth order. In the upper half of the table, pertaining to cohort estimates, the birth history estimates of TMFR and TFR are derived from PPRs calculated directly from first marriages and births by birth order. The birth history estimates of mean and median A_c are calculated directly from the ages of mother at childbirth associated with the births that occurred to women in the cohort. In the lower half of the table, pertaining to period estimates, the birth history estimates of TMFR and TFR are derived by Feeney's PPR-based method. The period TFR estimated by the conventinial birth history method, which is not shown in the table, is 3.57. The period estimates of mean and median A_c are calculated from ASFRs estimated by the conventional birth history method.

The cohort estimates with residence and educated omitted from the underlying CLL models in the first three columns of the upper half of Table 3 provide the best test of the methodology. As already discussed above, birth history estimates of the fertility measures are easily calculated directly from the birth histories of women belonging to the cohort, and these birth history estimates provide a baseline for comparison with estimates derived by the P_{it} and P_{ait} methods. We do not expect perfect agreement among estimates derived from the birth history, P_{it}, and P_{ait} methods, because, as already mentioned, the CLL models that underlie the P_{it} and P_{ait} methods impose functional forms on the data that are imperfect approximations of reality.

method, and for median age at first marriage, median closed birth intervals, and median ages at childbearing by the P_{ait} method.

When the underlying CLL models omit residence and education, however, those functional forms are quite flexible. In the case of the M-1 and higher-order transitions, the form of the CLL models underlying the P_{it} method is

$$P = 1 - \exp\{-\exp[b_0 + b_1T_1 + b_2T_2 + ... + b_9T_9]\} \tag{24}$$

and the form of the CLL models underlying the P_{ait} method is

$$P = 1 - \exp\{-\exp[b_0 + b_1T_1 + b_2T_2 + ... + b_9T_9 + A(c_0 + c_1 t + c_2 t^2) + A^2(d_0 + d_1 t + d_2 t^2)]\} \tag{25}$$

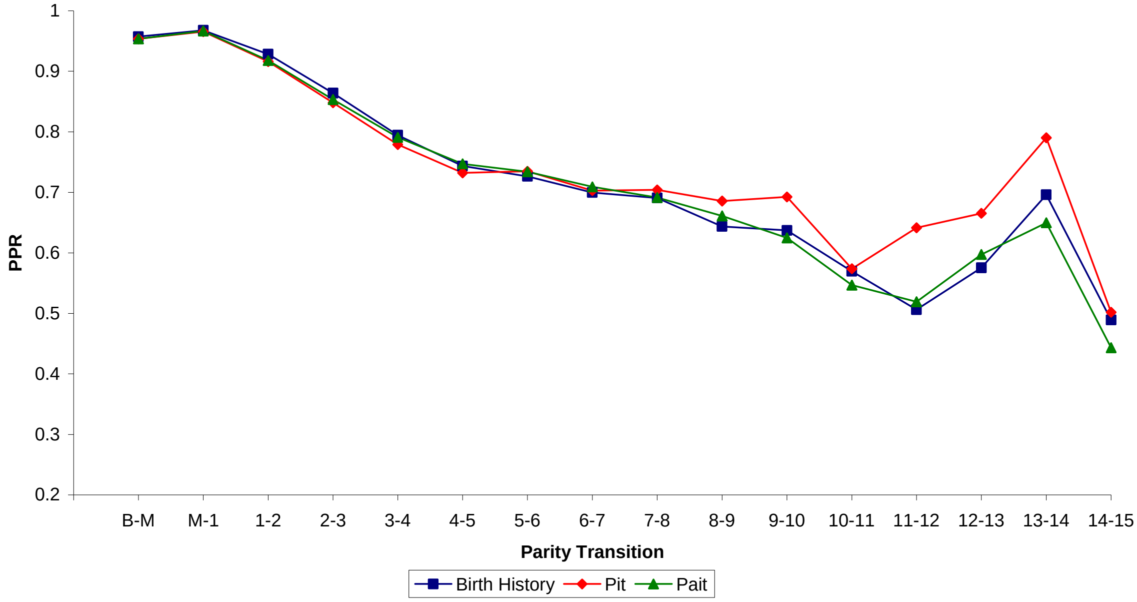
Equation (24) is very flexible, inasmuch as the only predictor variables are a set of dummy variables indicating duration in parity. Equation (25) is also very flexible, because it includes not only the dummy variables representing duration in parity but also starting age and all two-way interactions between the variables representing starting age and duration in parity.

The cohort estimates show that, when residence and education are omitted from the underlying CLL models, the birth history, P_{it}, and P_{ait} methods yield close to the same estimates of TMFR and TFR. Relative to the birth history estimates, the P_{it} method yields estimates that are 0.11 birth too low for TMFR and 0.12 birth too low for TFR. The P_{ait} method, which does slightly better, yields estimates that are 0.07 birth too low for TMFR and 0.09 birth too low for TFR. The P_{ait} method yields estimates of mean A_c and median A_c that are 0.2 year too low, relative to estimates calculated by the birth history method.

Figures 1-4 supplement the above cohort estimates with additional detail, based on the same underlying CLL models applied to the same data. Figure 1 shows that the P_{ait}-derived estimates of PPRs agree much more closely with corresponding birth history estimates than do the P_{it}-derived estimates, especially at the higher parities where numbers of births are small. Apparently the double-exponential specification in equation (24), which generates the P_{it} values from which a PPR is calculated, does not fit the data very well at the higher parity transitions. But when terms in A and A^2 are added to the set of predictor variables in equation (25), which generates the P_{ait} values, the fit is much better. Figure 2 shows that the P_{ait}-derived estimates of ASFRs agree closely with the birth-history estimates. Consistent with Figure 1, Figure 3 shows that the P_{ait}-derived estimates of mean closed birth intervals also agree much more closely with the birth history estimates than do the P_{it}-derived estimates, especially at the higher parity transitions. Figure 4 shows that the P_{ait}-derived estimates of mean age at first marriage and mean ages at childbirth by child's birth order agree closely with corresponding estimates derived by the birth history method.

We return now to the first three columns of the lower half of Table 3, pertaining to period estimates, again based on CLL models that omit residence and education from the set of predictor variables. In this part of Table 3, the birth history estimates of TMFR and TFR are derived by Feeney's PPR-based method. The birth history estimates of mean and median ages at childbearing, however, are calculated from ASFRs derived by the conventional birth history method. The P_{it}-derived estimates of TMFR and TFR in the second column agree very closely

Figure 1. Comparison of parity progression ratios (PPRs) derived by the birth history, P_{it}, and P_{ait} methods: Cohort analysis in which the CLL models that underlie the P_{it} and P_{ait} methods omit residence and education



Note: The open-ended parity transition for the P_{it} and P_{ait} methods is 13+ - 14+.

Figure 2. Comparison of age-specific fertility rates (ASFRs) derived by the birth history and P_{ait} methods: Cohort analysis in which the CLL models that underlie the P_{ait} method omit residence and education

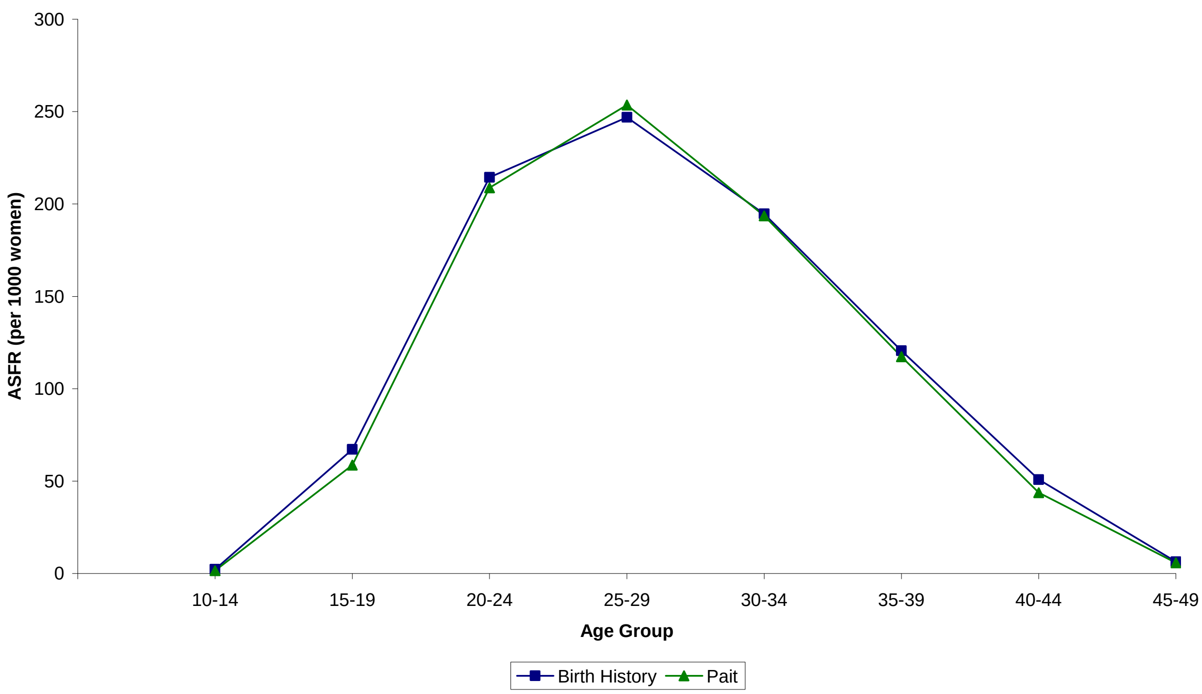


Figure 3. Comparison of mean closed birth intervals by child's birth order (CBI) derived by the birth history, P_{it}, and P_{ait} methods: Cohort analysis in which the CLL models that underlie the P_{it} and P_{ait} methods omit residence and education

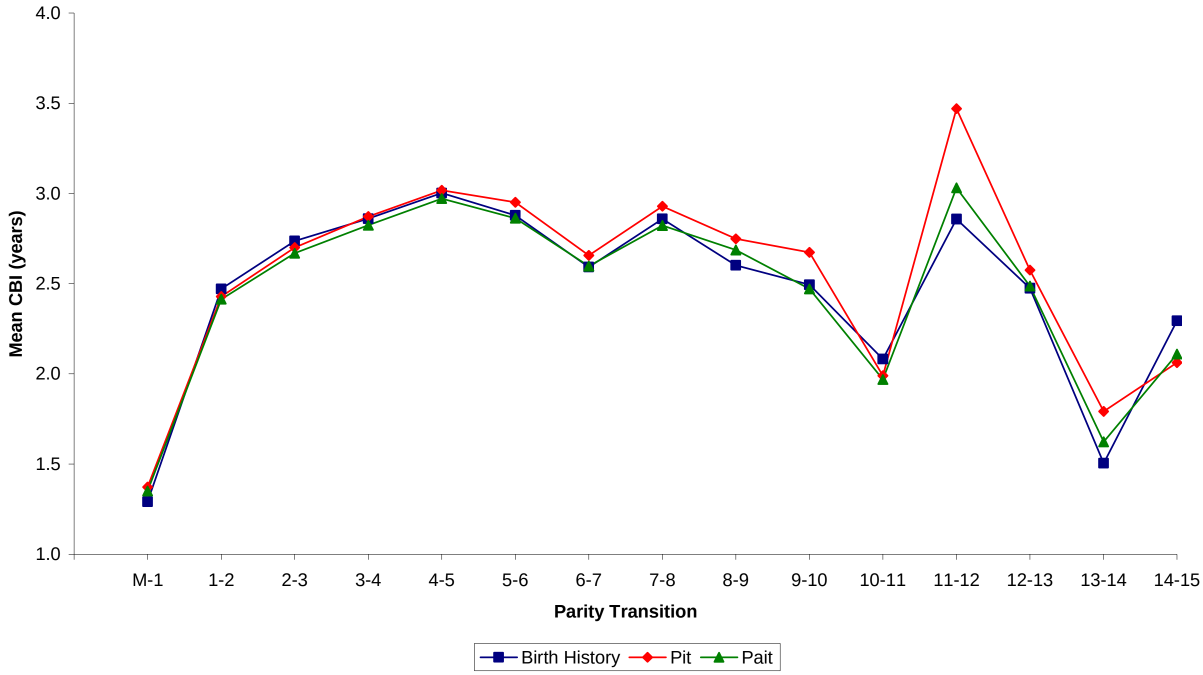
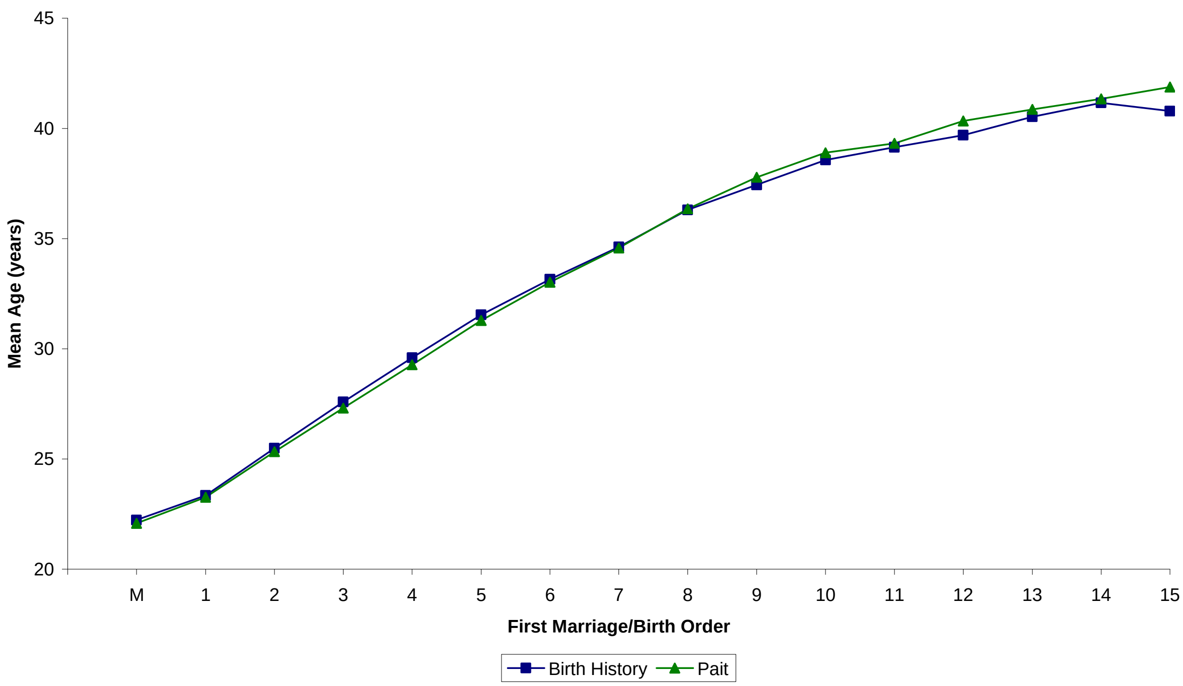


Figure 4. Comparison of mean age at first marriage and mean ages at childbirth by child's birth order, derived by the birth history and P_{ait} methods: Cohort analysis in which the CLL models that underlie the P_{ait} method omit residence and education



with corresponding estimates in the first column derived by the birth history method. By contrast, the P_{ait}-derived estimates of TMFR and TFR are considerably lower than either the corresponding P_{it}-derived estimates or the birth history-derived estimates. The P_{ait}-derived estimates of TMFR and TFR are lower than the birth history-derived estimates by 0.31 birth and 0.30 birth respectively. The main reason for these differences appears to be that, at any given age, the distribution of women by parity and duration in parity differs between the multidimensional life table population and the observed population. This part of the table also shows that, unlike the P_{ait}-derived estimates of TMFR and TFR, the P_{ait}-derived estimates of mean and median ages at childbearing (births of all orders) agree closely with birth history-derived estimates.[7]

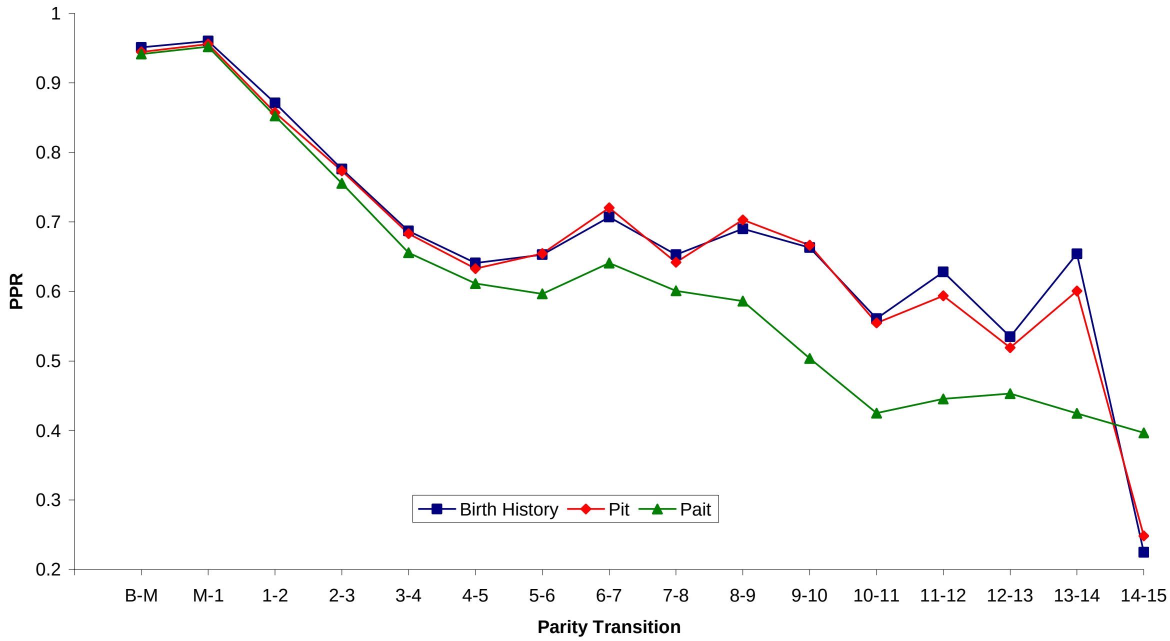
Figures 5-8 supplement the above period estimates with additional detail, based on the same underlying CLL models applied to the same data. Figure 5 shows that the P_{it}-derived estimates of PPRs agree closely with corresponding birth history estimates derived by Feeney's method, even at the higher parities. By contrast, the P_{ait}-derived estimates of PPRs are much lower than those derived by the P_{it} method or by Feeney's method, especially at the higher parities where fertility decline has been concentrated. (In this graph the P_{ait}-derived estimates of PPRs vary relatively smoothly from transition 11-12 onward because it was necessary to use an open parity interval 11+ to 12+, for which the underlying CLL model employed the variables I and I^2.) Figure 6 shows that the P_{ait}-derived estimates of ASFRs are lower than conventional birth history estimates of ASFRs, except at age 25-29. (It is not clear why 25-29 is an exception.) Figure 7 shows that the P_{it}-derived, P_{ait}-derived, and birth history estimates of mean closed birth intervals by child's birth order are close to one another at the lower parities but not so close at the higher parities where numbers of births are smaller. Figure 8 shows that the P_{ait}-derived estimates of mean age at first marriage and mean ages at childbirth by child's birth order are slightly higher than corresponding estimates calculated from age-specific first marriage rates and age-order specific birth rates derived by the conventional birth history method.[8] The graph shows almost no difference in mean age at first marriage. In the case of the P_{ait}-derived estimates of mean age at childbirth by child's birth order, however, the P_{ait}-derived estimates are almost always higher than the birth history estimates. This is not surprising, inasmuch as a lower TFR (as seen earlier in Table 3) is usually associated with delayed births.

The above findings suggest that, when residence and education are omitted from the underlying CLL models, the relatively low P_{ait}-derived estimate of the period TMFR (3.38, as shown in Table 3) may be more accurate than either the birth history estimate of the period TMFR derived by Feeney's method (3.69) or the P_{it}-derived estimate of the period TMFR (3.65). Likewise, the P_{ait}-derived estimate of the period TFR (3.18) may be more accurate than the conventional birth history estimate of the period TFR calculated from ASFRs (3.57), the estimate of the period TFR derived by Feeney's PPR-based method (3.48), or the P_{it}-derived estimate of the period TFR (3.44). The reason is that the P_{ait}-derived period estimates of TMFR and TFR are based on birth probabilities specific for age, parity, and duration in parity that are not affected by temporary distortions in population composition by parity and duration in parity at any given age in the observed population. By contrast, the estimates derived by the conventional birth history method and the P_{it} method are affected by these temporary distortions, which stem mainly from fertility variation in the past.

---

[7] See footnote 6.
[8] See footnote 6.

Figure 5. Comparison of parity progression ratios (PPR) derived by the birth history method (Feeney's method), the P_{it} method, and the P_{ait} method: Period analysis in which the CLL models that underlie the P_{it} and P_{ait} methods omit residence and education



Note: The open-ended parity transition is 13+-14+ for the P_{it} method and is 11+-12+ for the P_{ait} method .

Figure 6. Comparison of age-specific fertility rates (ASFRs) derived by the birth history and P_{ait} methods: Period analysis in which the CLL models that underlie the P_{ait} method omit residence and education

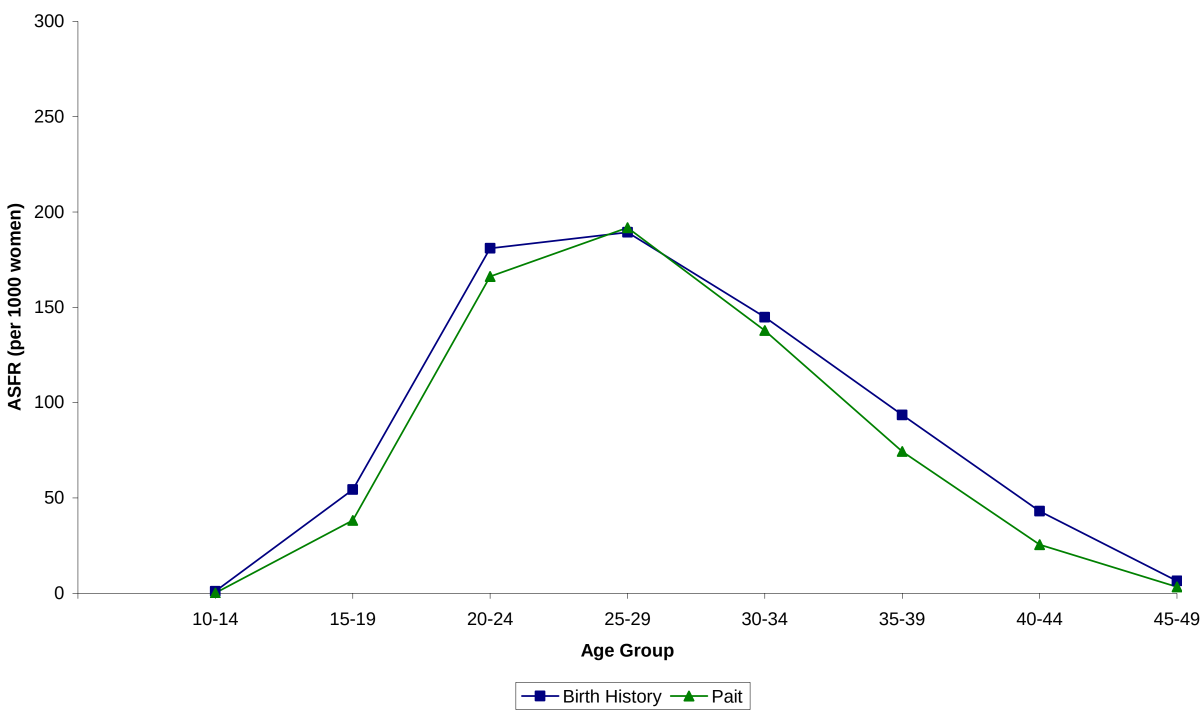


Figure 7. Comparison of mean closed birth intervals by child's birth order (CBI) derived by the birth history method (Feeney's method), the P_{it} method, and the P_{ait} method: Period analysis in which the CLL models that underlie the P_{it} and P_{ait} methods omit residence and education

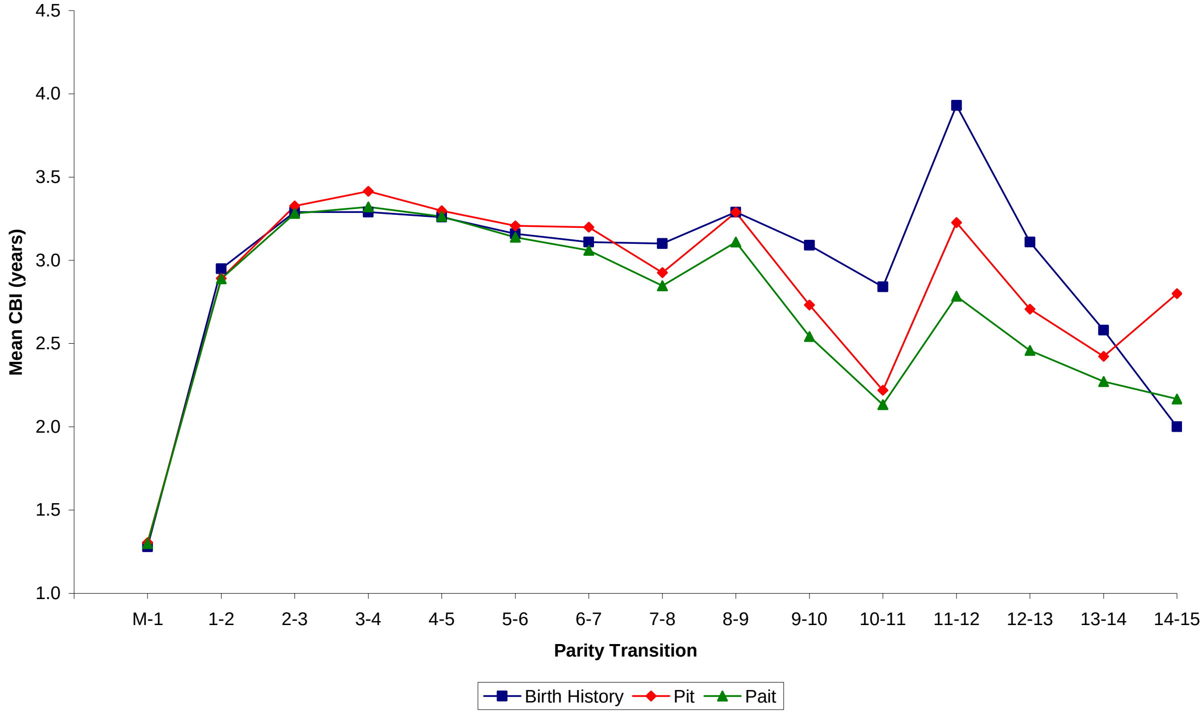
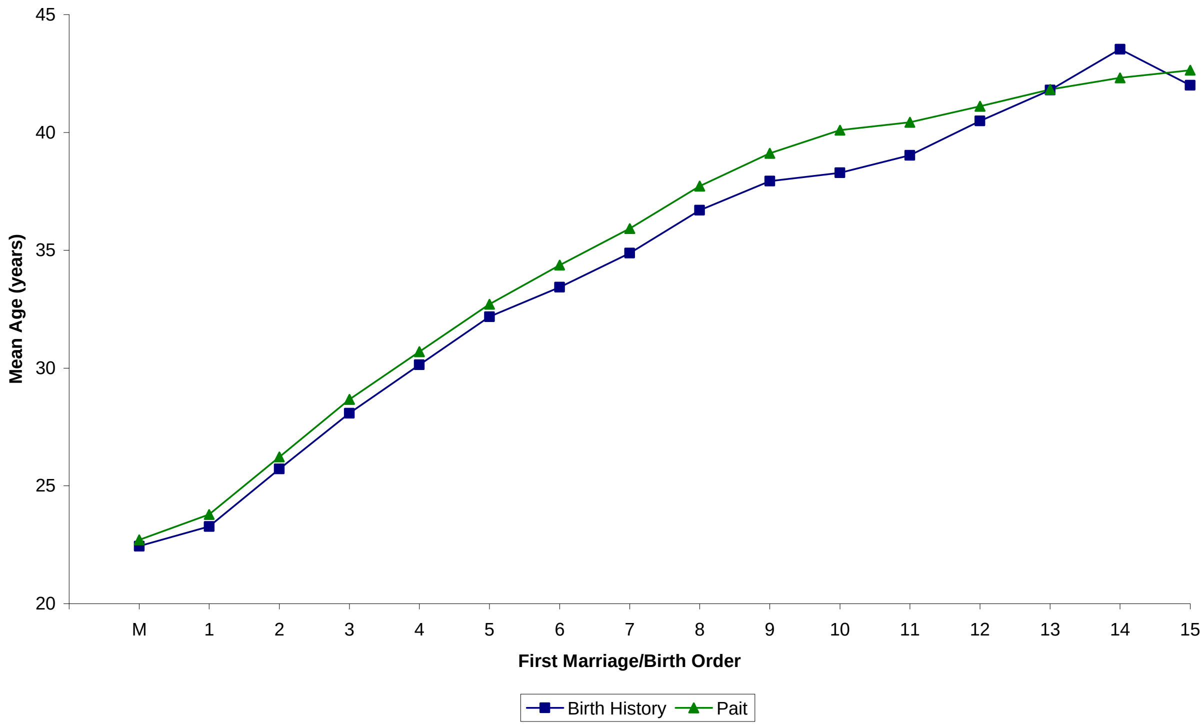


Figure 8. Comparison of mean age at first marriage and mean ages at childbearing by child's birth order, derived by the conventional birth history method and the P_{ait} method: Period analysis in which the CLL models that underlie the P_{ait} method omit residence and education



The last two columns of Table 3 show that, when the underlying CLL models include residence and education (as shown in equations (2) and (3) above) and these two variables are set to their interval-specific mean values, both the cohort and the period estimates of fertility measures derived by the P_{it} and P_{ait} methods usually agree less closely with estimates derived directly from the birth histories than they did when residence and education were not included in the CLL models. This is not surprising, given the non-linear nature of the CLL models and the multidimensional life table measures that are derived from the CLL models.[9]

The cohort estimates in the upper half of the table are of particular interest, again because they provide a better test of how well the CLL models that underlie the P_{ait} method fit the data. The P_{ait}-derived cohort estimates of TMFR and TFR in the last two columns, relative to the birth history estimates of these measures in the first column, are too low by 0.32 birth and 0.35 birth, respectively, compared with 0.07 birth and 0.09 birth too low when residence and education are omitted from the underlying CLL models. The P_{it}-derived estimates of TMFR and TFR in the last two columns, compared with the birth history estimates of these measures in the first column, show a similar pattern. On the other hand, the values of mean A_c and median A_c in the last column differ little from the values of mean A_c and median A_c in the first column, indicating that mean A_c and median A_c vary little by residence and education. The period estimates in the lower half of the table show a rather similar pattern, the major difference being that the P_{ait}-derived estimates of TMFR and TFR are considerably lower than the P_{it}-derived estimates of the TMFR and TFR.

Table 4 shows unadjusted and adjusted estimates of TFR, TMFR, mean A_c, and median A_c by residence and education, derived by the P_{ait} method. Similar tables for the other components of the TFR are voluminous and are not shown. Table 5, which is calculated from Table 4, shows the percentage of an unadjusted residence (or education) differential that is explained by education (or residence). Education explains much more of the residence differentials in TMFR and TFR than residence explains of the education differentials in TMFR and TFR, and this is so for both the cohort estimates and the period estimates. The picture is very different in the case of mean and median ages at childbearing, however. In the cohort case, the adjustments increase rather than decrease the unadjusted differentials, so that the percentage explained is negative. It is much more negative when explaining residence differentials than it is when explaining education differentials. In the period case, the percentage of the education differential that is explained by residence continues to be negative and rather small, but the percentage of the residence differential that is explained by education is reversed, from large and negative to very large and positive.

Summary and conclusion

This paper has further developed methodology for multivariate analysis of the total fertility rate and its components based on individual-level birth history data. The new, improved version of the methodology expands the set of components of the TFR to include age-based nuptiality and fertility measures as well as parity progression-based measures. The components of the TFR now

9 In a linear model, when one substitutes mean values of the predictor variables, one obtains a predicted value of the response variable that is identical to the observed mean of the response variable, but this is not generally true in non-linear models.

include parity progression ratios, age-specific fertility rates, total fertility rate, and total marital fertility rate as measures of the quantum of fertility, and mean and median ages at first marriage, mean and median closed birth intervals, and mean and median ages at childbearing (both overall and by child's birth order) as measures of the tempo or timing of marriage and fertility.

Table 4: Unadjusted and adjusted estimates of TFR, TMFR, mean A_c, and median A_c by residence and education, derived by the P_{ait} method

		TFR	TMFR	Mean A_c	Median A_c
	COHORT ESTIMATES				
Residence					
Urban	Unadjusted	3.83	4.03	28.2	28.1
	Adjusted	3.81	4.01	28.4	28.3
Rural†	Unadjusted	5.15	5.39	28.6	28.5
	Adjusted	4.63	4.91	29.0	29.0
Education					
Low†	Unadjusted	5.38	5.60	28.3	28.1
	Adjusted	5.27	5.49	28.3	28.1
Medium	Unadjusted	4.22	4.36	28.1	27.9
	Adjusted	4.24	4.38	28.2	27.9
High	Unadjusted	2.91	3.16	29.4	29.3
	Adjusted	3.03	3.29	29.4	29.4
	PERIOD ESTIMATES				
Residence					
Urban	Unadjusted	2.74	2.99	28.1	28.0
	Adjusted	2.68	2.94	28.3	28.2
Rural†	Unadjusted	3.87	3.97	27.9	27.6
	Adjusted	3.42	3.55	28.2	28.0
Education					
Low†	Unadjusted	4.53	4.63	27.4	26.9
	Adjusted	4.27	4.38	27.3	26.8
Medium	Unadjusted	3.27	3.42	27.3	27.1
	Adjusted	3.23	3.39	27.4	27.1
High	Unadjusted	2.44	2.69	29.2	29.1
	Adjusted	2.48	2.73	29.2	29.1

† Reference Category.

Note: Mean A_c and median A_c denote mean and median ages at childbearing (regardless of child's birth order).

Table 5: Percentage of an unadjusted residence (or education) differential in a fertility measure that is explained by adjustment for education (or residence)

Fertility differential	TFR	TMFR	Mean A_c	Median A_c
Cohort				
Rural minus urban	38	34	-54	-58
Low educ. minus high educ.	9	10	-9	-13
Period				
Rural minus urban	34	38	82	63
Low educ. minus high educ.	14	14	-4	-4

Notes: The percentage of a residence differential in a fertility measure that is explained by education is computed as $(1 - ((FR_{adj} - FU_{adj})/(FR_{unadj} - FU_{unadj}))) \times 100$, where FR_{unadj} denotes the unadjusted value of the fertility measure for rural, and FU_{unadj} denotes the unadjusted value of the fertility measure for urban. FR_{adj} and FU_{adj} are adjusted values. "Adjustment" in this case means that the dummy variables representing education are controlled by holding them constant at their interval-specific mean values in the underlying CLL models when comparing rural and urban. It does not matter whether the residence differential in a fertility measure is computed as FR - FU or FU - FR. The result is the same either way. The unadjusted and adjusted values of FR and FU are from Table 4, but more exact values than shown in Table 4 are used in the calculation of percentage explained. The formula for the percentage of an education differential that is explained by residence is similar to the formula for the percentage of a residence differential that is explained by education, the difference being that "adjustment" now means that the dummy variable representing residence is held constant at its interval-specific mean values in order to calculate the extent to which residence explains the education differential in the fertlity measure.

The improved methodology, like the original methodology, is based on a set of discrete-time hazard models of parity progression, with one such model for each parity transition and with duration in parity continuing to be the basic time dimension of each parity-specific model. For reasons having to do with the interpretation of effects, the hazard models have been specified as complementary log-log (CLL) models, although a discrete-time logit specification provides numerical results that are almost exactly the same. The CLL models in the improved methodology include age, as well as duration in parity and socioeconomic variables, in the set of predictor variables. Collectively, the improved models for the various parity transitions yield predicted parity progression probabilities that are specific not only for parity and duration in parity but also for age. These progression probabilities are denoted P_{ait}, where a denotes age, i denotes parity, and t denotes duration in parity. The probabilities P_{ait} are used to construct a three-dimensional "multidimensional life table" of nuptiality and fertility that follows women by age, parity, and duration in parity from age 10 to age 50. Because the probabilities P_{ait} are derived from multivariate hazard models, the entire multidimensional life table is multivariate, as are all measures (TFR and its components) derived from it. All measures derived from the multidimensional life table are consistent with one another.

For the methodology to be useful, the underlying CLL models of parity progression must fit the data well. Tests on birth history data for real cohorts indicate that they do fit well. The tests, which utilize data from the 2003 Philippines Demographic and Health Survey, show that model-predicted estimates of TFR and its components calculated from the multidimensional life table agree closely with estimates of TFR and its components calculated directly from the birth histories. To accomplish this good fit to the data, it has been necessary to include many non-linear variable specifications (quadratic specifications and interactions) in the underlying CLL models. The interactions include all two-way interactions among the predictor variables.

Because of the complexity of the underlying CLL models, coefficients of the predictor variables are numerous and difficult to interpret.[10] This difficulty has been the impetus for developing the new methodology, the goal being to transform coefficient estimates into simple bivariate tables and graphs that show how TFR and its components vary by categories of one socioeconomic variable while holding constant the other socioeconomic predictor variables included in the underlying CLL models.

Although the application to Philippines DHS data is intended to be illustrative, an important finding is that, when socioeconomic variables are dropped from the underlying CLL models so that age and duration in parity are the only predictors in the model for each parity transition, the P_{ait}-derived estimate of the period TFR (3.18) is 0.39 birth lower than the conventional birth-history estimate of the period TFR calculated from ASFRs (3.57). The main reason for this discrepancy appears to be that the conventional birth-history estimate of an ASFR is affected by temporary distortions in the composition of the age group by parity and duration in parity in the observed population, stemming primarily from past variation in the probabilities P_{ait}.[11] By contrast, the structure of the multidimensional life table population by age, parity, and duration in parity depends only on the probabilities P_{ait} from which the multidimensional life table is constructed, so there are no temporary distortions.

An implication is that, if current birth probabilities P_{ait} remain constant in the future, the P_{ait}-derived period TFR will not change in the future, but the conventional ASFR-derived period TFR will start out higher or lower than the P_{ait}-derived TFR and then gradually converge to the P_{ait}-derived TFR. Viewed from this perspective, the P_{ait}-derived TFR is more accurate than the conventional ASFR-derived TFR.

---

[10] The tables and figures presented in this paper are based on a large number of underlying CLL models that together contain a total of approximately 1,700 coefficients, including the constant term in each model.
[11] Past variation in mortality and migration can also contribute to these distortions.

Acknowledgments

Support for this research was provided by Grant 1R01HD057038 from the U.S. National Institutes of Health and by a grant obtained by the Nihon University Population Research Institute from the "Academic Frontier" Project for Private Universities: matching fund subsidy from MEXT (Japan Ministry of Education, Culture, Sports, Science and Technology), 2006–2010.

References

Allison, P. 1982. Discrete-time methods for the analysis of event histories. In S. Leinhardt (ed.), *Sociological Methodology*. San Francisco: Jossey-Bass.

Allison, P. 1995. *Survival Analysis Using SAS: A Practical Guide.* Cary, N. C.: SAS Institute Inc.

Andrews, F., J. Morgan, and J. Sonquist. 1969. *Multiple Classification Analysis*. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.

Feeney, Griffith, and Jingyuan Yu. 1987. Period parity progression measures of fertility in China. *Population Studies* 41:77–102.

Retherford, R. D., and M. K. Choe. 1993. *Statistical Models for Causal Analysis.* New York: John Wiley.

Retherford, R. D., N. Ogawa, R. Matsukura, and H. Eini-Zinab. 2010a. *Multivariate Analysis of Parity Progression-based Measures of the Total Fertility Rate and its Components Using Individual-level Data.* East-West Center Working Paper No. 119. Honolulu: East-West Center. http://www.eastwestcenter.org/fileadmin/stored/pdfs/POPwp119.pdf

Retherford, R. D., N. Ogawa, R. Matsukura, and H. Eini-Zinab. 2010b. Multivariate analysis of parity progression-based measures of the total fertility rate and its components. *Demography* 47: 97-124.