Supporting Information for 'Combining individual- and population-level data to develop a Bayesian parity-specific fertility projection model'

Joanne Ellison[1] (J.V.Ellison@soton.ac.uk), Ann Berrington[1], Erengul Dodd[1] and Jonathan J. Forster[2]
[1]University of Southampton, UK; [2]University of Warwick, UK

A | SAMPLE SELECTION PROCESS

The process of obtaining our sample of 18,218 women from the initial 27,792 interviewed in Wave 1 of the UKHLS is outlined in Figure 10. A very similar process is followed to select the UKHLS sample used in Ellison et al. (2022), the key differences being that birth cohort is not taken into account and a longer observation period is considered. To avoid repetition, we summarize the differences here and direct interested readers to Appendix A of the aforementioned paper for further details:

- Exclusion criterion 3: We exclude the comparatively smaller samples of women born in the pre-1945 and 1994 cohorts as their inclusion would introduce substantial uncertainty into our inferences.
- Exclusion criterion 7: Given our reproductive age range of 15-44, January 1960 is the first month where the 1945 cohort is observable (namely the January-born women). We choose the endpoint of November rather than December 2008 for the same reason as that in Ellison et al. (2022); however, a consequence of this is that we lose the 1993 cohort ($N = 282$ in Figure 10) from our sample, as only the January-born were observable at age 15 with the December cutoff.

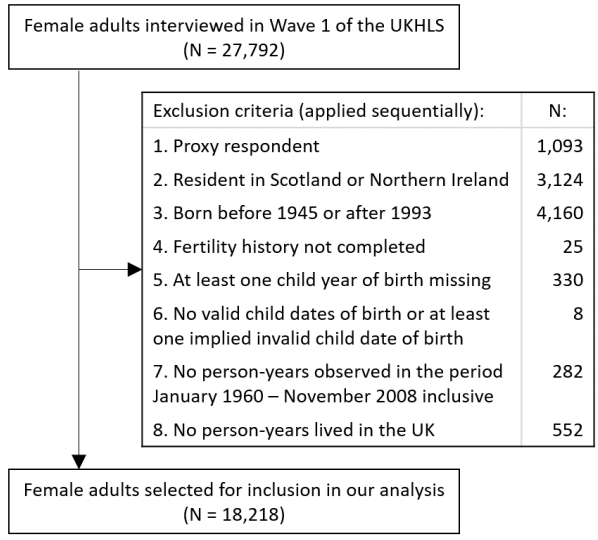

FIGURE 10 Exclusion flowchart to illustrate the selection of the women in our sample.

B | QUALIFICATION IMPUTATION MODEL

B.1 | Introduction

In Section 3.1 of the paper we mentioned the right-censoring of our highest educational qualification variable Q, caused by the youngest women still being enrolled at the time of survey and hence likely to achieve higher qualifications than those reported. To illustrate the problem, in Figure 11 we plot the unweighted proportion of women in the Q_4 categories ('< GCSE', 'GCSE', 'A Level', 'Degree'[1]) for each cohort in our sample. The proportions appear to change reasonably smoothly across cohort until we reach those born in the early 1980s (in their late twenties at the time of survey), after which the 'Degree' proportions decline to zero quite rapidly. This is because many of the women in the more recent cohorts have been interviewed before reaching the age at which we would expect a degree to have been completed if it was going to be. The locations of the change points for the lower qualification levels are staggered due to their successively younger average ages of completion. For example, the 'A Level' proportions show an unprecedented rise starting from the mid-1980s cohorts, counteracting the 'Degree' decline - many of these women are simply 'waiting' in the 'A Level' category until they are able to obtain a degree. The 'GCSE' and '< GCSE' proportions are stable up until the very youngest cohorts, where we see sharp rises in the proportions balanced by a steep decline in the 'A Level' proportion, as many of the women are now too young to have even reached *this* qualification level.

If we were to proceed to modelling with the qualification variable in its current form, we would not be able to reliably interpret the results due to this changing meaning across cohorts. Instead, we develop an imputation model that we use to assign more plausible values to the youngest women. We present the modelling approach and imputation method in Sections B.2 and B.3, and the results in Section B.4.

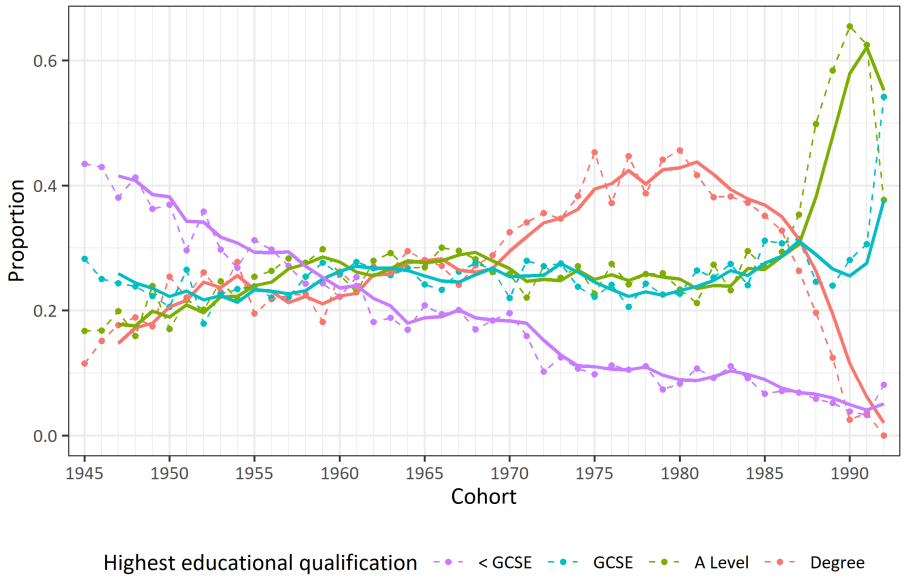

FIGURE 11 Plot of the unweighted proportion of women in the highest educational qualification categories ('< GCSE', 'GCSE', 'A Level' and 'Degree') for each of the cohorts in our sample; solid lines indicate the three-year moving average.

[1] Note that each category includes equivalent qualifications, e.g. O Levels, CSEs, which existed when the older cohorts were in education. Generally, GCSEs are taken at the end of secondary schooling (around age 16) and A Levels at around age 18; A Levels are typically a requirement for university admission.

B.2 | Modelling

To determine our imputation model, we fit various multinomial logistic regression models to the $Q = Q_4$ counts that we believe to be uncensored (we perform a sensitivity analysis to select our precise 'cutoff', i.e., the cohort after which we will impute). It is clear that C = cohort should be a regressor, and we also explore the inclusion of our remaining time-constant covariate, H = birth HDI, through its variants introduced in Figure 12.

	H_5	H_{4a}	H_{4b}	H_{4c}	H_{4d}	H_{3a}	H_{3b}	H_{3c}	H_{3d}	H_{3e}	H_{3f}	H_{2a}	H_{2b}	H_{2c}	H_{2d}
Low	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Medium	2	2	2	2	1	2	2	2	1	1	1	2	1	1	1
High	3	3	3	2	2	3	2	2	2	2	1	2	2	1	1
Very high	4	4	3	3	3	3	3	2	3	2	2	2	2	2	1
UK-born	5	4	4	4	4	3	3	3	3	3	3	2	2	2	2

FIGURE 12 Illustration of the variants of the categorical variable H (birth HDI); numbers and colours indicate the levels of each variant; H_{3b} indicates that the H variant is the second one (b) with three categories for example.

For a given cutoff c^*, we let N^* be the number of women with $C \leq c^*$. Let $\mathbf{Y}_i$ be the vector of indicator variables for each of the Q_4 categories for the ith woman, e.g., if the ith woman belongs to the 'GCSE' category then $\mathbf{Y}_i = (0, 1, 0, 0)$. Then we let $\mathbf{Y}_i \sim \text{Multinomial}(1, \pi_i^1, \pi_i^2, \pi_i^3, \pi_i^4)$, $i = 1, \ldots, N^*$, where π_i^j is the probability that the ith woman belongs to the jth Q category, $j = 1, \ldots, 4$, and $\sum_{j=1}^4 \pi_i^j = 1$. Letting Q category $j = 1$ ('< GCSE') be the reference category, we specify the model through the three equations:

$$\log\left(\frac{\pi_i^j}{\pi_i^1}\right) = \eta_i^j, \; j = 2, 3, 4, \tag{11}$$

which are fitted simultaneously. Letting $H = H_*$ be a given H variant with K categories, we experiment with various forms for the linear predictors η_i^j, which we specify in Table 3.

TABLE 3 Specifications of potential imputation models; η_i^j is the linear predictor from equation (11); $c_i \in \{1945, \ldots, c^*\}$ is the cohort of the ith woman and $h_i \in \{1, \ldots, K\}$ is her H value (in the order given in Figure 12).

Model (M)	Specification
1	$\eta_i^j = \beta_0^j \; \forall j$
2	$\eta_i^j = \beta_0^j + \beta_1^j c_i \; \forall j$
3	$\eta_i^j = \beta_0^j + \beta_1^j c_i + \sum_{k=1}^K \beta_{2,k}^j \mathcal{I}(h_i = k) \; \forall j$
4	$\eta_i^j = \beta_0^j + \beta_1^j c_i + \sum_{k=1}^K \beta_{2,k}^j \mathcal{I}(h_i = k) + \sum_{k=1}^K \beta_{3,k}^j \mathcal{I}(h_i = k) c_i \; \forall j$

We will refer to the models as M1-M4. In words, M1 just contains an intercept, M2 adds the main effect of C, M3 adds the main effect of H_* and M4 adds an interaction between C and H_*. We fit M1-M4 using the `mgcv` package in `R`, with M3 and M4 being fitted for all of the H variants in Figure 12, for a plausible set of potential cutoffs $c^* \in \{1980, \ldots, 1984\}$. We find that M4 with $H_* = H_{3e}$ gives the lowest BIC for each potential cutoff and so is our

chosen model. Solving the equations in (11) for each π_i^j (using the fact that $\sum_{j=1}^4 \pi_i^j = 1$), we obtain:

$$\pi_i^1 = \frac{1}{1 + \sum_{j=2}^4 \exp(\eta_i^j)}; \pi_i^j = \frac{\exp(\eta_i^j)}{1 + \sum_{j=2}^4 \exp(\eta_i^j)},\ j = 2, 3, 4.$$

Using this relationship to compute the fitted probabilities and forecasts generated by simple extrapolation, we find that the fits generally track the trends of the observed proportions well for each potential cutoff. We decide to set $c^* = 1982$ for the imputation because the gradients of the 'Degree' fits appear to decrease at a faster rate for the post-1982 cutoffs, particularly in the foreign-born categories; this implies that the forecasts are becoming more sensitive to the declining 'Degree' proportions close to the cutoff and therefore an earlier cutoff is preferable.

B.3 | Method

We illustrate our imputation method for a given post-1982 cohort $c \in \{1983, \ldots, 1992\}$ and H_{3e} value $h \in \{1, 2, 3\}$. For a hypothetical woman with these covariate values, let $\hat{\pi}_{c,h}^1, \ldots, \hat{\pi}_{c,h}^4$ be the extrapolated fitted probabilities generated from our chosen multinomial model. Let $n_{c,h}^{j,obs}$ be the number of women originally observed in Q category j, $n_{c,h}^{j,req}$ be the number of women required in Q category j, and $n_{c,h}^{j,cur}$ be the number of women currently in the category, initially set as the observed count $n_{c,h}^{j,obs}$. Let $n_{c,h}^{obs} = \sum_{j=1}^4 n_{c,h}^{j,obs}$ and $n_{c,h}^{req} = \sum_{j=1}^4 n_{c,h}^{j,req}$. In essence, our imputation process adjusts the number of women originally observed in each Q category to match the counts implied by the fitted probabilities as closely as possible; women can either stay in their current category or move up to a higher level, and certain checks are required to ensure that the total matches $n_{c,h}^{obs}$ throughout. We perform the following steps:

1. Let $n_{c,h}^{j,req} = \text{round}(n_{c,h}^{obs} \hat{\pi}_{c,h}^j)$, $j = 1, \ldots, 4$ be the initial required counts. If, due to the rounding, the sum of the required counts is one fewer than the sum of the observed counts (i.e., $n_{c,h}^{req} - n_{c,h}^{obs} = -1$), add 1 to $n_{c,h}^{j',req}$, where $j' = \text{argmin}_j \left(\left| \text{frac}\left(n_{c,h}^{obs} \hat{\pi}_{c,h}^j \right) - 0.5 \right| \right)$ is the category for which the decimal part of $n_{c,h}^{obs} \hat{\pi}_{c,h}^j$ is closest to 0.5 and is therefore the most borderline required count. If there is one too many in the required counts ($n_{c,h}^{req} - n_{c,h}^{obs} = 1$), subtract 1 from $n_{c,h}^{j',req}$ for the same reason.
2. Then, for $\tilde{j} = 1, 2, 3$, do the following:
   a. Compute $n_{c,h}^{\tilde{j},dif} = n_{c,h}^{\tilde{j},cur} - n_{c,h}^{\tilde{j},req}$, the difference between the current and required counts.
   b. If the difference is positive ($n_{c,h}^{\tilde{j},dif} > 0$), move the excess $n_{c,h}^{\tilde{j},dif}$ women currently in category $\tilde{j}$ to $\tilde{j}+1$, sampled at random.
   c. If the difference is negative ($n_{c,h}^{\tilde{j},dif} < 0$), fix the required count at its current value (set $n_{c,h}^{\tilde{j},req} = n_{c,h}^{\tilde{j},cur}$). Then, increase the required counts in the higher categories proportionally by computing

$$n_{c,h}^{j,req} = \text{round}\left(\frac{\hat{\pi}_{c,h}^j \sum_{k=\tilde{j}+1}^4 n_{c,h}^{k,cur}}{\sum_{k=\tilde{j}+1}^4 \hat{\pi}_{c,h}^k} \right), j = \tilde{j}+1, \ldots, 4$$

   and adjusting if $n_{c,h}^{req} \neq n_{c,h}^{obs}$ similarly to Step 1 (only possible for $\tilde{j} \in \{1, 2\}$).
   d. If the difference is zero ($n_{c,h}^{\tilde{j},dif} = 0$) move on to the next category (again, only possible if $\tilde{j} \in \{1, 2\}$), i.e., set $\tilde{j} = \tilde{j} + 1$.

We present the results of the imputation process in Section B.4.

B.4 | Results

The imputation process moves women up to higher Q categories according to our chosen multinomial model, allowing us to generate a more realistic series of proportions for the post-1982 cohorts compared to the observed proportions in Figure 11. We illustrate the impact on the unweighted (updating Figure 11) and weighted proportions by cohort in Figure 13. Regarding the former, the fact that our sample consists predominantly of UK-born women means that even though we have imputed at the level of H_{3e}, the resulting aggregate proportions are relatively smooth. Ideally we would have imputed on the weighted level directly, but this would require a more complicated process than that in Section B.3 in order to appropriately select the women who should move to higher Q categories given their different weights. Despite this, the weighted imputed proportions are only slightly more erratic than their unweighted counterparts, and definitely not at all implausible. So overall we are satisfied that this imputation suits our purposes in terms of providing a set of more appropriate and realistic Q values for the youngest women in our sample. We replace Q with this 'mean imputation' for all analyses performed in the paper.

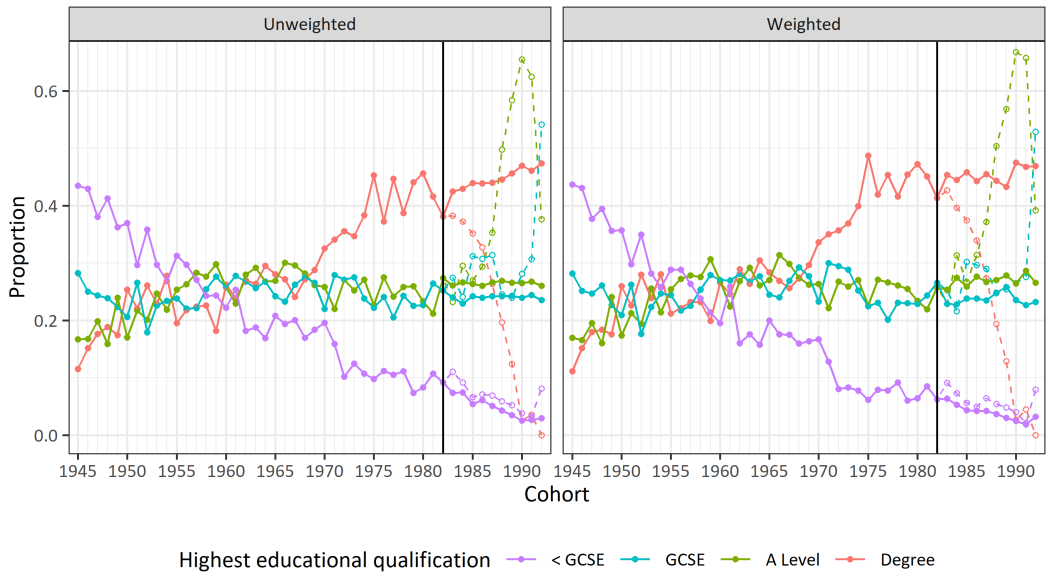

FIGURE 13 Plots of the unweighted and weighted proportions of women in the highest educational qualification categories for each of the cohorts in our sample (open circles and dashed lines), with the proportions resulting from the imputation process overlaid (filled circles and solid lines); vertical black lines indicate the 1982 cutoff.

C | FITTING GAMs IN R USING THE mgcv PACKAGE

In this appendix we give further details on the fitting process of GAMs in R using the mgcv package (Wood, 2017) and taking a P-spline approach (Eilers and Marx, 1996). We break our description down into three sections, the first two concerning the construction of the univariate and bivariate smooth functions. In each case there are two aspects to consider: the form of the function itself, and its degree of smoothness. Sections C.1 and C.2 explain how these aspects are controlled respectively via bases and penalties. In Section C.3 we then demonstrate how these components come together in the model fitting process, through the method of penalized likelihood maximization. We illustrate the concepts in the context of a binomial logistic GAM with univariate smooths of age and cohort, and a bivariate smooth of the two variables (see Section 3.2 of the paper). For convenience, we suppose we are fitting this GAM to the UKHLS dataset so that our covariate values, when taken together, constitute the set of age-cohort combinations for ages 15-44 and the 1945-1992 cohorts that are observable in the dataset; this gives $N = 1005$ covariate patterns.

C.1 | Bases

In the case of univariate smooths, the functions themselves are usually expressed as linear combinations of basis functions, where a spline basis is commonly used. A spline is a curve constructed from a series of piecewise polynomials joined together at knots, and is continuous up to its second derivative. Various spline bases exist, each with their own advantages; we opt for a cubic B-spline basis. B-splines are a popular choice due to the local nature of the basis functions, each being non-zero across the interval between a fixed number of knots (Wood, 2017). Let $\mathbf{A}$ be an $N \times k$ matrix, where a_{ij} is the jth cubic B-spline basis function evaluated at x_i, the ith value of the covariate N-vector $\mathbf{x}$. We present this unconstrained basis for dimension $k = 9$ and $x =$ age in the top-left panel of Figure 14.

A property of an unconstrained B-spline basis is that the sum of the values that the k basis functions take at any given x_i is 1 (i.e., $\sum_{j=1}^{k} a_{ij} = 1$ for $i \in \{1, \ldots, N\}$), which means that the intercept term is in the span of the basis. This is not a problem if there is only one smooth term in the model, as in this case a separate intercept term is simply not included; however it is highly undesirable if there are multiple smooth terms, as they cannot each estimate their own intercept (Wood, 2017). To this end, Wood (2017) imposes a sum-to-zero identifiability constraint on the basis using a method based on the QR decomposition.

For readers unfamiliar with the QR decomposition, we provide a brief introduction here. Let $\mathbf{Y}$ be an $s \times t$ real matrix with $s \geq t$. The QR decomposition of $\mathbf{Y}$ factors the matrix as the product of an $s \times s$ orthogonal matrix $\mathbf{Q}$ and an $s \times t$ upper triangular matrix $\mathbf{R}$, i.e., $\mathbf{Y} = \mathbf{QR}$. As the bottom $(s - t)$ rows of $\mathbf{R}$ consist entirely of zeroes, it is common to partition $\mathbf{R}$ or both $\mathbf{R}$ and $\mathbf{Q}$ as follows:

$$\mathbf{Y} = \mathbf{QR} = \mathbf{Q} \begin{bmatrix} \mathbf{R}_1 \\ 0 \end{bmatrix} = [\mathbf{Q}_1 : \mathbf{Q}_2] \begin{bmatrix} \mathbf{R}_1 \\ 0 \end{bmatrix} = \mathbf{Q}_1 \mathbf{R}_1 .$$

$\mathbf{R}_1$ is a $t \times t$ upper triangular matrix, 0 is an $(s - t) \times t$ zero matrix, and $\mathbf{Q}_1$ and $\mathbf{Q}_2$ have orthogonal columns and dimensions $s \times t$ and $s \times (s - t)$ respectively. We now return to discussing the sum-to-zero constraint.

Letting $\boldsymbol{\alpha}$ be the k-vector of basis function coefficients, the value obtained when the unconstrained smooth term is evaluated at x_i can be written as $f(x_i) = \mathbf{A}_i \boldsymbol{\alpha}$, where $\mathbf{A}_i$ is the ith row of $\mathbf{A}$. The N-vector of these values can then

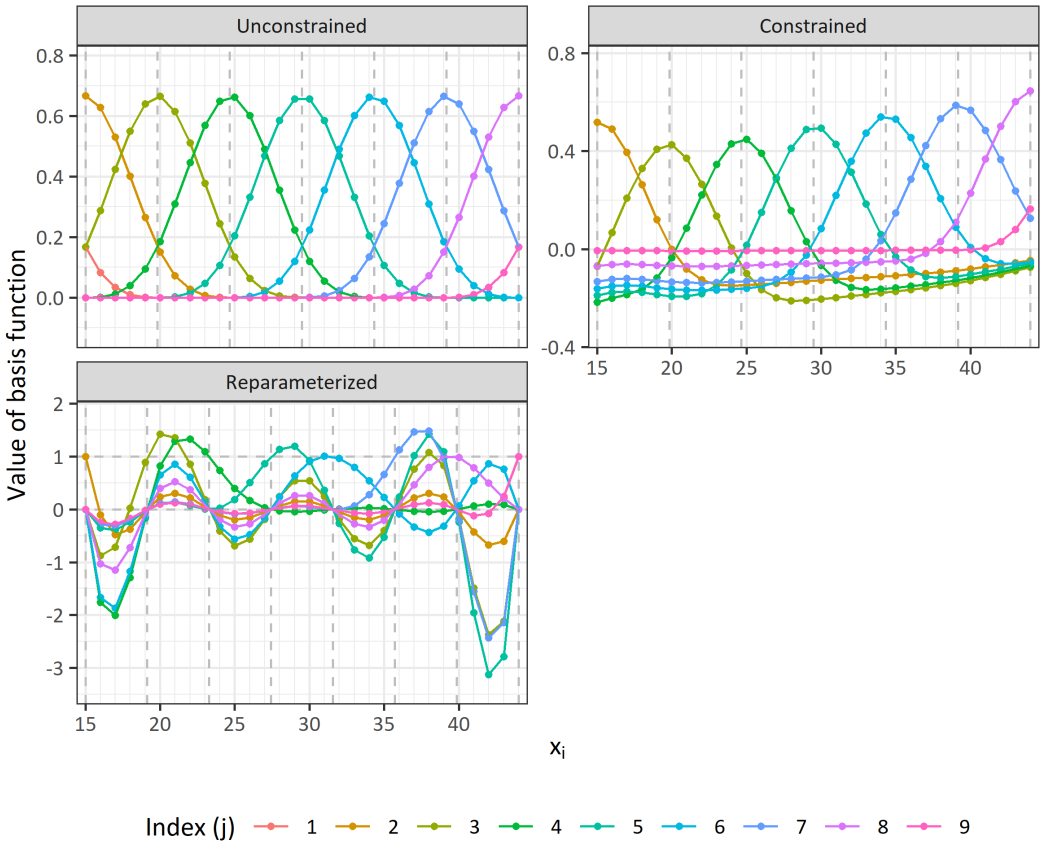

FIGURE 14 Unconstrained (a_{ij}), constrained (b_{ij}) and reparameterized (c_{ij}) cubic B-spline bases of dimension 9, 8 and 8 respectively for the covariate age; vertical dashed lines indicate knot locations for the unconstrained and constrained bases, and $\mathbf{x}^*$ for the reparameterized basis; horizontal dashed lines are at $y = 0$ and $y = 1$.

be written as $f(\mathbf{x}) = \mathbf{A}\boldsymbol{\alpha}$. A sum-to-zero constraint is enforced across f such that $\sum_{i=1}^{N} f(x_i) = 1'f(\mathbf{x}) = 1'\mathbf{A}\boldsymbol{\alpha} = 0$; this causes the smooth to be orthogonal to the intercept, allowing the "sharpest inference about the constrained components" (Wood, 2017) to be made. The constraint is implemented by finding a $k \times (k-1)$ matrix $\mathbf{Z}$ such that $1'\mathbf{AZ} = 0$. This is achieved by obtaining the QR decomposition of $\mathbf{Y} = (1'\mathbf{A})'$ and taking $\mathbf{Z} = \mathbf{Q}_2$. Then we have:

$$1'\mathbf{AZ} = (1'\mathbf{A})\mathbf{Z} = \mathbf{Y}'\mathbf{Q}_2 = (\mathbf{Q}_1\mathbf{R}_1)'\mathbf{Q}_2 = \mathbf{R}_1'\mathbf{Q}_1'\mathbf{Q}_2 = \mathbf{R}_1'0 = 0$$

as desired, with $\mathbf{Q}_1'\mathbf{Q}_2 = 0$ because $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are formed of distinct columns of $\mathbf{Q}$, which is an orthogonal matrix.

Having obtained the matrix $\mathbf{Z}$, the smooth is then reparameterized in terms of a $(k-1)$-vector $\boldsymbol{\gamma}$ such that $\boldsymbol{\alpha} = \mathbf{Z}\boldsymbol{\gamma} \Rightarrow \boldsymbol{\gamma} = \mathbf{Z}'\boldsymbol{\alpha}$, and an $N \times (k-1)$ matrix of constrained B-spline basis functions $\mathbf{B} = \mathbf{AZ}$. Then $f(\mathbf{x}) = \mathbf{A}\boldsymbol{\alpha} = \mathbf{AZ}\boldsymbol{\gamma} = \mathbf{B}\boldsymbol{\gamma}$ satisfies the constraint as $1'\mathbf{AZ} = 1'\mathbf{B} = 0$ and therefore $1'f(\mathbf{x}) = 1'\mathbf{B}\boldsymbol{\gamma} = 0\boldsymbol{\gamma} = 0$, as desired.

In the top-right panel of Figure 14 we illustrate the effect of the constraint. The degree of freedom used up

through enforcing the sum-to-zero constraint is absorbed into the new basis via the simultaneous removal of the first basis function and transformation of the rest, which introduces substantial asymmetry into the basis. This constrained basis is used by `mgcv` for the univariate smooth terms, allowing a separate intercept to be estimated without any identification issues.

Next we consider the implementation of bivariate smoothing of the continuous variables x and y. We note that if the bivariate smooth of a continuous variable x and a *discrete* variable y is desired, a separate univariate smooth function of x is estimated for each distinct value of y using the constrained basis discussed above.

There are two main ways to perform bivariate smoothing of continuous variables in `mgcv`, namely isotropic smoothing and tensor product smoothing: we use the latter approach as it naturally extends the univariate methods discussed thus far and also removes the need for scaling (Wood, 2017). To construct the tensor product basis, the univariate (or marginal) bases are first reparameterized for a second time so that the coefficients are the values of the smooth at the $(k-1)$-vector $\mathbf{x}^*$ of evenly spaced covariate values. This improves the interpretation of the smoothing penalties (Wood, 2017). Omitting the details, we let $\mathbf{C}$ be the $N \times (k-1)$ matrix of reparameterized B-spline basis functions with $\boldsymbol{\delta}$ the corresponding $(k-1)$-vector of coefficients. We represent $\mathbf{C}$ graphically in the bottom panel of Figure 14.

Letting $\mathbf{C}^*$ be the $(k-1) \times (k-1)$ matrix of the reparameterized basis functions evaluated at $\mathbf{x}^*$, it is clear from Figure 14 that $\mathbf{C}^* = \mathbf{I}_{k-1}$, i.e., the jth basis function is zero at all values of $\mathbf{x}^*$ except x_j^*, where it is 1. Consequently, $f(\mathbf{x}^*) = \mathbf{C}^*\boldsymbol{\delta} = \mathbf{I}_{k-1}\boldsymbol{\delta} = \boldsymbol{\delta}$, i.e., the coefficients of the smooth are the values it takes at $\mathbf{x}^*$, as desired.

Let $\mathbf{C}^x$ and $\mathbf{C}^y$ be the $N \times (k-1)$ matrices of reparameterized basis functions for x and y. To obtain the basis, we multiply each of the marginal basis functions for x with all of those for y, resulting in a basis dimension of $(k-1)^2$. We let $\mathbf{D}$ be an $N \times (k-1)^2$ matrix of tensor product basis functions. Mathematically, we have:

$$\mathbf{D} = \mathbf{C}^x \odot \mathbf{C}^y = \begin{bmatrix} c_{11}^x \mathbf{c}_1^y & c_{12}^x \mathbf{c}_1^y & \cdots & c_{1(k-1)}^x \mathbf{c}_1^y \\ c_{21}^x \mathbf{c}_2^y & c_{22}^x \mathbf{c}_2^y & \cdots & c_{2(k-1)}^x \mathbf{c}_2^y \\ \vdots & \vdots & \ddots & \vdots \\ c_{N1}^x \mathbf{c}_N^y & c_{N2}^x \mathbf{c}_N^y & \cdots & c_{N(k-1)}^x \mathbf{c}_N^y \end{bmatrix},$$

where $\odot$ represents the row-wise Kronecker (tensor) product and $\mathbf{c}_i^y$ denotes the ith row of $\mathbf{C}^y$. Each of the $(k-1)^2$ basis functions in $\mathbf{D}$ is now a 2D surface, with the tensor product smooth term formed from a linear combination of these surfaces in the same way as for the univariate smooths. Letting the $(k-1)^2$-vector of coefficients be denoted by $\boldsymbol{\zeta}$, we define $f(\mathbf{x}, \mathbf{y}) = \mathbf{D}\boldsymbol{\zeta}$ to be the N-vector of values obtained when the tensor product smooth is evaluated at each of the pairs $(x_i, y_i), i = 1, \ldots, N$.

In Figures 15a-15c we present the tensor product bases constructed from the unconstrained, constrained and reparameterized marginal bases respectively. Although only the latter (Figure 15c) is actually used in the model fitting process, it is interesting to observe the increasing complexity of the surfaces caused by the successive transformations being applied to the marginal bases. However, we note that the effect in terms of its smoothness is as it would be if the basis in Figure 15a were used - the transformations depicted in Figures 15b and 15c are purely carried out for the purposes of identifiability and interpretability. We motivate the use of penalties and describe their construction in Section C.2.

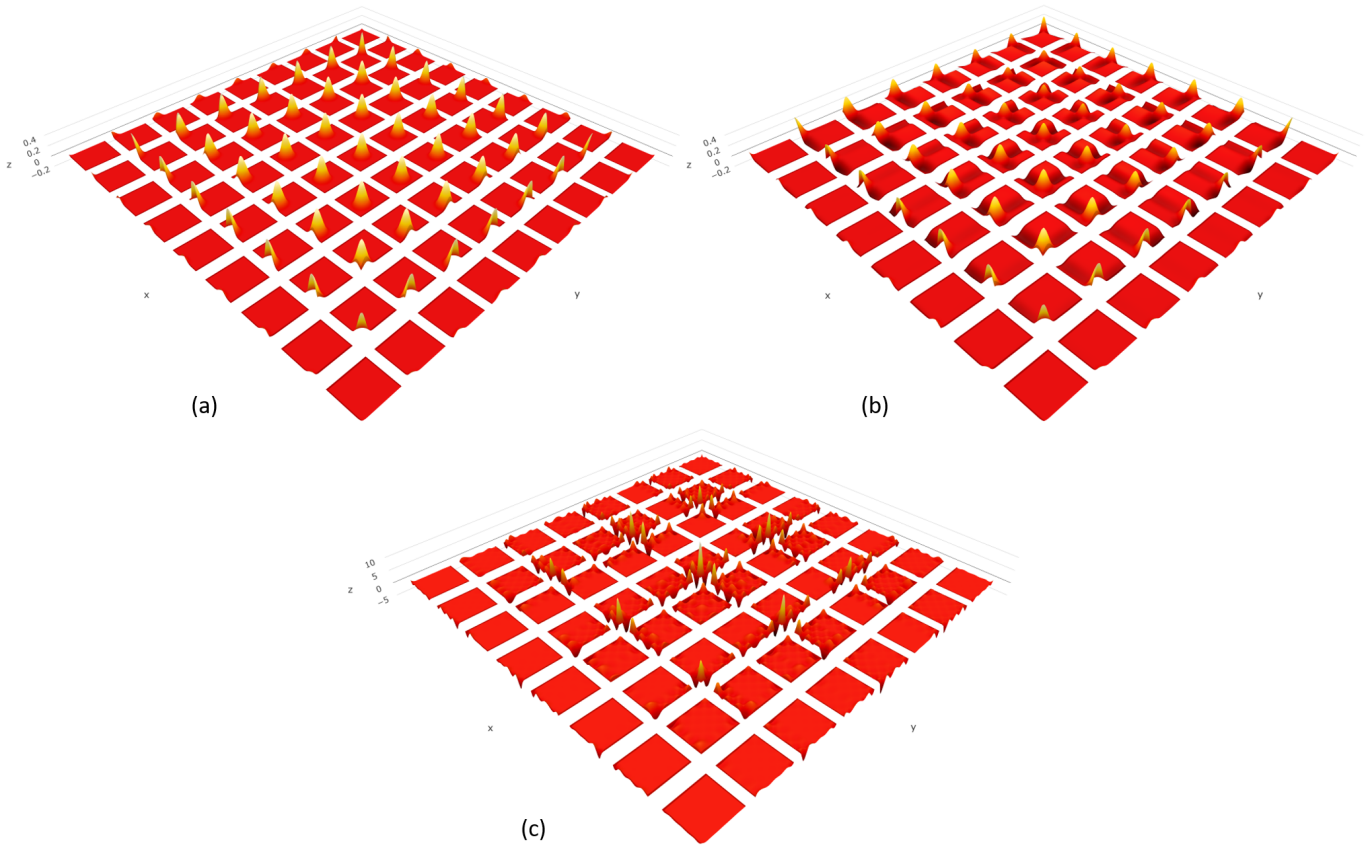

FIGURE 15 Tensor product bases for the smooth interaction between the covariates x = age and y = cohort constructed from the (a) unconstrained, (b) constrained and (c) reparameterized marginal basis functions.

C.2 | Penalties

Having discussed the forms of the univariate and bivariate smooth functions, the second aspect to consider is the way in which their smoothness can be controlled via penalties. To put this into context we briefly introduce the concept underlying the model fitting process, which we will give further details on in Section C.3. There are two opposing aims to balance: goodness of fit to the data and smoothness; putting too much weight on the former can lead to overfitting (and undersmoothing), whereas too much weight on the latter can lead to underfitting (and oversmoothing). Letting $\boldsymbol{\beta}$ and $\ell(\boldsymbol{\beta})$ be the vector of model parameters and model log-likelihood respectively, we estimate the model by maximising the *penalized* log-likelihood $\ell_p(\boldsymbol{\beta})$, defined below:

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{n_p} \lambda_j P_j(\boldsymbol{\beta}). \tag{12}$$

Here, n_p is the total number of penalties (each univariate smooth function has one penalty while a 2D bivariate smooth requires two, one for each dimension), $\lambda_j > 0$ is the smoothing parameter for the jth penalty, and $P_j(\boldsymbol{\beta})$ is the jth penalty term which depends on $\boldsymbol{\beta}$. The penalty terms measure the roughness of the function, with smaller values indicating greater smoothness. Each λ_j controls the aforementioned trade-off by weighting the contribution of its corresponding penalty term, $P_j(\boldsymbol{\beta})$, to the sum of the penalties. In order to maximize $\ell_p(\boldsymbol{\beta})$, it is clear that the larger the value of λ_j, the more we will favour values of $\boldsymbol{\beta}$ that give rise to smaller values of $P_j(\boldsymbol{\beta})$. We will now explain how the penalty terms are constructed, again for univariate and bivariate smooth terms as in Section C.1.

Returning to our unconstrained B-spline basis (see Figure 14), we note that the closer the values of the adjacent basis coefficients, the smoother the resulting function; indeed, identical coefficients give rise to a constant function. Therefore it is intuitive to penalize the (squared) differences between consecutive coefficients, i.e., apply difference penalties; combining B-splines with such penalties was termed 'P-splines' by Eilers and Marx (1996). We use a first-order difference penalty, with penalty term

$$P(\boldsymbol{\alpha}) = \sum_{i=1}^{k-1} (\alpha_{i+1} - \alpha_i)^2.$$

It is clear that $P(\boldsymbol{\alpha}) = 0 \iff \alpha_1 = \cdots = \alpha_k$, i.e., only a constant function invokes a zero penalty. We can write $P(\boldsymbol{\alpha}) = (\mathbf{P}\boldsymbol{\alpha})'(\mathbf{P}\boldsymbol{\alpha}) = \boldsymbol{\alpha}'\mathbf{P}'\mathbf{P}\boldsymbol{\alpha} = \boldsymbol{\alpha}'(\mathbf{P}'\mathbf{P})\boldsymbol{\alpha} = \boldsymbol{\alpha}'\mathbf{S}_\alpha\boldsymbol{\alpha}$, where

$$\mathbf{P} = \begin{bmatrix} -1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 1 \end{bmatrix} \text{ so that } \mathbf{P}\boldsymbol{\alpha} = \begin{bmatrix} \alpha_2 - \alpha_1 \\ \vdots \\ \alpha_k - \alpha_{k-1} \end{bmatrix}, \text{ as required.}$$

We define $\mathbf{S}_\alpha = \mathbf{P}'\mathbf{P}$ to be the $k \times k$ penalty matrix. In `mgcv`, $\mathbf{S}_\alpha$ is scaled by the quotient of the maximum absolute column sum of $\mathbf{S}_\alpha$ (in this case, 4) and the squared maximum absolute row sum of $\mathbf{A}$ (in this case, 1) - we will call this scaled version $\mathbf{S}_{\alpha_s}$.

To obtain the penalty matrix for our constrained B-spline basis (see Figure 14), we simply express $P(\boldsymbol{\alpha})$ in terms

of the new coefficients γ using the relationship $\alpha = \mathbf{Z}\gamma$:

$$P(\alpha) = \alpha' \mathbf{S}_{\alpha_s} \alpha = (\mathbf{Z}\gamma)' \mathbf{S}_{\alpha_s} \mathbf{Z}\gamma = \gamma' \mathbf{Z}' \mathbf{S}_{\alpha_s} \mathbf{Z}\gamma = \gamma' \mathbf{S}_\gamma \gamma = P(\gamma), \tag{13}$$

where $\mathbf{S}_\gamma = \mathbf{Z}' \mathbf{S}_{\alpha_s} \mathbf{Z}$ is the $(k-1) \times (k-1)$ penalty matrix.

For the bivariate smooth there are two penalty matrices, one for each covariate direction. We first obtain the two $(k-1) \times (k-1)$ reparameterized marginal penalty matrices for the covariates x and y, which we call $\mathbf{S}^x_{\delta_s}$ and $\mathbf{S}^y_{\delta_s}$ respectively. Note that $\mathbf{S}_\delta$ is obtained by expressing $P(\gamma)$ in terms of the new coefficients δ in a similar way to equation (13), and is then scaled by its largest eigenvalue to obtain $\mathbf{S}_{\delta_s}$. We then define the two $(k-1)^2 \times (k-1)^2$ tensor product penalty matrices as $\mathbf{S}^x_\zeta = \mathbf{S}^x_{\delta_s} \otimes \mathbf{I}_{k-1}$ and $\mathbf{S}^y_\zeta = \mathbf{I}_{k-1} \otimes \mathbf{S}^y_{\delta_s}$, where $\otimes$ denotes the Kronecker product. We rescale these similarly to $\mathbf{S}_\alpha$, except using $\mathbf{D}$ in the place of $\mathbf{A}$. We call the resulting matrices $\mathbf{S}^x_{\zeta_s}$ and $\mathbf{S}^y_{\zeta_s}$.

Having now determined our matrix of basis functions $\mathbf{B}$ and our penalty matrix $\mathbf{S}_\gamma$ for the univariate case, and our matrix of basis functions $\mathbf{D}$ and our penalty matrices $\mathbf{S}^x_{\zeta_s}$ and $\mathbf{S}^y_{\zeta_s}$ for the bivariate case, we discuss the model fitting process in Section C.3.

C.3 | Model fitting

We begin by letting $\mathbf{B}^x$ and $\mathbf{B}^y$ be the $N \times (k-1)$ matrices of constrained basis functions for x and y, and $\mathbf{S}^x_\gamma$ and $\mathbf{S}^y_\gamma$ be the corresponding penalty matrices. Before we fit the model, we test for linear dependence of the tensor product basis functions on the intercept and marginal basis functions for x and y, i.e., of $\mathbf{X}_2 = \mathbf{D}$ on $\mathbf{X}_1 = [1 : \mathbf{B}^x : \mathbf{B}^y]$. We follow the method described in Appendix D and implemented by the `fixDependence` function in `mgcv`. If any dependence is present, we remove the identified columns from $\mathbf{D}$, and the columns and corresponding rows from $\mathbf{S}^x_{\zeta_s}$ and $\mathbf{S}^y_{\zeta_s}$ - we denote these reduced matrices by $\mathbf{D}^r$, $\mathbf{S}^x_{\zeta^r}$ and $\mathbf{S}^y_{\zeta^r}$ respectively. Therefore the final model matrix in our example is $\mathbf{X} = [1 : \mathbf{B}^x : \mathbf{B}^y : \mathbf{D}^r]$ and the final penalty matrices are $\mathbf{S}^x_\gamma$, $\mathbf{S}^y_\gamma$, $\mathbf{S}^x_{\zeta^r}$ and $\mathbf{S}^y_{\zeta^r}$ - we will call these $\mathbf{S}_1, \ldots, \mathbf{S}_4$ for convenience.

Continuing from Section C.2, we let β be the parameter vector containing the intercept and the coefficients of the marginal and tensor product bases in $\mathbf{X}$. Specifying our binomial logistic GAM explicitly, we let $Y_i \overset{\text{ind}}{\sim} \text{Binomial}(n_i, r_i)$, $i = 1, \ldots, N$, where n_i is the number of women in the ith age-cohort group and r_i is the probability of a birth in that group. Then our GAM sets $\text{logit}(\mathbf{r}) = \mathbf{X}\beta$. We also set $n_p = 4$ and $P_j(\beta) = \beta' \mathbf{S}_{j'} \beta$, $j = 1, \ldots, 4$, where $\mathbf{S}_{j'}$ is $\mathbf{S}_j$ appropriately embedded as a diagonal block in an otherwise zero matrix, so that we obtain the correct penalty term. Our penalized likelihood $\ell_p(\beta)$, specified in equation (12) in Section C.2, then becomes:

$$\ell_p(\beta) = \ell(\beta) - \frac{1}{2} \sum_{j=1}^{4} \lambda_j \beta' \mathbf{S}_{j'} \beta. \tag{14}$$

For a given vector of smoothing parameters λ, the maximization problem can be solved (i.e., β can be estimated) by the method of penalized iteratively re-weighted least squares (PIRLS) - see Wood (2017) for details. In the case of logistic GAMs, which have a known scale parameter, λ is chosen to minimize the unbiased risk estimator (UBRE)

criterion (Craven and Wahba, 1979) by default:

$$\text{UBRE} = \frac{D}{N} + \frac{2\hat{\rho}}{N} - 1,$$

where D is the model deviance and $\hat{\rho}$ is the effective number of parameters. This latter quantity is defined as the trace of the projection or hat matrix, as in linear regression. It is also possible to disaggregate the total by the smooth terms and even the parameters themselves. Returning to the UBRE, we note that it is effectively just the Akaike Information Criterion, i.e., the AIC (= $2\hat{\rho} - 2\ell_m$ where ℓ_m is the log-likelihood of the fitted model) rescaled:

$$\text{UBRE} = \frac{\text{AIC}}{N} + \frac{2\ell_s}{N} - 1,$$

where ℓ_s is the log-likelihood of the saturated model which fits the data perfectly.

For various trial $\boldsymbol{\lambda}$'s, the UBRE is evaluated upon convergence of the PIRLS scheme - the $\boldsymbol{\lambda}$ that gives the smallest UBRE is then chosen, along with its corresponding estimate of $\boldsymbol{\beta}$. Alternative methods to estimate $\boldsymbol{\lambda}$ are available in `gam`, including a Laplace approximation to restricted maximum likelihood (REML) estimation and minimization of the generalized cross validation (GCV) criterion.

C.4 | Formulating a Bayesian GAM

An overview of the Bayesian formulation is given in Section 4.5 of the paper, however for completeness we note some additional details. To connect the description in Section 4.5 of the paper to this appendix, we let $\tau_j = 1/\lambda_j, j = 1, \ldots, 4$ be the inverses of the original smoothing parameters. The total smoothing penalty in equation (14) then becomes:

$$\sum_{j=1}^{4} \boldsymbol{\beta}' \frac{\mathbf{S}_{j'}}{\tau_j} \boldsymbol{\beta} = \boldsymbol{\beta}' \left(\sum_{j=1}^{4} \frac{\mathbf{S}_{j'}}{\tau_j} \right) \boldsymbol{\beta} = \boldsymbol{\beta}' S(\boldsymbol{\tau}) \boldsymbol{\beta},$$

where $S(\boldsymbol{\tau}) = \sum_{j=1}^{4} \frac{\mathbf{S}_{j'}}{\tau_j}$ is the combined penalty matrix as a function of the new smoothing parameters $\boldsymbol{\tau}$. Following Umlauf et al. (2018), we specify the below prior for $\boldsymbol{\beta}$, conditional on $\boldsymbol{\tau}$:

$$f(\boldsymbol{\beta}|\boldsymbol{\tau}) \propto |S(\boldsymbol{\tau})|^{\frac{1}{2}} \exp\left(-\frac{1}{2} \boldsymbol{\beta}' S(\boldsymbol{\tau}) \boldsymbol{\beta} \right). \tag{15}$$

Note that equation (15) is equivalent to combining equations (9) and (10) of the paper, which are written for general 1D and 2D smooth functions, for all smooth functions included in the particular model being fitted. The prior for $\boldsymbol{\beta}$ is based on the multivariate normal distribution, and its spread is determined by the smoothing parameters $\boldsymbol{\tau}$. The proportionality expresses our prior belief that smoother functions are more likely, as this gives higher prior probability density to values of $\boldsymbol{\beta}$ that give rise to smaller values of the total penalty $\boldsymbol{\beta}' S(\boldsymbol{\tau}) \boldsymbol{\beta}$ for fixed $\boldsymbol{\tau}$. However, the strength of this prior belief depends heavily on the value of $\boldsymbol{\tau}$: small values of $\boldsymbol{\tau}$ impose substantial smoothing whereas large values of $\boldsymbol{\tau}$ have negligible effect (note that this is the opposite interpretation to $\boldsymbol{\lambda}$ as a result of the inversion). Therefore the prior on $\boldsymbol{\tau}$ is critical. The weakly informative prior that we specify in the paper is not so vague that it provides no information, but not so precise that we are unable to learn from the data. In this way we give the data sufficient freedom to determine the best compromise between goodness of fit and smoothness, as was the case in the classical approach described in Section C.3.

D | NUMERICAL IDENTIFICATION OF DEPENDENCE

Let $\mathbf{X}_2$ be the model matrix for the tensor product smooth term, and $\mathbf{X}_1$ be the combined model matrix for the intercept and marginal smooth terms for the covariates involved in the tensor product smooth. We want to test for linear dependence of $\mathbf{X}_2$ on $\mathbf{X}_1$. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are separately of full column rank, we can do this by obtaining the QR decomposition of $[\mathbf{X}_1 : \mathbf{X}_2] = \mathbf{Q}^* \begin{bmatrix} \mathbf{R}^* \\ 0 \end{bmatrix}$ (see Section C.1). Then if $\mathbf{X}_2$ depends on $\mathbf{X}_1$ ($[\mathbf{X}_1 : \mathbf{X}_2]$ has reduced column rank), the upper triangular matrix $\mathbf{R}^*$ has reduced rank with the order of rank deficiency given by the size of the zero block at the lower right corner of $\mathbf{R}^*$. The use of column pivoting when computing the QR decomposition of a rank deficient matrix is preferable - to avoid the selection of any of the r columns of $\mathbf{X}_1$ for removal through the pivoting process, the QR decomposition of $[\mathbf{X}_1 : \mathbf{X}_2]$ is performed in two steps:

1. Obtain the QR decomposition of $\mathbf{X}_1 = \mathbf{Q}_1 \begin{bmatrix} \mathbf{R}_1 \\ 0 \end{bmatrix}$.
2. Let $\mathbf{Q}_1' \mathbf{X}_2 = \begin{bmatrix} \bar{\mathbf{G}} \\ \mathbf{G} \end{bmatrix}$, where $\bar{\mathbf{G}}$ contains the first r rows of $\mathbf{Q}_1' \mathbf{X}_2$ and $\mathbf{G}$ contains the rest.

   Then obtain the QR decomposition of $\mathbf{G} = \mathbf{Q}_2 \begin{bmatrix} \mathbf{R}_2 \\ 0 \end{bmatrix}$ with pivoting.

Noting that as $\mathbf{Q}_1$ is an orthogonal matrix, $\mathbf{X}_2 = \mathbf{Q}_1 \mathbf{Q}_1' \mathbf{X}_2 = \mathbf{Q}_1 \begin{bmatrix} \bar{\mathbf{G}} \\ \mathbf{G} \end{bmatrix}$, we can express the QR decomposition of $[\mathbf{X}_1 : \mathbf{X}_2]$ in the following way:

$$
\begin{aligned}
[\mathbf{X}_1 : \mathbf{X}_2] &= \left[\mathbf{Q}_1 \begin{bmatrix} \mathbf{R}_1 \\ 0 \end{bmatrix} : \mathbf{Q}_1 \begin{bmatrix} \bar{\mathbf{G}} \\ \mathbf{G} \end{bmatrix} \right] = \mathbf{Q}_1 \begin{bmatrix} \mathbf{R}_1 & \bar{\mathbf{G}} \\ 0 & \mathbf{Q}_2 \begin{bmatrix} \mathbf{R}_2 \\ 0 \end{bmatrix} \end{bmatrix} = \mathbf{Q}_1 \begin{bmatrix} \mathbf{I}_r & 0 \\ 0 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1 & \bar{\mathbf{G}} \\ 0 & \mathbf{R}_2 \\ 0 & 0 \end{bmatrix} \\
&= \mathbf{Q}^* \begin{bmatrix} \mathbf{R}^* \\ 0 \end{bmatrix},
\end{aligned}
$$

where $\mathbf{Q}^* = \mathbf{Q}_1 \begin{bmatrix} \mathbf{I}_r & 0 \\ 0 & \mathbf{Q}_2 \end{bmatrix}$ and $\mathbf{R}^* = \begin{bmatrix} \mathbf{R}_1 & \bar{\mathbf{G}} \\ 0 & \mathbf{R}_2 \end{bmatrix}$.

In this way, if dependence is present then then will be a zero block at the lower right corner of $\mathbf{R}_2$. The dependent columns are then easily identifiable as the $\mathbf{X}_2$ columns that were pivoted to these final columns in Step 2, and can subsequently be removed to ensure identifiability.

E | MODELLING QUALIFICATION GIVEN AGE AND COHORT: FURTHER DETAILS

Following on from Section 4.2 of the paper, in this appendix we provide further details on the modelling of qualification (Q) given age (A) and cohort (C), which for convenience we will refer to as modelling $Q|A, C$.

E.1 | Model specification

We begin by specifying the form of our multinomial logistic regression model in a similar way to the imputation model in Section B.2. For a given parity, age a, cohort c, and qualification variable Q with n_q categories (in ascending order), let $\mathbf{Y}_{a,c} = (Y^1_{a,c}, ..., Y^{n_q}_{a,c})$ be the vector of counts and $n^{obs}_{a,c}$ be the total observed number of records. Then we let $\mathbf{Y}_{a,c} \sim \text{Multinomial}(n^{obs}_{a,c}, \pi^1_{a,c}, \ldots, \pi^{n_q}_{a,c})$, where $\pi^j_{a,c}$ is the probability of belonging to the jth Q category, $j = 1, \ldots, n_q$, and $\sum_{j=1}^{n_q} \pi^j_{a,c} = 1$. With Q category $j = 1$ as the reference category, we specify the model through the equations:

$$\log\left(\frac{\pi^j_{a,c}}{\pi^1_{a,c}}\right) = \eta^j_{a,c},\ j = 2, \ldots, n_q, \tag{16}$$

which are fitted simultaneously. We experiment with various forms for the linear predictors $\eta^j_{a,c}$, which we specify in Table 4.

TABLE 4 Specifications of the initial $Q|A, C$ models; $\eta^j_{a,c}$ is the linear predictor from equation (16); n_q is the number of qualification categories; $\tilde{x}$ indicates that the variable x is centred around the median of its distinct values.

Model (M)	Specification
1	$\eta^j_{a,c} = \beta^j_0\ \forall j$
2	$\eta^j_{a,c} = \beta^j_0 + \beta^j_1 \tilde{c}\ \forall j$
3	$\eta^j_{a,c} = \beta^j_0 + \beta^j_{2,c}\ \forall j$
4	$\eta^j_{a,c} = \beta^j_0 + \beta^j_{2,c} + \beta^j_3 \tilde{a}\ \forall j$
5	M4 but with $\eta^{n_q}_{a,c} = \beta^{n_q}_0 + \beta^{n_q}_{2,c} + \beta^{n_q}_{4,a}$
6	M5 but with $\eta^{n_q-1}_{a,c} = \beta^{n_q-1}_0 + \beta^{n_q-1}_{2,c} + \beta^{n_q-1}_{4,a}$
7	$\eta^j_{a,c} = \beta^j_0 + \beta^j_{2,c} + \beta^j_{4,a}\ \forall j$

We will refer to the models as M1-M7. In words, M1 just contains an intercept, M2 adds the main effect of C as a linear term while M3 modifies this to be a set of cohort-specific parameters (essentially treating C as a discrete variable rather than continuous); we take the 1945 cohort to be the reference category. M4 adds a linear age effect to M3 while M5 again changes this to be a set of age-specific parameters for $j = n_q$ only (when $n_q \in \{3, 4\}$); M6 uses this more complex representation of A for $j \in \{3, 4\}$ when $n_q = 4$, while M7 allows this for all $j \geq 2$. We take age 15 as the reference category for parity 0, but age 44 for parities 1 and 2. This is because we found that the low exposures at the youngest ages led to great uncertainty in the parameter estimates if the youngest age was taken to be the reference category.

E.2 | Model fitting and comparison

For each parity, we fit these models in `R` using the `Stan` software package (Stan Development Team, 2019). We weight the contribution to the multinomial log-likelihood for each age-cohort combination so that the implied sample size is the weighted number of person-years (denoted $n_{a,c}^{obs_w}$) as opposed to the raw number $n_{a,c}^{obs}$ - this is achieved by using the weight $n_{a,c}^{obs_w}/n_{a,c}^{obs}$. Vague $N(0, 10^2)$ priors are specified for the model parameters. For each model we perform 1,000 warm-up iterations followed by 1,000 retained iterations, and find that the samples exhibit a strong level of convergence.

To compare models we use the Bayesian Information Criterion (BIC), which balances goodness of fit to the data with complexity, i.e., the number of parameters; smaller values are desirable. We perform a rough assessment of model fit using a Pearson chi-squared statistic, X^2:

$$X^2 = \sum_{a,c,j} \frac{(n_{a,c,j}^{obs_w} - n_{a,c}^{obs_w}\hat{\pi}_{a,c}^{j})^2}{n_{a,c}^{obs_w}\hat{\pi}_{a,c}^{j}}, \tag{17}$$

where $n_{a,c,j}^{obs_w}$ is the observed number of person-years with age a, cohort c and qualification level j multiplied by the weight above, and $\hat{\pi}_{a,c}^{j}$ is the corresponding posterior mean probability (this is also used in the BIC computation). For each model we compute this statistic and for calibration present the 95% quantile of the χ_ν^2 distribution ($\chi_{\nu,0.95}^2$), with ν the appropriate degrees of freedom if the model were unpenalized ($\nu = (n_q - 1) \times n_{ac}^{obs} - p$, where n_{ac}^{obs} is the number of observed age-cohort combinations and p is the number of model parameters).

E.3 | Initial results and variant descriptions

We present the results of this initial modelling in Table 5, noting that it is not possible to fit M5 for parity 1, and M6 for parities 1 and 2, as their n_q values are too small. The parity 0 results show M4-M7 (models including age as well as cohort) to have reasonable X^2 values; however the BIC is lowest for M5 and so we take this forward as our initial chosen model. For parity 1 M7 is our chosen model as it has the lowest BIC; its X^2 value is also low. Finally, for parity 2 the situation is identical to parity 0 (except with M6 absent) - M5 has the lower BIC and so is our preferred model.

Next, we build on these chosen models in order to allow us to obtain forecasts for the post-1982 cohorts; we call this the first variant and denote the corresponding models as Mx.1. This first variant approximates the cohort-specific parameters ($\beta_{2,c}^{j}$) with a straight line for $c \geq 1972$, which requires only an intercept and slope parameter for each j as opposed to the 11 independently estimated coefficients $\beta_{2,1972}^{j}, \ldots, \beta_{2,1982}^{j}$ in our initial models. Establishing this linear relationship between the cohort-specific parameters means that we can extrapolate it to future cohorts and thus forecast the age-cohort surface. We choose the 1972 cohort as the starting point for this linear approximation as it marks the start of a sharp decline in the parity 0 '< GCSE' observed proportions (this can be seen in the corresponding fitted probability surface in Figure 2 of the paper); this decrease is as a result of the replacement of O Levels by GCSEs in 1988. In Figure 16 we plot the $\beta_{2,c}^{j}$ parameter estimates from the initial chosen models with the first variant overlaid. The parity 0 plot evidences this 1972 change point - the jump in the $\beta_{2,c}^{j}$ values indicates the proportions increasing in the higher Q categories to counteract the decline in the lowest category. Following this the $\beta_{2,c}^{j}$ values exhibit a linear trend which is also the case for the higher parities, justifying our approach.

TABLE 5 Results of the initial modelling of $Q|A, C$ for parities 0, 1 and 2. M is the model number (see Table 4 for the model specifications); p is the number of parameters in the model; X^2 is the Pearson chi-squared statistic which we compare to $\chi^2_{\nu,0.95}$, the 95% quantile of the χ^2_ν distribution. The row of the chosen model for each parity is in **bold**. Values given to 2 decimal places.

Parity	M	p	BIC	ν	X^2	$\chi^2_{\nu,0.95}$
0	1	3	35297.45	2847	20568.62	2972.25
	2	6	22327.57	2844	7676.09	2969.18
	3	114	20586.94	2736	4681.47	2858.80
	4	117	18152.82	2733	2267.38	2855.73
	5	**145**	**17711.02**	**2705**	**1468.14**	**2827.11**
	6	173	17972.23	2677	1388.68	2798.48
	7	201	18224.61	2649	1306.28	2769.85
1	1	1	7371.36	910	3486.08	981.29
	2	2	6518.14	909	2679.24	980.25
	3	38	6103.00	873	1888.06	942.85
	4	39	5264.35	872	1071.19	941.81
	7	**67**	**5018.32**	**844**	**549.67**	**912.70**
2	1	2	14292.55	1646	7406.08	1741.50
	2	4	11428.35	1644	4716.15	1739.44
	3	76	11137.55	1572	3536.80	1665.35
	4	78	8995.13	1570	1601.53	1663.29
	5	**106**	**8737.65**	**1542**	**1200.83**	**1634.47**
	7	134	8882.70	1514	1018.96	1605.63

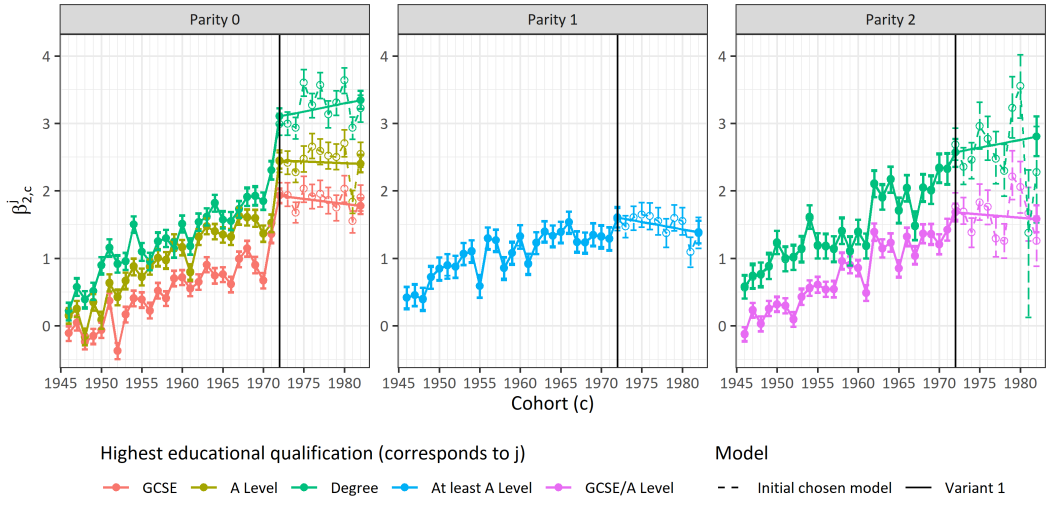

FIGURE 16 Parity-specific plots of the cohort-specific parameter ($\beta^j_{2,c}$) estimates for the initial chosen models from Table 5 and the first variant. Vertical lines indicate the 1972 change point; error bars indicate 95% CIs.

The second variant (models denoted Mx.2) modifies the first variant by additionally smoothing the age-specific parameters ($\beta^j_{4,a}$) via a first-order random walk with standard deviation parameter $\sigma_A \sim N(0, 0.01^2)$; we contrast the $\beta^j_{4,a}$ estimates for the two variants in Figure 17. This figure clearly illustrates the considerable uncertainty associated with the estimates for small a in the parity 1 and 2 plots, caused by small exposures at these ages - as such, we deemed it necessary to borrow strength across age in order to obtain more plausible CIs, which we have evidently achieved. In terms of the patterns exhibited by the parameter estimates in Figures 16 and 17, we can interpret the generally positive trends in the former as representing an increase in the likelihood of belonging to a higher qualification category as the cohorts become younger. In the case of the latter, the gradual increases with age are caused by women in lower qualification categories being more likely to have a child and so remove themselves from their current risk set at younger ages; this means that the women remaining in the said risk set will tend to be from higher qualification categories as age increases. However, the stabilising behaviour seen in the latter half of the age range across the parities is caused by these women *themselves* becoming more likely to have children compared to those in the lower categories due to postponement.

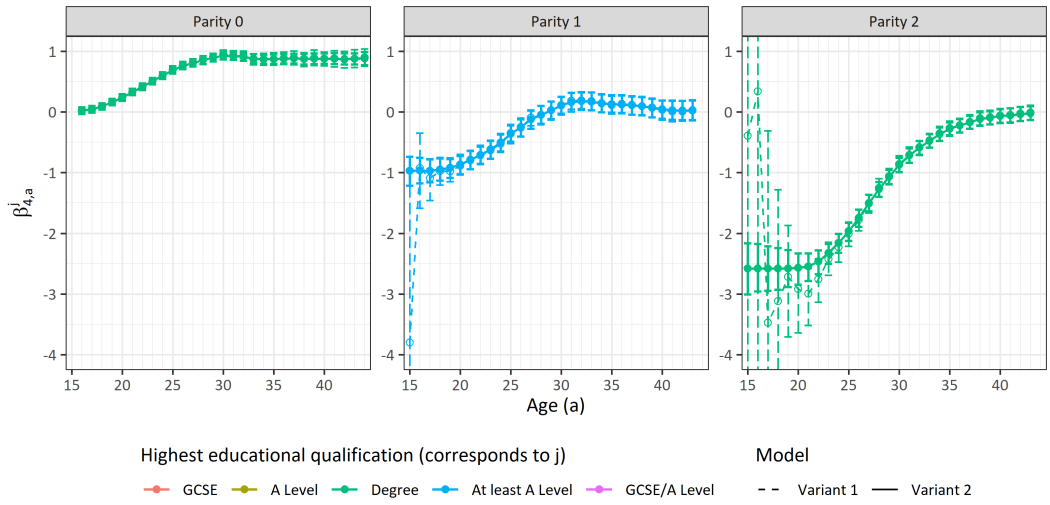

FIGURE 17 Parity-specific plots of the age-specific parameter ($\beta^j_{4,a}$) estimates for the first and second variants; error bars indicate 95% CIs.

E.4 | Variant results

In Table 6 we present the results of our initial chosen models (copied from Table 5 for convenience) and the first and second variants applied to these models. We note that the fit of the first variant is less close, which is expected as it greatly simplifies each of the initial chosen models; however, the X^2 values are still reasonable. The BIC increases for parity 0 but *decreases* for parities 1 and 2, due to the gains in parsimony outweighing the losses in fit in the case of the latter but not the former.

The effect of the second variant is slightly harder to assess because although technically we have added one standard deviation parameter (σ_A) in each case, in reality we have again simplified our current model (the first variant)

TABLE 6 Results of the additional modelling of $Q|A, C$ for parities 0, 1 and 2. M is the model number (see Table 4 for the model specifications; $x.y$ indicates Model x with variant y applied); p is the number of parameters in the model; X^2 is the Pearson chi-squared statistic which we compare to $\chi^2_{\nu,0.95}$, the 95% quantile of the χ^2_ν distribution. Values given to 2 decimal places.

Parity	M	p	BIC	ν	X^2	$\chi^2_{\nu,0.95}$
0	5	145	17711.02	2705	1468.14	2827.11
	5.1	118	17815.56	2732	1898.26	2854.71
	5.2	119	17828.83	2731	1898.60	2853.69
1	7	67	5018.32	844	549.67	912.70
	7.1	58	4947.59	853	575.87	922.06
	7.2	59	4964.25	852	574.10	921.02
2	5	106	8737.65	1542	1200.83	1634.47
	5.1	88	8615.57	1560	1292.66	1653.00
	5.2	89	8635.97	1559	1257.73	1651.97

by borrowing strength across age and so reducing the *effective* number of age-specific parameters. As there is no easy way to ascertain this quantity, we set p as its upper bound and hence work with a worst-case scenario; this should be taken into account when interpreting the BIC. The X^2 value has increased marginally for parity 0 while *decreasing* more substantially for parities 1 and 2. This is likely due to the greater impact the smoothing has had for these parities in terms of decreasing the uncertainty associated with age-specific parameters for which the data provided very little information, a problem that parity 0 did not suffer from (see Figure 17). Noting that the X^2 values are still low for the second variants, we take them forward as our final chosen models for each parity. We present the corresponding Lexis surfaces of fitted probabilities for each parity in Figure 2 of the paper.

F | MODELLING TIME SINCE LAST BIRTH GIVEN AGE AND QUALIFICATION: FURTHER DETAILS

Following on from Section 4.3 of the paper and in a similar way to Appendix E, in this appendix we provide further details on the modelling of time since last birth (T) given age (A) and qualification (Q), which for convenience we will refer to as modelling $T|A,(Q)$.

F.1 | Parity 3+ model specification and initial results

We begin with parity 3+ as there is no dependence on Q and so the specification is more straightforward - hence we are modelling $T|A$ for now. As discussed in Section 4.3 of the paper, the number of T categories at a given age is not constant. Consequently, we specify a separate multinomial model for each distinct maximum observable T value from $T = 3$ to $T = 11$, i.e., $A = 15$ to $A \geq 23$. However, we share T-specific parameters across A by fitting these models simultaneously to the T counts.

We now specify the form of our multinomial logistic regression model in a similar way to Section E.1. For a given parity, age a and maximum T value $t_a \in \{3, \ldots, 11\}$, let $\mathbf{Y}_a = (Y_a^1, ..., Y_a^{t_a})$ be the vector of T counts and n_a^{obs} be the total observed number of person-years. Then we let $\mathbf{Y}_a \sim \text{Multinomial}(n_a^{obs}, \pi_a^1, \ldots, \pi_a^{t_a})$, where π_a^j is the probability of belonging to the jth T category, $j = 1, \ldots, t_a$, and $\sum_{j=1}^{t_a} \pi_a^j = 1$. With T category $j = 1$ as the reference category, we specify the model through the equations:

$$\log\left(\frac{\pi_a^j}{\pi_a^1}\right) = \eta_a^j,\ j = 2, \ldots, t_a, \tag{18}$$

which are fitted simultaneously. We experiment with various forms for the linear predictors η_a^j, which we specify in Table 7.

TABLE 7 Specifications of the initial $T|A$ models; η_a^j is the linear predictor from equation (18); $\tilde{x}$ indicates that the variable x is centred around the median of its distinct values; $a_b = a - t$, i.e., the age at the last birth event.

Model (M)	Specification
1a	$\eta_a^j = \beta_0^j\ \forall j$
2a	$\eta_a^j = \beta_1^j \tilde{a}\ \forall j$
3a	$\eta_a^j = \beta_0^j + \beta_1^j \tilde{a}\ \forall j$
4a	$\eta_a^j = \beta_0^j + \beta_1^j \tilde{a} + \beta_2^j \tilde{a}^2\ \forall j$
5a	$\eta_a^j = \beta_1^j \tilde{a} + \beta_3^{a_b}\ \forall j$
6a	$\eta_a^j = \beta_1^j \tilde{a} + \beta_2^j \tilde{a}^2 + \beta_3^{a_b}\ \forall j$

We will refer to the models as M1a-M6a for reasons that will become clear when we discuss parities 1 and 2. In words, M1a contains just an intercept, M2a contains just the main effect of A as a linear term while M3a combines M1a and M2a; M4a additionally includes a quadratic effect of age. The final two models introduce the main effect of

A_b (the age at the last birth event) as a set of A_b-specific parameters for which a reference category is not necessary - M5a adds the linear age term while M6a additionally adds the quadratic age term. We fit these models as in Section E.2, adjusting the weighting process to be for each age rather than age-cohort combination. We also compare models and informally check their fit as before, again adjusting equation (17) to remove dependence on C; the degrees of freedom ν is simply the number of observed age-time-since-last-birth combinations with $t > 1$, minus p.

We present the results of the initial modelling of $T|A$ in Table 8. We see that only M5a and M6a have reasonable X^2 values, illustrating the importance of accounting for A_b - this is not surprising given our earlier discussion of the presence of the 'parallelograms' in the plots of the fitted probabilities in Figure 3 of the paper. M5a has the lowest BIC, so we will take this forward as our initial chosen model for now.

TABLE 8 Results of the initial modelling of $T|A$ for parity 3+. M is the model number (see Table 7 for the model specifications); p is the number of parameters in the model; X^2 is the Pearson chi-squared statistic which we compare to $\chi^2_{\nu,0.95}$, the 95% quantile of the χ^2_ν distribution. The row of the chosen model is in **bold**. Values given to 2 decimal places.

M	p	BIC	ν	X^2	$\chi^2_{\nu,0.95}$
1a	10	22067.78	245	19212.18	282.51
2a	10	12137.34	245	9605.52	282.51
3a	20	2513.52	235	900.28	271.76
4a	30	2041.95	225	423.29	260.99
5a	**41**	**1798.42**	**214**	**107.54**	**249.13**
6a	51	1869.66	204	72.79	238.32

F.2 | Parity 1 and 2 model specification and initial results

Next we will consider parities 1 and 2, which are more involved due to the presence of Q. It is straightforward to adjust our parity 3+ setup in Section F.1 to account for this by simply indexing our model components by A and Q rather than just A, e.g., $\pi^j_{a,q}$, $\eta^j_{a,q}$, etc. In terms of the models, we will focus on M5a and M6a only due to their superior performance for parity 3+. For a given model, we have more options as we can allow each model term to be either independent of Q and therefore shared across all Q categories, or to be Q-specific. Letting the equivalents of M5a and M6a for parity 1/2 be the cases where all terms are shared, we specify an additional 3 variants of M5a (M5b-M5d) and 7 variants of M6a (M6b-M6h). We illustrate these in Table 9, where ✗ and ✓ indicate that the relevant parameter is shared and Q-specific respectively.

TABLE 9 Illustration of the $T|A, Q$ models fitted for parities 1 and 2. M is the model number (see Table 7 for the specifications of M5a and M6a); β^j_1 is the coefficient of age, β^j_2 is the coefficient of the square of age, and $\beta^{a_b}_3$ is the age at last birth parameter; ✗ indicates that the parameter is shared across Q categories, while ✓ indicates it is Q-specific.

Parameter/M	5a	5b	5c	5d	6a	6b	6c	6d	6e	6f	6g	6h
β^j_1	✗	✓	✗	✓	✗	✓	✗	✗	✓	✓	✗	✓
β^j_2	-	-	-	-	✗	✗	✓	✗	✓	✗	✓	✓
$\beta^{a_b}_3$	✗	✗	✓	✓	✗	✗	✗	✓	✗	✓	✓	✓

Fitting these models similarly to those for parity 3+, and following slight adjustments to the BIC and X^2 definitions in order to incorporate the multiple AT surfaces that we have for parities 1 and 2, we present the results of fitting M5a-M6h in Table 10. First considering parity 1, we observe that all of the X^2 values are relatively large, with those for M6f and M6h the most reasonable. As M6f has the lowest BIC of all the models, we take this forward as our initial chosen model for parity 1. In contrast, for parity 2 there are four models with low X^2 values, namely M5d and M6f-M6h. Again it is one of these models, M5d, that also has the smallest BIC and so is our chosen model for parity 2. It is interesting that our chosen models, M5d and M6f, both have Q-specific coefficients of age ($\beta^j_{1,q}$) and age at last birth parameters ($\beta^{a_b}_{3,q}$) (with M6f further including shared coefficients of the square of age (β^j_2)).

TABLE 10 Results of the initial modelling of $T|A, Q$ for parities 1 and 2. M is the model number (see Tables 7 and 9 for the model specifications); p is the number of model parameters; X^2 is the Pearson chi-squared statistic which we compare to $\chi^2_{\nu,0.95}$, the 95% quantile of the χ^2_ν distribution. The row of the chosen model for each parity is in **bold**. Values given to 2 decimal places.

	Parity 1					Parity 2				
M	p	**BIC**	ν	X^2	$\chi^2_{\nu,0.95}$	p	**BIC**	ν	X^2	$\chi^2_{\nu,0.95}$
5a	41	5313.95	487	2428.22	539.45	41	10230.79	736	5947.72	800.22
5b	51	5011.52	477	1964.88	528.92	61	7427.15	716	2963.98	779.36
5c	72	4949.41	456	1697.30	506.78	103	6271.01	674	1484.56	735.51
5d	82	4735.66	446	1316.86	496.24	**123**	**5505.64**	**654**	**455.86**	**714.60**
6a	51	4688.47	477	1541.30	528.92	51	10257.42	726	5855.46	789.79
6b	61	4387.99	467	1135.83	518.38	71	7464.33	706	2905.95	768.92
6c	61	4450.56	467	1209.22	518.38	71	8288.26	706	3725.20	768.92
6d	82	4318.89	446	822.70	496.24	113	6316.81	664	1418.20	725.06
6e	71	4448.02	457	1088.81	507.84	91	6858.87	686	2124.43	748.04
6f	**92**	**4151.87**	**436**	**545.41**	**485.68**	133	5516.59	644	355.59	704.15
6g	92	4294.99	436	724.59	485.68	133	5785.53	644	640.01	704.15
6h	102	4243.73	426	513.31	475.12	153	5711.19	624	320.42	683.22

F.3 | Variant descriptions and results

As in Appendix E, we present two variants of these chosen models. The first is motivated by the Pearson residual plots for our initial parity 1 chosen model (not shown). They indicate that a partial cause of the large X^2 value is the presence of large positive residuals at $T = 11$. To mitigate this, we add an intercept parameter at this t value (β^{11}_0) shared across Q. We denote M6f fitted with this first variant by M6f.1 and present the results in Table 11. This extra parameter reduces the value of X^2 to something much more reasonable, and also improves the BIC considerably.

The second variant (denoted Mx.2) is similar to that used in Section E.3 in that we smooth the Q-specific/shared age at last birth parameters ($\beta^{a_b}_{3,(q)}$) via a first-order random walk, with standard deviation parameter $\sigma_{A_b} \sim N(0, 0.01^2)$ shared across the qualification levels q in the case of parities 1 and 2. The reason for this is again to borrow strength and reduce the uncertainty associated with the parameter estimates corresponding to very young ages at the last birth event. We fit this second variant to the first variant for parity 1 and the initial chosen models for parities 2 and 3+. Its impact is illustrated in Figure 18, where we see that the smoothing has the desired effect across the parities.

TABLE 11 Results of the additional modelling of $T|A, C, (Q)$ for parities 1, 2 and 3+. M is the model number (see Tables 7 and 9 for the model specifications; $x.y$ indicates Model x with variant y applied); p is the number of parameters in the model; X^2 is the Pearson chi-squared statistic which we compare to $\chi^2_{\nu,0.95}$, the 95% quantile of the χ^2_ν distribution. Values given to 2 decimal places.

Parity	M	p	BIC	ν	X^2	$\chi^2_{\nu,0.95}$
1	6f	92	4151.87	436	545.41	485.68
	6f.1	93	3991.47	435	370.10	484.63
	6f.2	94	4148.80	434	409.66	483.57
2	5d	123	5505.64	654	455.86	714.60
	5d.2	124	5672.35	653	538.63	713.56
3+	5a	41	1798.42	214	107.54	249.13
	5a.2	42	1878.82	213	151.55	248.05

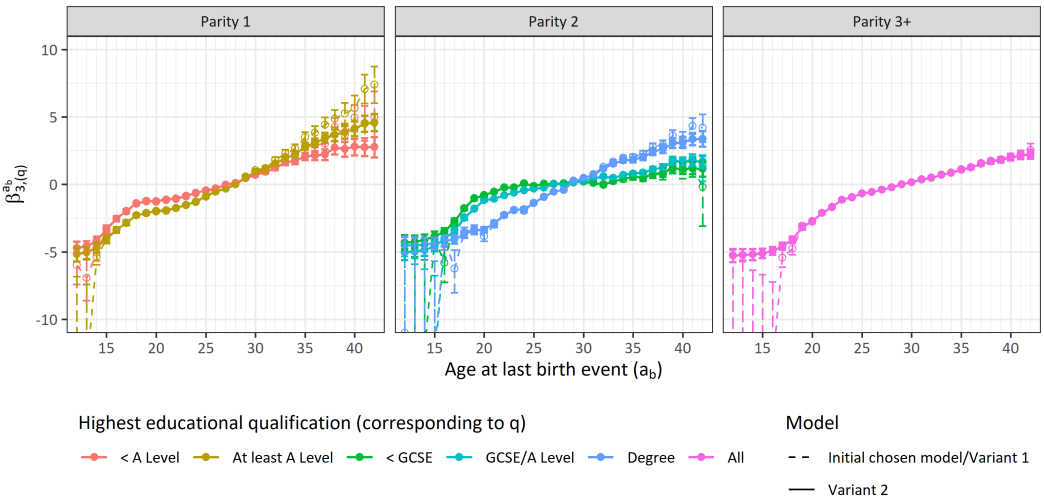

FIGURE 18 Parity-specific plots of the Q-specific/shared age at last birth parameter ($\beta^{a_b}_{3,(q)}$) estimates for the initial chosen models (parities 2 and 3+) or first variant (parity 1) and second variants; error bars indicate 95% CIs.

Inspecting the results for the second variants in Table 11, we see that that unlike in Section E.4 where the X^2 values either stayed basically the same or decreased, here they increase substantially. This is likely due to the original $\beta^{a_b}_{3,(q)}$ estimates tending to be extremely small at low A_b values (Figure 18). By smoothing these parameters we are forcing them to be considerably larger than they would otherwise choose to be, and hence causing the fit to be less close. However, the intended benefit of this second variant is to obtain fitted AT surfaces that would look more plausible for a larger population, even if they correspond less well with our particular survey population. It is also important to note that despite these increases in X^2, the values are still reasonable compared to $\chi^2_{\nu,0.95}$. We therefore take these second variants forward as our final chosen models for each parity. We present the corresponding Lexis surfaces of fitted probabilities for each parity in Figure 3 of the paper.

REFERENCES

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377–403.

Eilers, P. H. C. and Marx, B. D. (1996) Flexible Smoothing with B-splines and Penalties. *Statistical Science*, **11**, 89–121.

Ellison, J., Berrington, A., Dodd, E. and Forster, J. J. (2022) Investigating the application of generalized additive models to discrete-time event history analysis for birth events. *Demographic Research*, **47**, 647–694.

Stan Development Team (2019) *RStan: The R interface to Stan*. R package version 2.19.2. `http://mc-stan.org/`.

Umlauf, N., Klein, N. and Zeileis, A. (2018) BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond). *Journal of Computational and Graphical Statistics*, **27**, 612–627.

Wood, S. N. (2017) *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Press, 2 edn.