ECONOMIC GROWTH CENTER

YALE UNIVERSITY

P.O. Box 208269
New Haven, CT 06520-8269
http://www.econ.yale.edu/~egcenter/

CENTER DISCUSSION PAPER NO. 869

TWO STATISTICAL PROBLEMS IN THE PRINCETON PROJECT ON THE EUROPEAN FERTILITY TRANSITION

John C. Brown
Clark University

and

Timothy W. Guinnane
Yale University

September 2003

Notes: Center Discussion Papers are preliminary materials circulated to stimulate discussions and critical comments.

An earlier version of this paper was circulated under a slightly different title, and was revised while Guinnane was visiting the Faculty of Economics at the University of Cambridge. The paper is part of a larger project; see Brown, Guinnane, and Lupprian (1993) and Brown and Guinnane (2002) for additional discussion. For comments and suggestions we are grateful to Jan van Bavel, Theo Engelen, Avery Guest, Sriya Iyer, Christopher Meissner, Carolyn Moehling, Mary MacKinnon, Sheilagh Ogilvie, Cormac Ó Gráda, Barbara Okun, Matthias Schuendeln, and seminar participants at the Cambridge Group for the History of Population and Social Structure and the N.W. Posthumus Institute. This research was supported by the National Institute of Child Health and Human Development. We are especially grateful to Patrick Galloway for his comments and for helping us to construct extracts from his Prussian data.

Two Statistical Problems in the Princeton Project

On the European Fertility Transition

John C. Brown
Department of Economics, Clark University
Worcester, MA 01610
Jbrown@Clarku.Edu

Timothy W. Guinnane
Department of Economics, Yale University
New Haven, CT 06520
Timothy.Guinnane@Yale.Edu

Abstract

The Princeton Project on the Decline of Fertility in Europe (or European Fertility Project, hereafter EFP) was carried out at Princeton University's Office of Population Research in the 1960s and 1970s. This project aimed to characterize the decline of fertility that took place in Europe during the nineteenth and early twentieth centuries. The project's summary statements argued that social and economic forces played little role in bringing about the fertility transition. The statement stresses instead a process of innovation and diffusion. A central feature of the EFP argument is a series of statistical exercises that purport to show that changes in economic and social conditions exerted little influence on fertility. Two recent papers on Germany for this period have used similar data and methods to draw different conclusions. These findings echo those of researchers working in other contexts, who increasingly find that economic and social factors play a strong role in fertility. We show that one reason for the new findings is some general statistical problems in the Princeton methodology. These are reason to temper acceptance of the Princeton project's larger message.

Keyword: fertility transition

JEL codes: J13, N33, O15

The Princeton Project on the Decline of Fertility in Europe was a large-scale research project undertaken by the late Ansley Coale and his collaborators at Princeton's Office of Population Research in the 1960s and 1970s. The project compiled measures defined at the level of administrative areas for most western European countries and used this data to study the patterns of fertility decline and its correlation with possible explanatory factors. This research has been extremely influential,because of the project's scope and the skill and ingenuity of the individual studies. The project's overall conclusion, often called the "Princeton view," downplayed the importance of economic and social change in causing the fertility transition in Europe, and instead stressed a process of innovation and diffusion, driven by similar attitudes and communication networks.

Two recent studies of Germany in the late nineteenth and early twentieth centuries come to quite different conclusions. Patrick Galloway, Eugene Hammel, and Ronald Lee (hereafter GHL) studied Prussia, the largest of the German states, while our own research (hereafter BG) focused on Bavaria, which was the next-largest state. Both projects find a clear role for the economic and social forces that the Princeton project (hereafter "EFP") downplayed. These new results for Germany build on a much earlier paper by Toni Richards, who used the data underlying John Knodel's EFP monograph on Germany to come to conclusions at variance with his own. The difference in results is surprising because both of the more recent studies bear strong similarities to the EFP approach, while Richards' study is based directly on the EFP data.

This paper argues that much of the difference can be attributed to two problems in the statistical methods used by the Princeton authors. These problems are quite general and would affect studies other than those for Germany.[1] Our argument suggests caution in accepting the EFP conclusions. But our critique should not be exaggerated; we cannot show that the other EFP studies were wrong, we can only demonstrate the problems in the German case and note that related problems might affect other studies. On a similar note, we want to stress at the outset that the weaknesses we identify here reflect in a real way the EFP's strengths. The first problem we dis-

[1] Galloway et al (1998b, pp.195-208) surveys the methods used in recent research on the fertility transition. A recent paper by Potter *et al* (2002) uses methods similar to those advocated here, and also comes to the conclusion that social and economic forces have not been given due weight in explaining the fertility transition. The context for that paper is modern Brazil.

cuss is a natural consequence of the EFP's scope, and the second reflects its pioneering status.

We focus on two distinct issues.

- The use of aggregate data necessarily leads to a loss of efficiency in estimation. The GHL and BG studies both use aggregate data, but the administrative areas used in these studies are much smaller than those in the EFP, reducing the problem considerably.
- Coale's vision of the fertility transition focused very much on *change over time*. He argued that cross-area, pre-transition fertility *levels* might be interesting in their own right, but that these levels were not informative about the transition itself. Much of the statistical analysis actually undertaken by the EFP authors was, however, cross-sectional. Some of its analysis that was not cross-sectional still does not deal with important issues that arise in analysis of change over time. As we demonstrate, some of the differences between the Richards, GHL, and BG studies on the one hand, and the original Knodel monograph on Germany on the other, reflect the use of statistical models that more accurately differentiate cross-sectional from time-series changes.

The paper is not intended as an omnibus discussion of the European fertility transition or even the methods of the Princeton project. We focus on our two points and the interpretative issues they raise, and leave to other works (including our own) larger issues of interpretation and explanation.

1 The EFP and its vision

The project, it should be noted, was not originally conceived to address the questions we raise here. Coale and his collaborators originally designed the EFP to test the validity of the classical demographic transition theory in the European historical context. Coale noted as early as 1967 that the EFP data did not support transition theory, and a conference held in 1968 focused on what appeared to be the futility of any unified theory of the fertility transition. Subsequent work by the EFP scholars moved increasingly towards the regional and cultural arguments found in the summary statements.[2] Thus

[2] For discussion of the EFP's history, see Friedlander, Okun, and Segal (1999, pp.497-500). This exceptionally clear and comprehensive review warrants reading by anyone

the project's original conception, which is to say its original goals and the intellectual atmosphere in which it was initiated, played a strong role in creating the problems we identify here. The large administrative areas used in the project were probably sufficient to determine the inadequacy of the classic demographic transition model, as Coale argued. Only from the vantage of hindsight can one see, as we argue, that the statistical tools used by the project were not consistent with Coale's ultimate view of the fertility transition.

The EFP studies all took a slightly different approach to their country, reflecting data availability and the judgments of the individual author. We will not discuss any single work in detail, and by focusing on the statistical work we are overlooking thoughtful discussions of the fertility transition. In our characterization of the statistical methods used in the EFP we are thinking primarily of three monographs because these three contain the most extensive statistical analysis: Germany [Knodel (1974)], Belgium [Lesthaeghe (1977)], and Italy [Livi-Bacci (1977)].[3] Each of these studies shaped itself to the available information, the interests and concerns of the author, and the specifics of the country under study. There was a common EFP methodology in the following sense:

Units of analysis: All Princeton studies were based on aggregate data defined over administrative areas. The primary effort of the project was to compute a common series of fertility indices for each of these provinces, starting before the fertility transition. The EFP studies then examined the pace of change in the several provinces.

Measuring rods: The EFP authors relied on a set of four inter-related indices that in effect compare fertility in the population under study to the fertility of the Hutterites, a North American Anabaptist sect with well-documented, very high fertility. These indices were devised specially for use with the project. The index of marital fertility (I_g) can be thought of as the ratio of legitimate births in the population to the number of births one would expect in a Hutterite population with the same number of married women of the same ages. The index of proportions married (I_m) weights the pro-

interested in these issues.

[3] The Office of Population Research's website (www.opr.princeton.edu) has a list of all publications associated with the project. Our bibliography lists those most relevant to our discussion.

portions married at each age by the age-specific Hutterite fertility schedules. The index of non-marital fertility (I_h) is defined analogously to (I_g).[4] Overall fertility (I_f) is thus defined as the weighted sum of marital and non-marital fertility, where the weights are the index of proportions married:

$$I_f = I_m * I_g + (1 - I_m) * I_h \tag{1}$$

Measures of social and economic change: Most EFP studies use published information defined over the administrative areas to study the correlation between change in fertility on the one hand and social and economic development on the other. These measures include measures of urbanization, literacy, religion, and workforce allocation between agricultural and other pursuits, and other variables.

Statistical methods: All of the monographs estimate correlation or regression models intended to ask whether some variables or combination of variables can explain the patterns of fertility decline.

None of the monographs confined themselves to this methodology alone. All of them introduce other information and follow-up on issues suggested by the subject-matter or the author's own interests, and several of them make subtle use of data that did not fit into the common EFP methodology.

1.1 Conceptions of the fertility transition

Many discussions of the EFP turn on stated or unstated disagreements about what the fertility transition was, rather than what caused it. We have some reservations about the EFP image of a fertility transition. But to maintain our focus in this paper we set them aside in favor of asking how well the various studies applied Coale's notion to the concrete historical circumstance. Coale (1986) describes the major elements. Prior to the fertility transition all populations were characterized by natural fertility, Coale argued, so the fertility transition is the point at which some significant part of the population has adopted fertility-control measures. Coale, like Louis Henry before him, defined natural fertility as the absence of parity-specific fertility control. Parity-specific control means that the probability that a woman has her

[4] The indices lie between 0 and 1, but many authors find it convenient to multiply them by 1000. Appendix A provides definitions.

$N + 1st$ birth t months after the last birth depends on N. Natural fertility is consistent with a wide array of completed families sizes or fertility levels.

The EFP found that most provinces in Europe experienced a plateau in the level of marital fertility for some years prior to the transition. The level of this plateau varied widely; the mean of I_g was about .72, but ranged from .5 to nearly 1. Why did the level of marital fertility vary so much, even in populations that were (by assumption) not controlling fertility? As Coale explained it, "Marital fertility varied from one population to another because of differences in the prevalence and average duration of breast-feeding, periodic separation of spouses, etc." (Coale 1986, p.35). Coale argued that such variations are consistent with natural fertility.

These variations are not inconsistent with natural fertility. Rather, variations in the level of natural fertility were driven by local differences in behaviors that affected fertility but that did not, in Coale's vision, constitute conscious fertility control. The fertility transition itself is simpler:

> In the typical history of marital fertility in Europe, the plateau of I_g was interrupted by a decline that began at the time of the initiation of contraception or abortion (or both) among a large enough segment of the population to affect aggregate marital fertility; I_g then continued to fall, reaching a minimum in almost all instances ... of at least 50 percent below the plateau. An important feature of the history of I_g within each province is the date at which the sustained decline began. The decline is characterized as *sustained* because it was generally monotonic, except for postwar reversals, and continued to fall until a greatly reduced level was reached (Coale 1986, p.37).

Coale suggested that a convenient operational definition of the fertility transition was the date at which I_g had first fallen by 10 percent. This cutoff was selected on the grounds that once marital fertility declined this much it never rose again, so a 10-percent decline was safely irreversible. Several EFP monographs experiment with different (operational) definitions of the transition, but most focus on a ten-percent decline in I_g.

Coale's vision of the fertility transition, then, is that different administrative areas had very different pre-transition levels of natural fertility, depending on breast-feeding, spousal separation, etc. But he focused on *changes* in fertility.

1.2 Findings and interpretations

The EFP view on the relative unimportance of social and economic change is best understood within the context of a distinction laid out in a paper that was not part of the project itself. Carlsson (1966)'s two alternatives motivate many studies of the fertility transition. He put explanations of the fertility transition into one of two categories — innovation/diffusion or adaptation.[5] The innovation/diffusion view claims that the adoption of fertility control within a population represents a new behavior. The underlying reasons for the new behavior could be new medical knowledge, or new ways of communicating old knowledge, or changes in notions about the role of women in families or the moral acceptability of contraception. The adaptation view, on the other hand, claims that fertility control reflects couples' adaptation to changing economic and social circumstances. This distinction may not be as useful today as it was when Carlsson published his paper, but it is important to understanding the framework used by the Princeton studies.

The approach taken in the EFP monographs, and the broader interpretation advanced in the summary statements, are consistent with the following operational approach, which is a basic strategy in all empirical social science.[6] The project uses as its null hypothesis "changes in an indicator or set of indicators that proxy for social and economic change cannot explain the change in fertility." The alternative is "they can." The point of the empirical work is to construct statistical tests of the null; we see if the data can reject the null hypothesis that changes in some X have no effect on fertility. The EFP summary statements say that the studies could not reject the null.

This view that the proxies for social and economic change do not explain variations in the fertility transition is sometimes stated directly. A more common way to state the EFP conclusions is to say that the fertility transition occurred at virtually the same time across the provinces of Europe. Coale (1986) notes that for Europe as a whole, 53 percent of all administrative areas experienced their fertility transition between 1890 and 1920. For an event to have a common cause across European societies, the logic goes, there must

[5] Bean et al (1991) is an important study by historical demographers who stress the adaptation hypothesis. We are following their use of the term "adaptation" rather than Carlsson's term, "adjustment."

[6] One can ask serious questions about what the two different types of fertility transition would look like in practice, and what kind of data one would need to distinguish the two. But devising perfect tests is not our aim here. What is important is the relationship between the ideas and the empirical work.

be common features of those societies at the time of the fertility transition. But the project says there was not, or at least not enough to account for this apparent simultaneity in transitions. Knodel and van de Walle (1986) draw this inference from the information they summarize in their Table 10.1. This kind of observation is probably the source of the scholarly shorthand that says the EFP concluded that the only variable that explains the fertility transition is "date." In their more general criticism of "demand theories" of the fertility transition, Cleland and Wilson make a similar point: "clearly the simultaneity and speed of the European transition makes it highly doubtful that any economic force could be found which was powerful enough to offer a reasonable explanation" (Cleland and Wilson 1987, p. 18).

The EFP and its individual authors brought to bear a common set of techniques and evidence, along with many original contributions. But the statistical work about which we have reservations was central to the rejection of the adaptation view of the transition. As one of the more influential participants in the project put it, "Given the rough coincidence of modernization and the demographic transition, and the persuasiveness of the stories that were told to explain their relation, it is surprising that in country after country, the tests of the hypotheses embedded in demographic transition theory produced no certain confirmation of the theory" (Watkins, 1986, pp. 436-437). The repeated finding of the same, negative result for country after country played an important role in convincing the EFP participants, and others, of the validity of their view.

Are these conclusions warranted? In our view they must be tempered by an appreciation of the methodological problems we describe here. Our two points suggest that the EFP studies all reached similar conclusions in part because they all used methods that suffer from serious flaws. To provide concreteness we will use two historical data sets and some very simple statistical models in an effort to replicate the main tools of the Princeton project. If we could, we would estimate what we think are the right statistical models using the EFP data, and compare our new results to what was reported in the monographs. This is unfortunately not possible, for two different reasons. To address our first point we would need completely different datasets than those assembled by the Princeton project. We only have the "right" data for Prussia and Bavaria, that is, from the GHL and BG projects. Addressing our second point would be possible if the Princeton project had made all its data publicly available, but it did not.The data are unavailable not because the project researchers are uncooperative, but because the various studies were handled individually and at a time that predates widespread sharing of data in this form. With one exception, the project has made available the fertility indices but not the right-hand side variables required to estimate the explanatory models.[7] The exception is Germany, and for that exception the models we think are correct have been estimated and published by Toni Richards. We discuss her paper below.

1.3 Illustrative examples

Our datasets are from the German kingdoms of Prussia (1875-1910) and Bavaria (1880-1910). Galloway, Hammel and Lee have used the Prussian data in their published work, while we have used the Bavarian data for a substantive paper and use it here to provide concrete examples.[8] Both German datasets are based on units of observation that are much smaller than in Knodel's study. His data set is based on published information from 71 administrative areas in Germany, 30 of which are in Prussia. The Prussian dataset is based on the *Kreis*, the smallest administrative unit for which most data is available. The GHL team created 407 constant-territory *Kreise*, with observations every 5 years for the period 1875-1910. Bavaria in the late nineteenth century had eight provinces, and in Knodel's dataset Bavaria contributed eight observations. Our dataset is based on the *Bezirksamt*, the smallest administrative unit in the seven Bavarian provinces right of the Rhine.[9] We focus on the 138 *rural* districts, for which we have observations on 1880, 1885, 1895, 1900, and 1910. A full description of the source and more detailed variable definitions can be found in the published papers. For convenience we will refer to the small units (whether *Kreise* or *Bezirksämter*) as "districts." In both Prussia and Bavaria the larger unit that corresponds to Knodel's unit of analysis was called a *Regierungsbezirk*, which we will call a "province." We have and will continue to use the term "administrative area" in a neutral sense.

The differences in degrees of aggregation here are very large. In Knodel's dataset, the average Prussian province has a population of more than 900

[7] All available data have been posted at: http://opr.princeton.edu/archive/eufert.

[8] The Prussia data were kindly provided by Patrick Galloway. An earlier version of this paper also used simulated datasets. The results of those exercises are available upon request.

[9] Bavaria had an eighth province, the Palatinate, whose districts are not comparable to those in the rest of Bavaria. Descriptive statistics can be found in Appendix B.

thousand. In the GHL data the average district in 1900 has a population of about one-twelfth that figure. In the BG data the average Bavarian district is one-twentieth the size of the counterpart provinces in the Knodel dataset.

Our examples take a simple form: we regress the general marital fertility rate (GMFR) on the proportion of the district that is Catholic and the proportion that is urban.[10] This serves as a sort of "ideal type" of regression from this literature. Catholicism was expected to have a positive impact on fertility levels, and a negative impact on its decline, and in our examples can be thought of as the "cultural" or "ideational" variable. Urbanization's expected impact is just the opposite, and can be viewed here as the "social structure" or "economic" variable. No participant in debates about the fertility transition will find this an adequate model. There are many other variables to consider, and both the GHL and BG papers demonstrate the importance of richer economic information. But this simple model works nicely to illustrate the purely statistical points at issue here. The problems we illustrate here would affec both a richer model and a very simple model with different right-hand side variables. The model we use has the great virtue of having variables with the same definitions in Prussia and in Bavaria.[11]

2 The effects of aggregation

The EFP was based, perforce, on analysis of aggregate data. The project's scope made use of individual-level data impractical. There are some drawbacks to ecological analysis that cannot be surmounted with any type of aggregate data, but when aggregate data are all that is available (or, in the case of the Princeton project, all that is really compatible with the project's aims)

[10] The GMFR is defined as the number of legitimate births per married woman aged 15-49. Some of the EFP studies use the framework of partial correlation instead of linear regression. The two approaches are very similar, and what we say here would also apply to models of partial correlation.

[11] There is one difference. The Prussian districts comprise all of Prussia. Some districts, in fact, are 100 percent urban (such as the city-*Kreis* of Berlin). The Bavarian districts, on the other hand, are the *rural* administrative units of the kingdom. "Rural" for Bavaria meant "not having the legal status of a city" and in some cases our Bavarian districts are quite urban. This would be a serious problem if our aim were to make historical statements about Prussia and Bavaria, but that has been done already elsewhere. In all the examples given below, dropping the most urban Prussian districts did not materially affect the results.

those drawbacks must be accepted as the price of scope. The monograph authors were all aware of one statistical problem implicit in using aggregate data. Suppose we regress I_g on the proportion Catholic and the proportion urban. The estimates would tell us nothing about whether Catholic city-dwellers have higher or lower fertility than Protestant city-dwellers. We cannot claim that the regression coefficients from the aggregate data can recover individual effects. Claiming otherwise is to commit the "ecological fallacy," which none of the Princeton studies do.[12]

There is a different, serious problem caused by the large size and internal heterogeneity of the districts used. (Size is actually not the issue, but in many circumstances large size implies internal heterogeneity.) In some EFP studies, the administrative areas that are the units of analysis could be quite large. Some if not most of these provinces were quite heterogeneous internally. For example, one of Knodel's German provinces is Oberbayern, in Bavaria. This province covers 16,700 square miles and contains both the city of Munich and some of the most agricultural areas of Germany at the time. We can use our Bavarian district-level data to examine the heterogeneity missed with the larger units. The proportion Catholic is fairly uniform across Oberbayern's districts (ranging from 91 percent to almost 100 percent), but the proportion urban varies widely, from 0 percent to 62 percent. In Knodel's dataset Oberbayern's internal heterogeneity is all lost.

The EFP monographs, along with the summary volume, do address the question of aggregation. Watkins (1986, p.441) argued that the units used by the Princeton studies were sufficiently homogenous in their patterns of fertility decline that most of the variation in the decline was between provinces, not within provinces. Our examples show this not to be the case for Prussia and Bavaria. Others noted the possible *benefits* of aggregation. Livi-Bacci (1977, pp.137-142) noted that with very large units of observation, it is likely that short-term migration takes place within, rather than across, the units. Thus one possible benefit of high levels of aggregation is that it avoids problems caused by migration, problems we note below. Whether this small benefit is worth the larger problems we demonstrate is an empirical question. The examples we provide suggest not.

[12] There have been several advances in the statistical methods for use of ecological data since the EFP completed its work. These new methods are not addressed to an issue that is our concern. The main reference is King (1997). *Historical Methods* 34(3), 2001, is a special issue on the topic.

2.1 Aggregation and efficiency

Aggregation into internally heterogeneous units poses a serious potential trap. Suppose we wanted to estimate the relationship between an individual woman's fertility and some independent variable X. Assume first that we have individual-level data on N women. We could estimate a regression of the following form:

$$Fert = \alpha + \beta X + \varepsilon \tag{2}$$

Ignoring the impact of other influences, β could be estimated by ordinary least-squares (OLS). Suppose instead we take all of the individual women in the sample used to estimate (2), and assign them to the district in which they live. We then take means by district for both the right- and the left-hand sides and use the districts as the units of analysis. This is very much like what the EFP did, by necessity, although in their case the aggregation was done by the statistical authorities. If there are M districts, then our new regression will have M observations:

$$\overline{Fert} = \alpha_0 + \beta_0 \overline{X} + \varepsilon_0 \tag{3}$$

where the bars now denote that the observation is the mean value for a district. The naughts on α, β and ε will help us to remember that (2) and (3) are different equations.

What is the relationship between (2) and (3), especially between β and β_0? Many econometrics textbooks include a discussion that shows that an OLS estimate of β_0 is an unbiased estimator for the ungrouped case.[13] But β_0 is a less efficient estimator than β; the standard errors for β_0 will be larger than for β. Consider the expression for the standard error of the jth OLS regression coefficient:

$$SE(\beta_j) = \left[\frac{e'e}{n-K}(X'X)_j^{-1}\right]^{\frac{1}{2}} \tag{4}$$

where e is the vector of OLS residuals, n is the number of observations, K is the number of parameters estimated, and X is the matrix of independent

[13] In Johnston (1963) the discussion is on p.228-238. Cramer (1964) is a very clear discussion of the implications of aggregation in an applied context. There is a further complication that is not our point. Suppose the error term in (2) is homoscedastic. Even so, the error term in (3) is almost certainly heteroscedastic.

variables, and the subscripts j indicate the appropriate elements of the coefficient vector and the X'X matrix. Part of the loss of efficiency in aggregation results from the reduction in the degrees of freedom, n-K. More complicated changes result from changes in the regression's fit (e'e) and in the variation in the Xs (X'X).

Given the way the EFP authors set up their statistical tests, this point is crucial. A larger standard error means that it is harder to reject any particular null hypothesis. And this means that by using large units, *the EFP pre-disposed itself to concluding that any given variable on the right-hand side would not affect fertility.* That is, if we estimated equation (2), we might well conclude that X had a statistically significant effect on Fert, but if we estimated equation (3) we could conclude that $\overline{X}$ did not.

In addition, the R^2 goodness-of-fit measure from (3) will often (but not necessarily) be larger than the R^2 for (2). Intuitively, this happens because by aggregating we may be disposing of variation in the independent variables that is not strongly correlated with the dependent variable. Estimating equation (3) in preference to (2) could well lead to the conclusion that even in an equation that apparently explains the data well, $\overline{X}$ did not affect fertility. Usually we only estimate (3) when we cannot estimate (2), but it is important to bear this point in mind when thinking of (3) as a proxy for (2).

2.2 An example of the effects of aggregation

We can illustrate this problem using simple examples from the Prussian and the Bavarian data. For each German state, we estimate a regression using the district-level data, and then the parallel regression using the provincial-level data. The latter is analogous to what we would obtain using the EFP data. Table 1 reports results. The regressions are cross-sectional, to keep matters simple. The point at issue here does not depend on whether the regression is cross-sectional. Notice first that the point estimates for the district-level regression are similar to those for the province-level regression, with the exception of the urbanization variable for Bavaria. This just confirms what we noted before, that OLS estimates are unbiased for the grouped case. Now look at the effect of the grouping on standard errors. In both Prussia and Bavaria, moving to the larger units increases the standard errors considerably. (Recall that the Bavarian province regressions have seven observations.) In Prussia the standard error on proportion urban increases by nearly four-fold, and in Bavaria, by a factor greater than 40. In neither case does the aggregation affect in 1880 alter a substantive conclusion (proportion Catholic matters either way, proportion urban does not) but the very large effect on the standard errors warns that in other circumstances we could be failing to reject a null hypothesis for the wrong reasons. The R^2 measures show, in most cases, increases from aggregation.

Aggregation will tend to produce this problem in any circumstance. There are two separate forces at work. First, the provincial-level regressions have far fewer observations than their district-level counterparts. In the Prussian case, moving from 407 districts to 35 provinces increases the standard error by a factor of about 3.5 just because of the loss of degrees of freedom. Second, the loss of efficiency and the increase in R^2 both reflect the way the districts have been grouped into provinces. An old literature in econometrics studied the consequences of deliberately grouping individual observations to reduce computational burdens, a common practice prior to the the advent of cheap computing power. Cramer (1964) is a convenient summary of the main results. We can draw on those results to understand the implications of aggregation here. If the observations are grouped such that similar Xs are within a group, then there is a relatively small loss of efficiency and a relatively large increase in R^2 relative to the ungrouped case. This is because aggregation that puts observations with similar Xs in the same group preserves relatively more of the variation in that X.

Simple experiments with the Prussia data illustrate the point. We start with the cross-sectional regression for the Prussian districts in 1910, as reported in Table 1. Next, we sort the data by the value of Catholic and construct 25 groups. These 25 groups preserve as much the variation in Catholic as possible, because similar values of Catholic are assigned to a single group. Running a regression on the grouped data, we find Cramer's result: the standard error for Catholic is virtually unchanged from that reported in Table 1, while the standard error for Urban more than doubles (to .016). The R^2 for this regression rises to .93. Then we reverse the procedure: we sort the data by the value of Urban and construct 25 groups of districts. This time the standard error for Urban rises only slightly (to .009) while that for Catholic more than doubles (to.015). R^2, as we expect, increases to .97 An aggregation scheme that preserves relatively more of the variation in Catholic, we find, will affect the standard errors of that variable relatively less.

In the Princeton project, the grouping was not deliberate, it was produced by the historical processes that led to the regional distribution of religion, urbanization, and other potential explanatory variables. The spatial organization of German society implies that the high degree of aggregation in the Princeton project is relatively more likely to downplay factors such as urbanization. German provinces were generally either Catholic or not, for historical reasons. The same was not true of urbanization. Put statistically, in a one-way analysis of variance for 1880, province "explains" 69 percent of the variation in proportion Catholic in Prussia and 67 percent in Bavaria. The analogous ANOVAs for proportion urban explain 25 percent in Prussia and less than 1 percent in Bavaria. Thus when we aggregate up to the provincial level, we lose little of the variation in Catholicism, because that variation is mostly at the province level. The same is not true for urbanization, and we lose most of that variation via aggregation. Put differently, and referring back to Cramer (1964), the grouping of the Princeton project's provinces was, because of the historical record, less harmful for efforts to estimate the impact of Catholicism than of Urbanization.

Our results pertain, strictly speaking, to Prussia and Bavaria alone. But we suspect that a similar problem affects virtually all of the EFP studies. The problem we identify is inherent in the nature of city formation, in Germany and elsewhere in Europe. The centripetal forces of economies of scale at the level of firms and cities, and increasing specialization driven by declines in transportation prices, promoted increased differentiation at a local level. Some areas were increasingly urban, while others, quite near by, remained entirely rural and relied on the urban centers for the products and services of the city.[14]

The only way to know how much this aggregation problem affects results from other countries would be to replicate the sort of studies now available for Prussia and Bavaria. The efficiency losses depend on the amount of aggregation and the losses in variation between observations, and that is a strictly empirical question. Where data is available at a lower level of aggregation, it can be used to check on the results reported in the EFP.

3 Change over time

The other statistical problem in the EFP was the way it modeled change over time, that is, the fertility transition. Modelling change is difficult, and no

[14] This argument, which is hardly controversial in economic history, is stated forcefully by Hohenberg (Forthcoming).

single approach is uncontroversial. We cannot propose the single “correct” way, but we can note the drawbacks in the approach taken by the EFP.

Today most approaches to modelling change are a variant on panel-data techniques. The EFP data are all panel datasets, but the project itself never used these tools. Again, it is fair to note that the approach we suggest was not in widespread use when the EFP was conducting its research. Richards (1977) marks one of the first uses of panel models in demography.

3.1 How the EFP modelled change

The EFP monographs took four different approaches to the statistical problem of modelling change. First, many studies relied heavily on bivariate correlations. These correlations suffer from the problem of omitted variables bias. One may conclude incorrectly that X and Y are or are not correlated simply because of the correlation of X and Y with some omitted variable Z. This, of course, is also true of the illustrative models we report here, but presumably less so of the more complex models reported in the GHL or BG papers. Second, many of the exercises the Princeton project reports are purely cross-sectional; they regress fertility on some other variables at a point in time. This approach, which Thornton (2001) has called “reading history sideways,” is not consistent with Coale’s vision of the fertility transition, as is clear in light of our earlier discussion. These first two approaches were if anything more widely used than the third and fourth. Since they are inherently incorrect, reliance on them calls into question most of the tests of the causes of the fertility transition reported in the project volumes.

A third approach regresses the percentage change in fertility over a given period on the *levels* of several variables at the outset of the period:[15]

$$\frac{Y_0 - Y_1}{Y_0} = \alpha + \beta X_0 + \varepsilon \tag{5}$$

where Y_0 is the fertility measure in the first period, etc.[16] Here the Xs are all defined as of the first period. If we are examining the change in fertility between 1880 and 1900, then, the left-hand side would be the percentage

[15] Some studies distinguish percentage change from percentage decline. The distinction amounts to truncating the variable at zero; thus if fertility *rose* between the first and second dates, the value for its “decline” is entered as 0. Here we will ignore that distinction.

[16] Most EFP monographs used I_g as the fertility measure. Our point does not depend on the precise definition of Y.

change in fertility over the twenty-year period, while the right-hand side variables would be the level of cultural, social, and economic variables in 1880. This type of specification is consistent with testing certain types of models of fertility change, *but it is not a meaningful test of the adaptation hypothesis.* The adaptation hypothesis says that couples reduce the number of children they have as result of changes in their environment. Equation (5) asks whether fertility declines when, say urbanization reaches a certain level. Results based on this kind of model may be interesting, but cannot be used to address the ideas Carlsson (1966) laid out in his seminal paper.

The fourth approach used in the EFP studies is a variant on the following:

$$\frac{Y_0 - Y_1}{Y_0} = \alpha + \beta \frac{X_0 - X_1}{X_0} + \varepsilon \tag{6}$$

Equation (6) asks whether a change in a right-hand side variable is associated with a change in fertility. (In some specification, the dependent and independent variables are multiplied by 100 to make them percentage changes; in others, as in (6), they are estimated as proportionate changes. The difference is irrelevant to our point.) The percentage-change specification probably has two origins. On the one hand, it might be motivated by the criterion for the onset of the fertility transition (a 10-percent decline in I_g); on the other, it appears to remove the effect of initial levels by converting X and Y to percentage changes. At a general level this specification is entirely consistent with Coale's vision, and in principle is a direct test of the adaptation hypothesis. Unfortunately, there is an additional statistical complication that arises in modeling change. This complication was, in fact, implicit in the way Coale described the fertility transition. In the examples we show below, this problem is severe enough to call the results into serious question. Whether the same problems are present in all the EFP results we could not say without actually estimating new models with the other datasets

3.2 Panel approaches to modelling change

To see the problem it helps to step back to consider the data and the question we want to address. The datasets collected for the EFP studies all consist of repeated observations on the same districts. Suppose we have N districts and T years of data, so there are N x T observations in the dataset. A general way to study the relationship between an X and Y in such data would be to run the following equation:

$$Y_{it} = \alpha + X_{it}\beta + \varepsilon_{it} \tag{7}$$

where i subscripts the district and t the time period. We could estimate this model by OLS and, subject to the usual concerns, the results would be informative. But there are two, related reasons to estimate a different model.

First, equation (7) uses both the cross-sectional and the time-series variation. That is, the coefficients are estimated by taking account of the differences between districts at a point in time, and the changes in districts over time. But as Coale noted, we are primarily interested in the changes over time. We want to remove, as much as possible, the effect of cross-sectional differences at a point in time. Our examples below show that in at least our applications, models such as equation (7) can be driven mostly by cross-sectional variation, producing results that are misleading when interpreted in terms of change.

The second reason to estimate a different model is that it offers an opportunity to deal with a serious problem that Coale implicitly noted in calling attention to the differences in pre-transition fertility levels. We never have all the information we would like about any historical situation. If there is a variable that is missing but important, it can bias our results. Because we have repeated observations on these districts, however, there are ways to remove the influence of some forms of unobserved heterogeneity.

Suppose there is some variable D that is not in our dataset, but which influences fertility, as follows:

$$Y_{it} = \alpha + X_{it}\beta + D_i\delta + \varepsilon_{it} \tag{8}$$

If D is correlated with both Y and any X, then if we leave out D (that is, if we estimate (7) instead of (8)) our estimates of β will be biased. Suppose for the moment that D is what causes those large differences in pre-transition fertility levels across districts. If D is fixed over time for each district, we can in effect remove D by subtracting each value of X and Y from the within-district mean. This amounts to estimating:

$$Y_{it} = \sum_{i=1} \eta_i + X_{it}\beta + \varepsilon_{it} \tag{9}$$

where we have replaced D with a different constant term for each district (the η terms). This is called a fixed-effects estimator. The fixed-effects estimator is one of several different "panel" models.[17] Estimating some version of (8) is, in our view, both preferable on purely statistical grounds, and more true to Coale's idea of the fertility transition. Equation (9) strips out the initial differences across the districts, and focuses on *changes* in both X and fertility. The approach abstracts from whether fertility was high or low in the first period, and from whether the district was urban or not in the first period. It asks instead whether districts where urbanization increased also witnessed a decline in fertility.

How is this approach different from equation (6)? At first glance it might seem that the approach many EFP studies use pulls out the differences between districts, as well. The percentage changes used in (6) in effect standardize all variables in terms of percentage deviations from their initial levels. But that is the problem. There are three intuitive ways to think about the drawback to (6). First, consider the role of the constant term in equation (6) The specification forces it be to the same for all districts. This means that the baseline rate of change in fertility is the same for all districts. (To see that, consider a district where X did not change between the first and second period.) Second, the equation requires the relationship between Y and X to approximate that of a constant-elasticity function. Now consider two hypothetical districts. District one has a very high level of pre-transition fertility, while district two has a low level of pre-transition fertility. Suppose both experience an identical percentage change in the X variable. Can one model fit both cases? Only if district one has a much larger absolute decline in fertility, to produce a percentage decline equal to that of district two. That is, the same change in these two districts will not fit a simple model like this, because the initial fertility levels are used to scale those changes. This is just another way of saying that equation (6) does not pull out the effects of the initial fertility levels.

A third way to see this is to re-write (5) by multiplying through by Y_0:

$$Y_0 - Y_1 = \alpha Y_0 + \beta \frac{Y_0}{X_0}(X_0 - X_1) + \varepsilon Y_0 \tag{10}$$

Inspection of (10) shows that the equation requires that the *change* in fertility be a fixed proportion of the *initial level* of fertility. In addition, even

[17] One excellent introduction to panel models is contained in Greene (2000, Chapter 14). To simplify exposition here we do not discuss random-effects or other panel models. In our other work we found that the fixed-effects estimator was the best model for the Bavarian data.

if the estimated $\widehat{\alpha}$ is zero, so that $\widehat{\alpha}Y_0$ =0 for all values of Y_0, the second term on the right-hand side makes the change proportional to the ratio of the initial Y and X. (The regression would also be heteroskedastic, but that problem has straightforward solutions.) The approach taken in the EFP studies does not abstract from the initial levels to study change, as Coale argued; it conceals the effects of those initial levels.[18]

3.3 Some examples of the panel approach

Any early demonstration of the power of this approach came in 1977, when Toni Richards used Knodel's data to estimate panel models of the fertility transition in Germany. Her results are striking. She shows that the panel approach improved the model's explanatory power, sometimes dramatically. More importantly, it shifts the interpretation considerably. Without this approach she would have concluded that economic and social change explained almost none of the German fertility decline. Using the explicit panel framework, she concluded that economic and social change actually explains *most* of the German experience. This paper unfortunately never received the attention it deserved. Both the GHL and BG papers use fixed-effects models similar to (9). In our own work we experimented with a version of (6) but rejected it early on because it did not fit the data as well as the fixed-effects model.

We can get a sense of the problem by examining another set of simple models. We proceed in two stages to get a clearer idea of what is causing the problems. Table 2 presents both pooled and fixed-effects regressions that use all of the years available in both of our datasets. (A "pooled" model is like (7); it takes all N x T observations and treats them the same.) All four of these models use the district-level data. The point of the examples in Table 2 is to illustrate the importance of pulling out the fixed effects. The pooled regressions ask how Catholicism is related to fertility. The fixed-effects regressions ask how differences in Catholicism over time, within a district, are related to differences in fertility over time, within a district. The two models imply very different results. Most of the effects are sharper with the fixed-effects model, and the impact of Catholicism in fixed-effects model for Prussia has the "wrong" sign. We return to this point below.

[18] Equation (10) is not in principle objectionable, although it is a bit odd. Relative to a full panel specification, however, it incorporates several restrictions that are testable. Our concern amounts to saying that those restrictions should not be imposed *a priori.*

Table 2 also reports three different versions of the R^2 goodness-of-fit statistic for the fixed-effects models. The "within" measure is what we obtain if we estimate (8) by OLS. This highlights the model's ability to explain within-district change over time, which is our primary interest. The "between" version of the R^2 essentially discards all variation that is within a district over time, and runs OLS on the district means. This measure highlights the model's ability to explain the differences between the districts. The "overall" R^2 is obtained by running OLS on (7), the pooled model. This measure makes no distinction between explanation of variation within districts as opposed to between districts. Note that in Bavaria, most of the model's fit arises from its ability to explain cross-sectional differences; the model does a relatively poor job of explaining change over time. Again, the goodness-of-fit statistic here is not telling us much about what Coale emphasized, which is how changes in Xs explain changes in fertility.

Table 3 reports some examples that are a direct comparison of the EFP approach, equation (6), to a fixed-effects estimator. Here we have limited the sample for the fixed-effects estimator to the first and last years, to make the results directly comparable to the EFP approach. (Table 2 reports the same model with the full sample). For both Prussia and Bavaria the fixed-effects estimator fits the data much, much better. This should not be surprising; the fixed-effects estimator places much less structure on the data. Note that the fixed-effects specification noticeably sharpens the impact of urbanization.

Our fixed-effects estimators all produce results quite different from those that come out of either pooled models or simple cross-sections. This implies that the fixed-effects, which try to sweep out the effect of unobserved heterogeneity, are playing an important role. What are they? We cannot really say, because they are proxies for something unobservable. But at a mechanical level we can say that in the Prussian data the correlation between the estimated fixed effects and proportions Catholic is about .6. In the Bavarian data the correlation is reversed, about -.8. Correlations with proportion urban are much smaller. One way to think of this is to say that Catholicism is correlated with other factors that in Prussia imply initially higher fertility, and in Bavaria, initially lower fertility than one would expect given the observables. Whatever the interpretation, it is clear that failing to account for the unobservables yields a model that places misleading weight on Catholicism or any other variable that would be correlated with the unobservables.

3.4 More general issues

No single model is the obviously correct choice for modelling fertility change. There are two problems with the fixed-effect estimator that our own discussion has already highlighted. First, in deriving equation (9) from equation (8) we assumed that D, the unobservable, was fixed over time.What is D? If we think of it as the missing variable that explains cross-sectional differences in the pre-transition level of fertility, then assuming it is fixed over time might make sense. Coale mentioned as possible explanations for these pre-transition differences behaviors such as breast-feeding or spousal separation. Are these behaviors likely to be fixed in time? One could argue they are. If (as many claim, and Coale certainly thought) practices regarding breast-feeding and so forth are driven by deep cultural notions of the appropriate way to treat infants, or relations between the sexes, one would not expect them to change radically over a period as brief as that modelled here.

But this might not be true, which would pose a problem. The way we have modelled the fixed effects might be too strong; it might be that the unobservables are changing, as well. Our exposition relies on a very strong assumption that is not strictly necessary. One could assume, for example, that D is the same in the first two periods, then the same in the third and fourth, but can change between the second and third periods. But some assumption about the constancy of D is required to make the model work.

A second issue is implicit in the unexpected finding on Catholicism in the fixed-effects model for Prussia. The result implies that an *increase* in the proportion Catholic leads to a *decrease* in fertility. This is a most unexpected finding! In our BG paper with a much richer model than we report here,we found that the impact of Catholicism on fertility was positive, as expected, but implausibly large. The magnitudes of the impact were so large that it made us wonder whether Catholicism was not in part a proxy for something else, or a sign that something was amiss in our equation.[19] We began by noting that very few people in Bavaria changed their religions. With a fixed-effects estimator, the variation in Catholicism the model is picking up has to be within a district. What caused the changes in proportions Catholic? *Differential migration rates.* Areas that became more Catholic over time were those with strong in-migration caused by economic development. The devel-

[19] Galloway et al (1994, p.151) report a similar result, and interpret it as we do here. We also found that the impact of Catholicism declined over time, which is what one would expect.

oping areas had initially been mostly Protestant. After including measures of net migration we found that Catholicism still had the expected positive impact on fertility, but the magnitude was less. (We also used proxies for religiosity, which vary more over time and address the cultural hypothesis more directly.) More generally, results such as these are a warning for the methods we use, and would also be a problem with the approach the EFP used. We always have to ask where the variation over time is coming from. If people do not change their religions, then the variation in the proportion Catholic within a district over time has to be caused by religious differences in migration, fertility, or mortality. This would be true of any attempt to estimate the impact of a variable that does not change rapidly over time. Using smaller districts may in some cases exacerbate the problem, but that need not be the case.

Two smaller points are worth noting for their role in the literature. First, we have treated both of our explanatory variables as exogenous. Some variables important in fertility studies are arguably endogenous and should be approached as such. In both our study of Bavaria (Brown and Guinnane 2002) and one of the GHL team's works on Prussia (1998a) this issue was explored in detail. Second, one sometimes sees the claim that aggregated data are important because they are the only way to study the impact of phenomena that are in themselves aggregative. This is simply not true; the best way to study the effect of, say, a local religious ethos on fertility is to use data at the lowest possible level of aggregation, and to include in the statistical models variables that measure the religious ethos. This is what multi-level modelling is all about. If there is variation across individuals in that environmental variable, this effect be identified with individual-level data without the loss of efficiency that comes with aggregation.

4 Conclusions

The Princeton studies have been justly famous since their completion over twenty years ago. We can thank the Princeton authors, and especially Coale, for setting out an ambitious agenda and devising a methodology that would leave us with a broad vision of the fertility transition in Europe. Since their publication the individual monographs, and especially the summary statement, have been the subject of detailed discussion, praise, and criticism.

This paper emphasizes two general statistical problems that affect all of the Princeton studies. One reflects the project's scope. Aggregate data was all the EFP could work with, given its aims, but in general statistical estimation with aggregate data is liable to conclude that relationships are not important when they are. This problem alone can account for the EFP's rejection of the role of economic and social change in the fertility transition. The spatial organization of most European societies was similar to the German cases we have studied; the problem is especially severe. A second problem reflects the project's pioneering status. The largely cross-sectional nature of the statistical work reported in the EFP studies is not consistent with Coale's vision of the fertility transition, and does not constitute a clear test of the adaptation hypothesis the summary statements rejected. Panel approaches, which allow direct study of the effect of changes in social variables on changes in fertility, suggest quite different results.

Recent studies of the fertility decline in Prussia and Bavaria have used methods similar in spirit to the EFP. Both reach conclusions that are at odds with the "Princeton view." The reasons for the different German results illustrate these two limitations of the EFP. The level of aggregation in the two recent studies is much lower than in Knodel's study of Germany. Both of the recent studies also use panel approaches, which show a stronger role for economic and social change than was found by Knodel. This latter point echoes results reported by Toni Richard many years ago, on the basis of a panel approach and Knodel's own data.

In his presidential address to the Population Association of America, Arland Thornton discusses the influence and pitfalls of what he calls "reading history sideways" (Thornton 2001). He does not stress this connection, but one can view the EFP as an example of a project that read history sideways. (This is surely the justification for using cross-sectional regressions to test models of demographic change.) Thornton's clear, nuanced judgement on reading history sideways can be applied to most projects:

> . . . we, like our ancestors, frequently must rely on problematic data and assumptions. In this context, reading history sideways is simply a method that requires strong assumptions; violation of these assumptions can lead to faulty conclusions. Social scientists today, of course, are far more methodologically sophisticated than our ancestors. Moreover, our methodological humility should be increased by remembering the enormous negative impact of reading history sideways on the history of family and demographic

> studies. Thus I can conclude that cross-sectional approaches may be acceptable for exploratory purposes if we are clear about the assumptions and exceptionally cautious about the results (p.461).

As an exploratory project the EFP was unusually fruitful, ambitious, and influential. But if we are "clear about the assumptions and exceptionally cautious about the results," we will recognize that the summary statements rely in part on statistical analysis we should no longer trust. This is reason enough to press on with new sources and new methods.

References

Alter, George, 1992. "Theories of Fertility Decline: A Nonspecialist's Guide to the Current Debate." in John R. Gillis, Louise A. Tilly, and David Levine, editors, *The European Experience of Declining Fertility, 1850-1970.* Cambridge MA: Blackwell.

Bean, Lee, Geraldine Mineau, and Douglas Anderton. 1990. *Fertility Change on the American Frontier: Adaptation and Innovation.* Berkeley: University of California Press.

Brown, John C., Timothy W. Guinnane and Marion Lupprian, 1993. "The Munich *Polizeimeldebögen* as a Source for Quantitative History." *Historical Methods* 26(3): 101-118.

Brown, John C. and Timothy W. Guinnane, 2002. "Fertility Transition in a Rural, Catholic Population: Bavaria 1880-1910. *Population Studies* 56(1):35-49.

Carlsson, Gösta. 1966. "The decline of fertility: innovation or adjustment process."*Population Studie*s 20:149-174.

Cleland, John, and Christopher Wilson. 1987. "Demand theories of the fertility transition: An iconoclastic view."*Population Studies* 41:5-30.

Coale, Ansley J.; Watkins, Susan Cotts [editors], 1986. *The Decline of Fertility in Europe: the Revised Proceedings of a Conference on the Princeton European Fertility Project.* Princeton, NJ: Princeton University Press.

Coale, Ansley J.; Anderson, Barbara; Harm, Erna, 1979. *Human Fertility in Russia since the 19th Century.* Princeton, NJ: Princeton University Press.

Coale, Ansley and Susan C. Watkins, eds. 1986. *The Decline of Fertility in Europe.* Princeton: Princeton University Press.

Cramer, J.S., 1964. "Efficient Grouping, Regression and Correlation in Engel Curve Analysis." *Journal of the American Statistical Association*, 59(305), pp.233-250.

Friedlander, Dov, Barbara S. Okun, and Sharon Segal, 1999. "The Demographic Transition Then and Now: Processes, Perspectives, and Analyses." *Journal of Family History* 24(4): 493-533.

Galloway, Patrick R., Eugene A. Hammel and Ronald D. Lee, 1994. "Fertility Decline in Prussia, 1875-1910: A Pooled Cross-Section Time Series Analysis." *Population Studies* 48(1): 135-158.

Galloway, Patrick R., Ronald D. Lee, and Eugene A. Hammel 1998a. “Urban versus Rural: Fertility Decline in the Cities and Rural Districts of Prussa, 1875 to 1910.” *European Journal of Population* 14:209-264.

Galloway, Patrick R., Ronald D. Lee, and Eugene Hammel, 1998b. "Infant mortality and the fertility transition: macro evidence from Europe and new findings for Prussia," in *From Death to Births: Mortality Decline and Reproductive Change*, Washington D.C.: National Academy of Sciences, eds. Cohen, B. and Montgomery, M., Chapter 6, pp. 182-226.

Guinnane, Timothy W., Barbara S. Okun, and James Trussell, 1994. “What do We Know about the Timing of the European Fertility Transition¿‘ *Demography* 41(1).Knodel, John E, 1974. *The Decline of Fertility in Germany*, 1871-1939. Princeton, NJ: Princeton, University Press.

Hohenberg, Paul M., Forthcoming, “The Historical Geography of European Cities: An Interpretive Essay,” in V. Henderson and J.F. Thisse, *Handbook of Regional and Urban Economics*, vol. 4. Amsterdam: Elsevier Science.

King, Gregory, 1977. *A solution to the ecological inference problem: Recovering individual behavior from aggregate data.* Princeton: Princeton University Press.

Knodel, John, 1988. *Demographic behavior in the past: A study of fourteen German village populations in the eighteenth and nineteenth centuries.* New York: Cambridge University Press.

Knodel, John and Etienne van de Walle. 1986. “Lessons from the past: Policy implications of historical fertility studies.“in Ansley J. Coale and Susan C. Watkins, eds. *The Decline of Fertility in Europe.* Princeton: Princeton University Press.

Lesthaeghe, Ron J, 1977. *The Decline of Belgian Fertility*, 1800-1970. Princeton, NJ: Princeton, University Press.

Lesthaeghe, Ron and Chris Wilson. 1986. “Modes of production, secularization, and the pace of fertility decline in western Europe, 1870-1930.“in Ansley J. Coale and Susan C. Watkins, eds. 1986. *The Decline of Fertility in Europe.* Princeton: Princeton University Press.

Livi Bacci, Massimo, 1971. *A Century of Portuguese Fertility.* Princeton, NJ: Princeton University Press.

Livi Bacci, Massimo, 1977. *A History of Italian Fertility during the Last Two Centuries.* Princeton, NJ: Princeton University Press.

Potter, J. E., C. Schmertmann, and S. M. Cavenaghi. 2003. "Fertility and Development: Evidence from Brazil." *Demography*39(4): 739-762.

Richards, Toni, 1977. "Fertility Decline in Germany: An Econometric Appraisal." *Population Studies* 31(3): 537-553.

Teitelbaum, Michael S., 1984. *The British Fertility Decline: Demographic Transition in the Crucible of the Industrial Revolution.* Princeton, NJ: Princeton University Press.

Thornton, Arland, 2001. "The Developmental Paradigm, Reading History Sideways, and Family Change." *Demography 38(4):449-465.*

Van der Walle, Etienne, 1974. *The Female Population of France in the Nineteenth Century.* Princeton, NJ: Princeton University Press.

Watkins, Susan C. 1986. "Conclusions."in Ansley J. Coale and Susan C. Watkins, eds. 1986. *The Decline of Fertility in Europe.* Princeton: Princeton University Press.

Appendix A: Definitions of the Princeton indices

The definition of the index of marital fertility I_g is the ratio of legitimate births to a weighted sum of the number of married women in the population:

$$I_g = \frac{B_m}{\sum\limits_{a=15-19}^{a=45-49} m_a F_a}$$

where B_m is the number of births to married women, m_a is the number of married women in the age group a, and F_a is the Hutterite fertility schedule given below. The index I_h is defined by analogy, substituting births to unmarried women for B_m, and the number of unmarried women for m_a. The index of nupitality I_m is the ratio of a weighted sum of the number of married women in the population to a weighted sum of the number of total women in the population:

$$I_m = \frac{\sum\limits_{a=15-19}^{a=45-49} m_a F_a}{\sum\limits_{a=15-19}^{a=45-49} w_a F_a}$$

where w_a is the total number of women in the population. The schedule F_a as used in the project is:

Age	15-19	20-24	25-29	30-34	35-39	40-44	45-49
F$_a$	.300	.550	.502	.447	.406	.222	.061

Appendix B: Descriptive Statistics for the Prussian and Bavarian datasets

Means and standard deviations of variables used in examples

	Prussia	Bavaria
District-level datasets		
Fertility (GMFR)	.26 (.04)	.28 (.05)
Proportion Catholic	.35 (.37)	.78 (.32)
Proportion in urban areas	.30 (.19)	.12 (.12)
Province-level datasets		
Fertility (GMFR)	.26 (.04)	.28 (.03)
Proportion Catholic	.34 (.32)	.75 (.28)
Proportion in urban areas	.36 (.16)	.12(0.04)

Other characteristics of the datasets :

	Prussia	**Bavaria**
Number of district-level observations in a single cross-section	407	188
Number of province-level observations in a single cross-section	36	7
Number of cross-sections in the full panel	8	5

TABLE 1
The effects of aggregation

Sample	Constant	Proportion Catholic	Proportion Urban	Adjusted R-square
Prussia,1880 districts	.256 (.003)	.062 (.003)	-.007 (.007)	.44
Prussia,1880, provinces	.254 (.011)	.072 (.012)	-.015 (.026)	.53
Bavaria,1880, districts	.226 (.009)	.081 (.009)	-.011 (-.39)	.36
Bavaria,1880, provinces	.261 (.045)	.083 (.031)	-.390 (.399)	.50
Prussia,1910, districts	.229 (.003)	.093 (.004)	-.103 (.008)	.66
Prussia,1910, provinces	.252 (.016)	.089 (.016)	-.152 (.032)	.66
Bavaria,1910, districts	.189 (.008)	.102 (.009)	-.084 (.021)	.50
Bavaria,1910, provinces	.193 (.034)	.122 (.037)	-.214 (.223)	.60

Note: OLS estimates, standard errors in parentheses.

Source: Estimated from Prussian and Bavarian datasets described in the text.

TABLE 2

Fixed-effects and Pooled Regressions of Panel Fertility Data

	Type of regression	**Constant**	**Catholic**	**Urban**	$\underline{R^2}$
(1)	Prussia, districts with fixed effects 1875-1910	.506 (.009)	-.478 (.027)	-.265 (.010)	.308 .401 .285
(2)	Prussia, districts, pooled regression 1875-1910	.252 (.001)	.070 (.002)	-.053 (.003)	.440
(3)	Bavaria, districts with fixed effects 1880-1910	.119 (.094)	.216 (.121)	-.062 (.015)	.033 .444 .380
(4)	Bavaria, districts, pooled regression 1880-1910	.215 (.004)	.088 (.004)	-.035 (.011)	.381

Note: Standard errors in parentheses. The three values of R^2 in models (1) and (3) are the within, between, and overall measures discussed in the text.

Source: Estimated from the Prussian and Bavarian datasets described in the text.

TABLE 3

Modelling changes in fertility

	State and type of regression	Constant	Catholic	Urban	R^2
(1)	Pr ussia, percent changes 1875 to 1910 *N*=394	20.333 (.893)	-.003 (.002)	0.034 (.019)	.015
(2)	Bavaria, percent changes 1880 to 1910 *N*=135	17.47 (1.534)	.020 (.020)	.009 (.002)	.153
(3)	Pr ussia: fixed effects estimator 1875 and 1910 *N*=814	.546 (.023)	-.513 (.067)	-.379 (.027)	.440 .345 .193
(4)	Bavaria, fixed effects estimator 1880 and 1910 *N*=138	.215 (.162)	.104 (.207)	-.201 (.029)	.263 .410 .378

Note: Standard errors in parentheses.The three values of R^2 in models (3) and (4) are the within, between, and overall measures discussed in the text.

Source: Estimated from the Prussian and Bavaria datasets discussed in the text.