Routledge
Taylor & Francis Group

Econ. Innov. New Techn.

ISSN: 1043-8599 (Print) 1476-8364 (Online) Journal homepage: www.tandfonline.com/journals/gein20

# IDENTIFYING AGE, COHORT, AND PERIOD EFFECTS IN SCIENTIFIC RESEARCH PRODUCTIVITY: DISCUSSION AND ILLUSTRATION USING SIMULATED AND ACTUAL DATA ON FRENCH PHYSICISTS

Bronwyn H. Hall, Jacques Mairesse & Laure Turner

**To cite this article:** Bronwyn H. Hall, Jacques Mairesse & Laure Turner (2007) IDENTIFYING AGE, COHORT, AND PERIOD EFFECTS IN SCIENTIFIC RESEARCH PRODUCTIVITY: DISCUSSION AND ILLUSTRATION USING SIMULATED AND ACTUAL DATA ON FRENCH PHYSICISTS, Econ. Innov. New Techn., 16:2, 159-177, DOI: 10.1080/10438590600983010

**To link to this article:** https://doi.org/10.1080/10438590600983010

Published online: 29 Mar 2007.

*Econ. Innov. New Techn.*, 2007, Vol. 16(2), March, pp. 159–177

Routledge
Taylor & Francis Group

# IDENTIFYING AGE, COHORT, AND PERIOD EFFECTS IN SCIENTIFIC RESEARCH PRODUCTIVITY: DISCUSSION AND ILLUSTRATION USING SIMULATED AND ACTUAL DATA ON FRENCH PHYSICISTS

BRONWYN H. HALL[a,*], JACQUES MAIRESSE[b] and LAURE TURNER[c]

[a]*Department of Economics, University of California, 549 Evans Hall # 3880, Berkeley, CA 94720-3880;* [b]*CREST-ENSAE, 15 Bd Gabriel Péri, 92245 Malakoff Cedex, France;* [c]*CREST-LEI, 28 rue des Saints-Pères, 75007 Paris, France*

*(Received 10 May 2004; revised 30 May 2005; in final form 15 October 2005)*

The identification of age, cohort (vintage), and period (year) effects in a panel of individuals or other units is an old problem in the social sciences, but one that has not been much studied in the context of measuring researcher productivity. In the context of a semi-parametric model of productivity, where these effects are assumed to enter in an additive manner, we present the conditions necessary to identify and test for the presence of the three effects. In particular, we show that failure to specify, precisely, the conditions under which such a model is identified can lead to misleading conclusions about the productivity–age relationship. We illustrate our methods using data on the publications 1986–1997 by 465 French condensed-matter physicists who were born between 1936 and 1960.

*Keywords*: Scientific productivity; Age; Identification; Panel data; Bibliometrics

*JEL Classification*: C23; O31; J44

## 1 INTRODUCTION

Empirical studies in the social sciences often rely on data and models where a number of individuals born at different dates are observed at several points in time, and interest centers on the identification of age, cohort, and time or period effects in the relationship of interest. However, modeling and identification of such relationships has proved to be problematic, largely because of the obvious impossibility of observing two individuals at the same point in time who have the same age but were born at different dates. The identification problem is further aggravated if one uses standard panel data estimators in which one takes into first account the differences (or within individual differences) of the variables, in order to control for unobserved individual effects. In this case, the cohort effect disappears completely (because it is collinear with the individual effects), which obscures but does not eliminate the problem of identifying year and age effects simultaneously.

* Corresponding author. E-mail: bhhall@econ.berkeley.edu

ISSN 1043-8599 print; ISSN 1476-8364 online 
DOI: 10.1080/10438590600983010

A number of 'solutions' to this identification problem have been offered in the literature in different contexts (e.g., Hall, 1971; Mason *et al.*, 1973; Rodgers, 1982a,b; Mason and Fienberg, 1985; Berndt and Griliches, 1991), all of which assume restrictions on the specification of the general underlying model, usually by imposing some sort of functional form of assumption on the way the three effects enter. Hall (1971) was concerned with disentangling depreciation (the age effect), embodied technical change (the cohort effect), and disembodied technical change (the period effect) in a vintage capital model applied to trucks, in which he imposed the constraint that the two most recent vintages were identical in order to identify the model. Berndt and Griliches (1991) were interested in a problem similar to that confronting Hall: the construction of a hedonic pricing model of personal computers that incorporates technical change, vintage, and age effects. Unlike Hall, they explored and exposited the full range of assumptions available for identification of the additive dummy variable model.

During the same time this economic research was being undertaken, the problem had not gone unnoticed in the sociological literature, especially as it related to the interpretation of cohort effects. In a series of papers, William Mason and his co-authors proposed estimating cohort–age–period models using identification assumptions similar to the one used by Hall (1971). This work culminated in a conference volume published in 1985 (Mason and Fienberg, 1985) that provides an excellent overview of the state-of-the-art and the views of sociologists, statisticians, and economists on the problems associated with this kind of modeling, both conceptual and methodological.

One of the many domains in which this identification problem is prevalent is the study of the scientific productivity of researchers, where we would like to take simultaneous account of differing productivity over time, as a function of age, and as a function of the vintage of the researcher. Scholars in the sociology of science, and more recently economists, have tried to measure the age-related productivity curve, and to purge it of effects due to the vintage of the researchers and the periods in which they are being observed. A major problem in such analysis is the need to take into account two major tendencies: the exogenous increase of publications with time and with cohort. Descriptive statistics on scientific publications suggest that they tend to increase over time, more or less rapidly in many scientific fields, overall but also per researcher. A way to capture such time effects, as well as any general changes in the state-of-art and work environment is simply to introduce period (year) indicators in the model. In the same manner, it seems that younger cohorts tend to publish more than older ones when they were of the same age, which may be related to the fact that there are increased incentives and competition for the younger generations, and/or they are more motivated and better trained, and/or that the cost of publishing is less (with the use of computers and internet, and growing numbers of journals, etc.). However, including cohort indicators (or for that matter individual effects) together with period indicators in the model introduces the aforementioned identification problem with the age variable.

In this article, we give an overview of the general identification problem of age, cohort, and period effects in a panel data regression model, of the estimation and interpretation difficulties it raises, and propose what we think as a practical approach to deal with them (Sec. 2); we illustrate these difficulties and the suggested solutions on simulated data (Sec. 3), and on a rich longitudinal database of the publications over 20 years (1980–2002) of about 500 French condensed-matter physicists (Sec. 4). We have three goals in undertaking this work: (1) to illustrate the potential for such data to lead to misleading inference if the identification problem is overlooked or not confronted; (2) to discuss the estimation and interpretation of cohort–age–period models when there are individual effects; (3) to apply our methods to a panel of real data in order to draw some conclusions about the evolution of scientific research productivity over time and age.

We want to emphasize that we do not break new statistical ground on these questions here. Instead, we outline how to apply the methods proposed by previous researchers to the problem of scientist productivity and we explore the implications of the resulting estimates for substantive research questions, in particular to highlight the ambiguous nature of some of the previous results in this area. To put it in another way, we want to underline the importance of *a priori* assumptions in interpreting results from a cohort–age–period regression. Our research questions are the following: How do we interpret results when there are more than one way to achieve identification? What happens when we remove the individual effects (effectively removing the cohort information), and how do we interpret the results in this case?

## 2 THE AGE, COHORT, AND PERIOD IDENTIFICATION PROBLEM

### 2.1 Problem Statement

It is well-known that the identity age = year (period) − year of birth (cohort) implies that all three effects cannot be identified in a linear model. It is somewhat less well-known that identification can be achieved in a dummy variable model by dropping a small number of variables (e.g., see Berndt and Griliches, 1991). In fact, no experiment can be devised to identify a completely general model with cohort ($C$), age ($A$), and period ($P$) effects. Given the identity $A = P - C$ that exists in the data, it is obvious that any function $f(C, P, A)$ can be written as $f(C, P, P - C) = f(C, P)$, so that it does not depend on the value of $A$. Therefore, it will always be necessary to impose some constraints or prior information on $f(.)$, if we wish to identify an age effect that is parsimonious and not simply derived from the cohort–period behavior.

The requirement for parsimony is another way of saying that we expect the age effect to be rather smooth and slow to change, and that we would like to impose that belief on the model. Conceptually, if it were not for our *a priori* belief that things change slowly with age, we could simply derive the age effect from the observed cohort and period effects via the identity. In fact, some scholars (notably Rodgers, cited below) would argue that, in any case, this is the only solution available without external prior information such as the macro-economic environment. That is, the identification problem is fundamental, given the impossibility of observing $A$ such that $P - C \neq A$. Or alternatively, one could argue that the age effect is whatever we obtain from the interaction of period and cohort effects, and therefore is identified simply from the combination of those two effects. Unfortunately, an age effect identified in this way is not stationary among each cohort and cannot be presumed to apply when we look at a different time period. So, we might prefer to find a way to identify a more parsimonious age effect via reasonable assumptions on the cohort and period effects.

There exists a large body of prior research and debate in sociology, demography, and economics over the question of exactly identifying all three effects using suitable constraints on the functional form of the relationship or other prior information. In sociology, a rather heated debate over the identification between William Mason and his co-authors and Willard Rodgers was conducted in the pages of the *American Sociological Review* in 1982 (Mason *et al.*, 1973; Rodgers 1982a,b; Smith *et al.*, 1982). Mason *et al.* (1973) had proposed a method of identifying a model with three sets of dummy variables for age, period, and birth by constraining some of the coefficients, and Rodgers critiqued their approach strongly because of its *ad hoc* nature, arguing that a better method of identification was to replace one of the sets of dummy variables with 'real' variables that were correlated with that particular aspect of the relationship (i.e., replacing period dummies with variables describing the macro-economy during the period).

Part of his critique was based on the argument that modeling the effects as additively separable, already imposed too many constraints on the model and did not allow for interactions between, for example, cohort and changes over time.
Nevertheless, most researchers who are interested in identifying three separate effects, do begin by assuming that they are additive, that is, that

$$f(C, P, A) = f_C(C) + f_P(P) + f_A(A) \tag{1}$$

Clearly, when the $\{f_J, J = C, P, A\}$ are linear, we have the well-known case that one of the three functions is not identified. However, Heckman and Robb (1985) show more generally that when the $f_J$ are polynomials of order $J$, only $\binom{J+1}{J}$ combinations of the $\binom{J+2}{J}$ coefficients on the terms of order $J$ are identified. That is, for the linear model, only two of the three linear coefficients are identified. For a quadratic model, only three of the six quadratic coefficients are identified, and so forth. So, although low-order polynomials seem to be an attractive way to model these effects because of their smoothness, in practice, they have not been much used because the lack of identification is so obvious.

Given additivity, the most general semi-parametric model is a model that simply includes dummy variables for all three effects. However, if we do not impose additivity, a more general model is available, one which is simply the means of the dependent variable for each cohort–period combination. If there are no covariates other than cohort, age, and period, these means are the sufficient statistics for the data.[1] In the next section of the article, we begin with this model as our baseline and then present a series of models that are nested within it.

### 2.2 Model with Age, Cohort, and Year Dummies

Suppose that we have data on a variable of interest $Y_{it}$ on $N$ individuals from $C$ cohorts, observed for $P$ periods. If we have no prior information on the relationship of $Y$ to cohort and period other than assuming that it is multiplicative in levels (and therefore additive in logarithms), the natural semi-parametric regression model simply includes a dummy variable for each cohort–period combination. Such a method uses all the information available from the means of the data by cohort and period (and therefore age), and exhausts the degrees of freedom.

Using lower case $y$ to denote the logarithm of $Y$, this model can be written as saturated.

$$y_{it} = a_{ct} + \varepsilon_{it} \tag{2}$$

where $i = 1, \ldots, N$ individuals; $t = 1, \ldots, P$ periods; and $c = 1, \ldots, C$ cohorts. We are implicitly assuming that the data are balanced across $P$ and $C$ (although not necessarily across $N$). That is, for each cohort we observe a complete set of $P$ periods.[2] Given the assumption of balance in the $P$ and $C$ dimension, when we observe $P \times C$ cells, we are observing $A = P + C - 1$ ages. This model, which we call the saturated model, allows us to identify PC means of $y$, one for each cohort–period combination. However, writing the model this way neither provides estimates of age, cohort, or period effects separately, nor it imposes constancy on these effects.

---

[1] Strictly speaking, we would also need the total and within variances of the dependent variable and an assumption of normality and conditional homoskedasticity for these means to be sufficient.

[2] Symmetrical treatments where the data are balanced for $C$ and $A$ (we observe the same number of ages for each cohort, and therefore periods are unbalanced), or where the data are balanced for $P$ and $A$ (we observe the same number of ages in each period, and therefore cohorts are unbalanced) are possible. We present the $C$ and $P$ case here because that is the way our data is organized: the changes necessary to estimate with data balanced for $C$ and $A$ or $P$ and $A$ are obvious, but tedious.

As the saturated model is the most general model that can be estimated using this type of data, it is a useful starting point, but most researchers prefer to impose constancy of the coefficients across the same ages, cohorts, and periods, which leads to a model, which we call the three-way or CAP model

$$\text{CAP: } y_{it} = \mu + \alpha_c + \beta_t + \gamma_a + \varepsilon_{it} \tag{3}$$

We know that one cannot estimate Eq. (3) directly: the coefficients of the different indicators can only be estimated relative to a reference value for each of the three dimensions. Therefore, one imposes (for example) nullity on the coefficients $\alpha_1$, $\beta_1$, and $\gamma_1$, which are, respectively, those of the first cohort, the first period, and the first age. However, the collinearity between the indicators of age, period, and cohort has not been removed by this procedure: in fact, it is easy to show that even the variables in this new equation will not be linearly independent. How can one then estimate this model? As discussed earlier, several methods of identification have been proposed in the past; here, we focus our discussion on those that involve simple restrictions on the dummy variables, rather than the addition of new variables such as macro-economic indicators.

Mason *et al.* (1973) proposed determining the number of restrictions which are necessary to impose on Eq. (3) in order to eliminate the problem of collinearity and identify the model. They demonstrated that one possible sufficient condition is to constrain two coefficients in the same dimension (age, period, or cohort) to be equal. For example, by imposing that the effects of the first and last ages are equal, one can identify the model, provided that there are at least 12 cohort–period combinations. The number of coefficients that can be estimated is, therefore, $1 + (P - 1) + (C - 1) + (A - 1) - 1 = 2(P + C) - 4$, as compared with PC of the saturated model. When $P = C = 2$, the three-way model coincides with the saturated model, implying that at least one of $P$ or $C$ must be larger than 2 for this model to impose meaningful constraints.

Naturally, the problem with identifying the three-way model using an equality constraint on two of the coefficients is that the different equality constraints will correspond to different estimates of the coefficients. The explanatory power of the models (measured by the $R^2$) estimated under the different equality constraints will be the same. As a consequence, in the absence of an equality constraint that is preferred *a priori*, identifying the model in this way does not allow the selection of a 'good' model. A secondary problem is that the identification may be fairly weak, relying as it does on a single equality constraint between coefficients.

However, as Berndt *et al.* (1995) showed, when the number of periods and cohorts is large enough, the three-way model imposes a number of constraints on the saturated model that can be tested in order to determine its plausibility. Similarly, models with only two or one set of dummies are nested within the three-way model, so that it is possible to test their validity using either the saturated or the three-way model as the maintained hypothesis. We write the two-way models as follows:

$$\begin{aligned} \text{CP: } y_{it} &= \mu + \alpha_c + \beta_t + \varepsilon_{it} \\ \text{CA: } y_{it} &= \mu + \alpha_c + \gamma_a + \varepsilon_{it} \\ \text{PA: } y_{it} &= \mu + \beta_t + \gamma_a + \varepsilon_{it} \end{aligned} \tag{4}$$

and the one-way models similarly

$$\begin{aligned} C\text{: } y_{it} &= \mu + \alpha_c + \varepsilon_{it} \\ P\text{: } y_{it} &= \mu + \beta_t + \varepsilon_{it} \\ A\text{: } y_{it} &= \mu + \gamma_a + \varepsilon_{it} \end{aligned} \tag{5}$$

For example, testing the CP model against the CAP model is equivalent to testing whether the $A-1$ coefficients $\gamma_2, \gamma_3, \ldots, \gamma_A$ are equal to zero, which corresponds to testing the constraints on the saturated model given in Table I (similar tables apply for the other models).

TABLE I Constraints for the cohort-period model.

| *Periods/ cohorts* | $P_1$ | $P_2$ | $P_3$ | $P_4$ |
|---|---|---|---|---|
| $C_1$ | $a_{11} = \mu$ | $a_{12} = \mu + \beta_1$ | $a_{13} = \mu + \beta_2$ | $a_{14} = \mu + \beta_3$ |
| $C_2$ | $a_{21} = \mu + \alpha_1$ | $a_{22} = \mu + \alpha_1 + \beta_1$ | $a_{23} = \mu + \alpha_1 + \beta_2$ | $a_{24} = \mu + \alpha_1 + \beta_3$ |
| $C_3$ | $a_{31} = \mu + \alpha_2$ | $a_{32} = \mu + \alpha_2 + \beta_1$ | $a_{33} = \mu + \alpha_2 + \beta_2$ | $A_{34} = \mu + \alpha_2 + \beta_3$ |

The implication of this particular set of constraints is that the change in $y$ from period-to-period is the same for each cohort, but that the change in $y$ from age-to-age is different for each cohort. The number of constraints relative to the saturated model is equal to $PC - (P-1) - (C-1) - 1 = (P-1)(C-1)$, which can be a sizable number. In Section 4 of the article, we present empirical results for a panel of French physicists which has 25 cohorts and either 12 or 21 periods. For the shorter sample using these data, the number of implied constraints for the *CP* model is equal to 264 out of 300 coefficients. Table II gives the general formulas for the number of constraints in all the models when the data are balanced in the cohort and period dimension. Table III illustrates the computations for our two panels, where we have $i = 1, \ldots, N$ individuals ($N = 465$ for the short panel and 418 for the long); $p = 1, \ldots, P$ years ($P = 12$ or 21); $c = 1, \ldots, C$ cohorts ($C = 25$); and therefore age $a = t - c$ ($A = P + C - 1 = 36$ or 45).

TABLE II Number of parameters and constraints for the different models.

| *Model* | *Free parameters* | *Number of constraints* | *Minimum P, C for over identification* |
|---|---|---|---|
| Saturated | $P \cdot C$ | 0 | NA |
| CAP | $P + C + A - 3 = 2(C + P) - 4$ | $(P-2)(C-2)$ | $P = 3, C = 3$ |
| CP | $P + C - 1$ | $(P-1)(C-1)$ | $P = 2, C = 2$ |
| CA | $C + A - 1 = 2C + P - 2$ | $(C-1)(P-2)$ | $P = 3, C = 2$ |
| PA | $P + A - 1 = C + 2P - 2$ | $(P-1)(C-2)$ | $P = 2, C = 3$ |
| *C* | $C$ | $C(P-1)$ | $P = 2, C = 1$ |
| *P* | $P$ | $P(C-1)$ | $P = 1, C = 2$ |
| *A* | $A = P + C - 1$ | $(P-1)(C-1)$ | $P = 2, C = 2$ |

TABLE III Number of parameters and constraints for the data.

| | *Short sample* | | *Long sample* | |
|---|---|---|---|---|
| *Model* | *Free parameters* | *Number of constraints* | *Free parameters* | *Number of constraints* |
| Saturated | 300 | 0 | 525 | 0 |
| CAP | 70 | 230 | 88 | 437 |
| CP | 36 | 264 | 45 | 480 |
| CA | 60 | 240 | 69 | 456 |
| PA | 47 | 253 | 65 | 460 |
| *C* | 25 | 275 | 25 | 500 |
| *P* | 12 | 288 | 21 | 504 |
| *A* | 36 | 264 | 45 | 480 |

It is clear from these tables that when there are a large number of years or cohorts, there are a large number of implied constraints. The implication is that even though it is not possible to identify a model with a full set of cohort, year, and age dummies, it is still possible to test for the presence of any one set of these dummies conditional on including the other two sets. That is, because only one additional constraint is required to identify the model with all three effects, when more than one additional constraint is implied by dropping a set of dummies, we can still perform a test. As mentioned earlier, in the case of data balanced in the cohort and period dimension, this will be true when either the number of periods or the number of cohorts is at least three.

### 2.3 Including Individual Effects

In many situations, it is desirable to control not only for effects due to the cohort to which an individual belongs, but for permanent differences in individuals as well, leading to a variation of the CAP model:

$$\text{IAP: } y_{it} = \mu_i + \alpha_c + \beta_t + \gamma_a + \varepsilon_{it} \tag{6}$$

It is obvious that this will create a further identification problem: given any individual $i$, the cohort $C$ to which he belongs is known, and the cohort effect $\alpha_c$ is, therefore, completely unidentified in a model with individual effects. In addition, some of the identification strategies discussed above (specifically those involving constraining the cohort dummies) are unavailable, because including individual effects necessarily involves including a complete set of cohort dummies. One additional danger in including individual effects in these models (and as a consequence differencing out the cohort effect) is that the identification problem itself is therefore obscured and may be missed by the researcher.

Heckman and Robb (1985) discuss the identification issue in CAP models with individual effects and suggest an alternative identification strategy using a variance components decomposition. That is, they propose modeling using random effects in cohort, age, and period, and then estimating the model using the moment matching methods associated with Joreskog's LISREL program (2005).

## 3 AN ILLUSTRATION USING SIMULATED DATA

In order to illustrate the identification problem and the difficulties it creates for measuring age effects in researcher productivity, we performed a simulation using data calibrated to match the panel of French physicists analyzed in Turner and Mairesse (2003) and also in section 4 of this article. That dataset had observations on the publications of 465 individuals who were born between 1936 and 1960 (25 cohorts), for the period 1986–1997 (12 years). In this section of the article we present the results of a series of structured statistical tests on the simulated data that are designed to choose the correct model from among the various dummy variable alternatives discussed earlier. In addition, we show the results of one draw from our simulation graphically. The model we chose for simulation illustrates the potential for a model that has only cohort and period effects to generate data that may appear to have peak in productivity at a certain age in spite of there being no age effect in reality.

Our approach here is to generate data, that looks like the real data, using a negative binomial model (so we obtain counts with overdispersion), but to estimate using the log-linear dummy variable model that is common in the literature. Given the generally small values of the dependent variable and the fact that we are using dummy variables, the differences in using OLS or using the more correct ML on a negative binomial model for estimation are likely

to be slight. Figures 1(a)–(c) show the results of one simulation draw of the model given below:

$$\begin{aligned} y_{it} &\sim \mathrm{NB}(\lambda_{it}, \sigma^2) \\ \lambda_{it} &= \exp[\mu_0 + \alpha t + \beta c + \gamma c^2] \end{aligned} \tag{7}$$

where NB denotes the negative binomial distribution, $t$ is the period (1986–1997, centered at 1991.5), $c$ is the cohort (1936–1960, centered at 1948), $\alpha = 0.022$, $\beta = 0.001$, $\gamma = -0.0015$, $\mu_0 = 1.09$, and $\sigma = 3.1$. $\mu_0$ and $\sigma$ were chosen so that the logarithms of the simulated data had the same mean and variance as the actual data. These parameter values imply that the quadratic in $C$ reaches its maximum in about 1948–1949, in the middle of our data period, but that the slope ranges from a $-4\%$ growth rate to a $+4\%$ growth rate in publications per year for the observed cohorts and is usually much lower, of the same order of magnitude as the year effect. In levels and at the mean of the data (2.7 publications per year), 4% corresponds to a growth rate of about 0.1 article per year.

Each panel of Figure 1 shows the resulting data from one draw of this simulation, plotted in three different ways. Figure 1(a) shows the means by age, Figure 1(b) the means by year, and Figure 1(c) the means by cohort. In each case, we also show the best fit line for the dummy variable model that excludes the variable on the $X$-axis, as a guide to the eye. Note that any dummy variable model which includes a set of dummies for the $X$-axis variable will fit the means of the data perfectly. For example, in Figure 1(a) we show the fit from a model that includes only the cohort and period dummies (the CP model). Any model that includes age dummies (i.e., the CAP, CA, PA, or $A$ models of Section 2) would have matched the overall age means exactly. Of course, were we to examine the fit of the age distribution for particular cohorts or particular years, only the saturated model would be able to match the data exactly. This fact is illustrated in Figure 2, which shows the data and the fit of the various models for three separate cohorts (1936, 1948, and 1960) that have three sets of non-overlapping ages (50–61, 38–49, and 26–37).

The main message of Figure 1 is that although the year and cohort distributions look the way we would expect, given the simulation, the resulting age distribution exhibits smooth behavior with peaking during the 40s, even though there is no age effect in our simulated model. As we expected, the year distribution shows a modest trend increase of about 0.06 publications on an average throughout the 12-year period and the cohort distribution a slight peaking tendency in the late 1940s. Our conclusion is that for samples of our size, averaging approximately 17 observations per period–cohort cell, it would be possible to observe a peaked age effect even if one is not there, at least if there is curvature in the cohort or period dimension.[3] That is, the observed age effect can be generated simply by the interaction of period and cohort effects.

What are the implications of this 'age' effect for model selection? That is, even though we observe something that looks like an age effect in Figure 1(a), the testing strategy outlined in section 2 of the article may allow us to choose correctly among the many possible models that are given in Tables II and III, at least when the number of cells or observations are large enough, and to reject models that are inappropriate for the data. The tests corresponding to the eight different models in Table II are nested in the way shown in Figure 3: the three-way CAP model is nested within the saturated model, the three two-way CP, CA, and PA models are nested within the CAP model, and the three one-way $C$, $P$, and $A$ models are nested within either of their corresponding two-way models. Thus, we can test for the correct specification

[3] In this investigation, we have focused on a quadratic age effect because that is a typical finding of the human capital literature and is therefore of considerable interest to researchers of scientific productivity. Spurious linear age effects would be even easier to generate using trends in period and cohort.

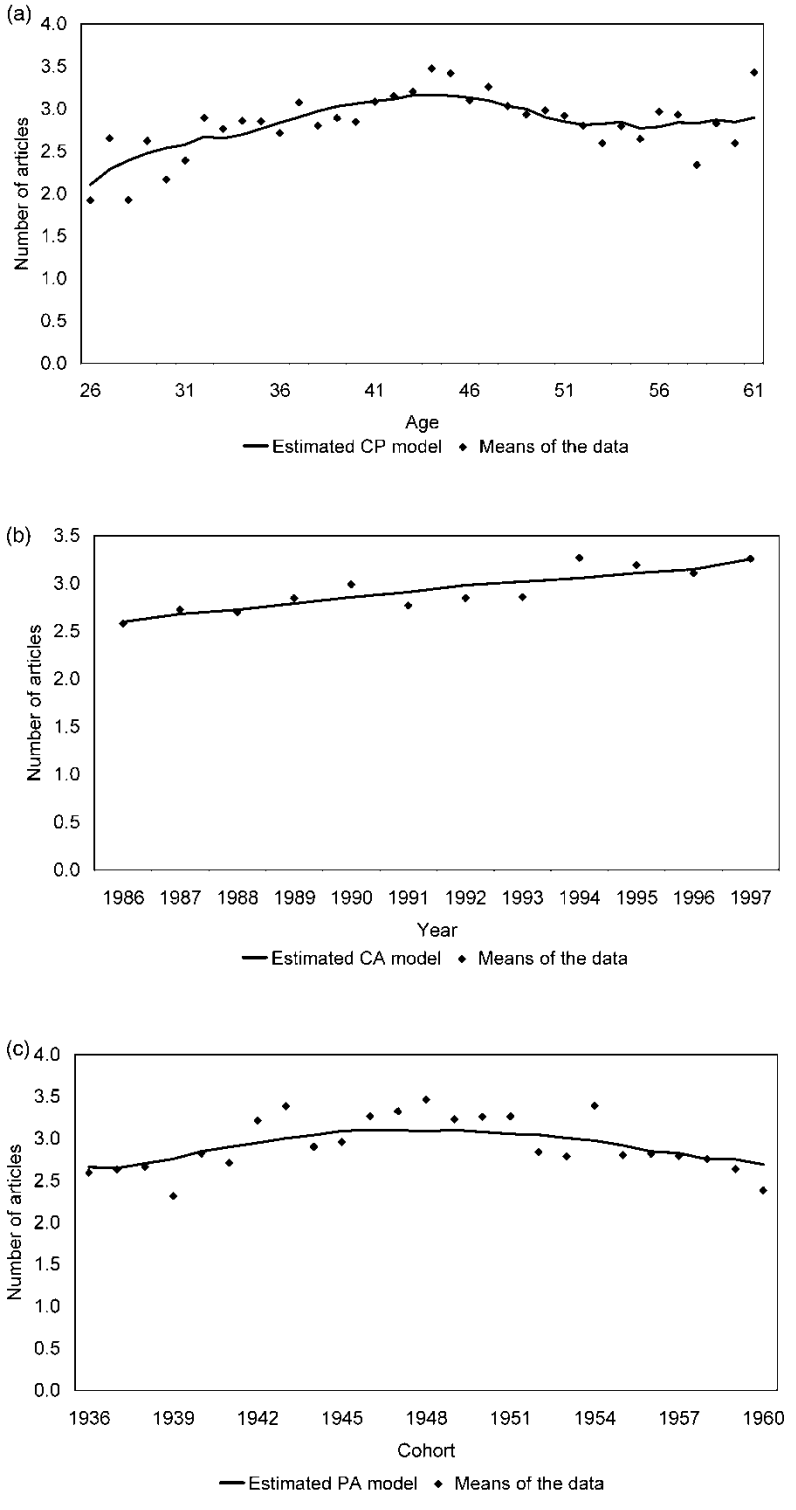


FIGURE 1 (a) Data simulation with cohort and period effects only, plotted versus age; (b) Data simulation with cohort and period effects only, plotted versus period; (c) Data simulation with cohort and period effects only, plotted versus cohort.

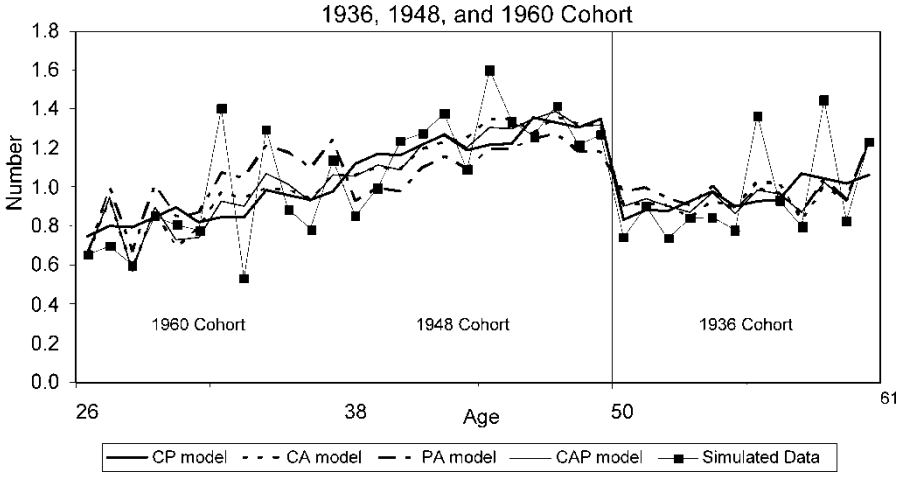


FIGURE 2 Estimated models for data simulated with cohort and period effects.

using a general-to-specific sequence of tests: first, we test the CAP model using the saturated model as the maintained hypothesis, and if we accept, then we can test the three two-way models (CP, CA, and PA) using the CAP model as the maintained hypothesis. Each of the three two-way models has nested within it two one-way models and each one-way model ($C$, $A$, and $P$) can be obtained from two different two-way models. Because this is a sequence of nested tests, one might want to adjust the corresponding significance levels when conducting

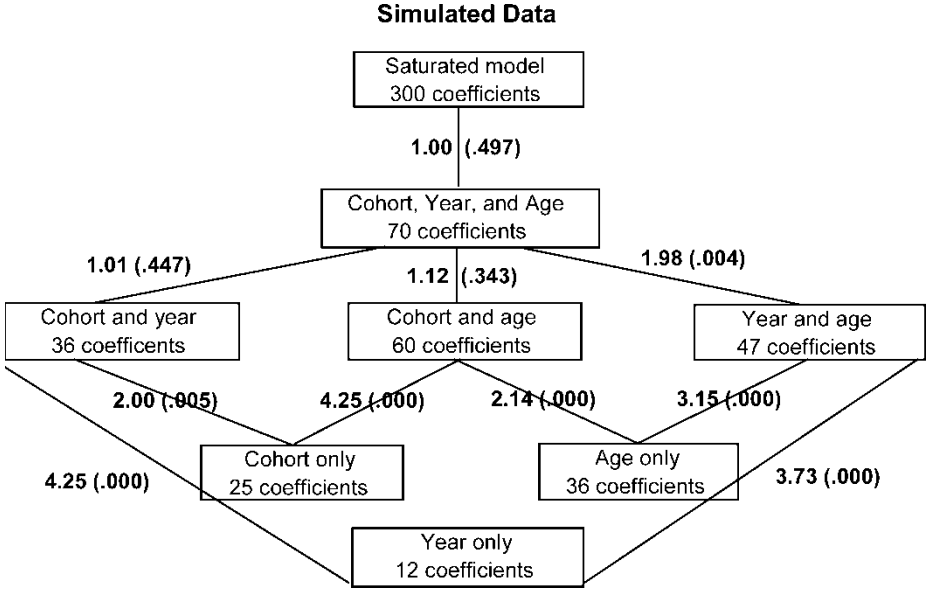


FIGURE 3 $F$-tests for cohort, age, and period models.

the tests. In the case shown here, the tests are sharp enough (either very significant or very insignificant) that such an adjustment would make little difference.[4]

As we suspected from Figures 1 and 2, for the simulated data the results of this model selection approach are somewhat ambiguous. Given the saturated model, we can easily accept a model with cohort, year, and age effects only [$F$-statistic (230, 5280) = 1.00], but conditional on that model, there are two models that will describe the data accurately: one is the cohort-year model [$F$-statistic (34, 5050) = 1.01], which is consistent with the data generating process we used for the simulation, and the second is the cohort-age model [$F$-statistic (10, 5050) = 1.12], which is not. The year–age model and the three one-way models are clearly rejected, regardless of the model that is taken as the maintained hypothesis. Our conclusion is that for data like ours, it may be difficult to discriminate between some of the models using samples of the size available to us, although clearly we are able to reject the more restrictive specifications. In particular, we are likely to find age effects in a model that has only a linear time trend and a smooth quadratic cohort effect. In the next section of the article, we apply the same sequence of tests, this time to the real data, and reach similar conclusions.

## 4 AN APPLICATION USING DATA FOR A PANEL OF FRENCH PHYSICISTS

There are many studies of age and/or gender differences in research production in the sociology of science and in scientometrics (for example Cole, 1979; Cole and Zuckerman, 1984; Cole and Singer, 1991; Bonaccorsi and Daraio, 2003). Economists have also investigated them in the framework of cumulative advantage models and/or life cycle models (Diamond, 1984; Levin and Stephan, 1991; David, 1994; Stephan, 1996 and 1998). These models reveal the consequences of events arriving in the early career of the scientist on the one hand and of the anticipation of the coming end of career on the other on the allocation of research efforts over time and individual productivity. However, there has been relatively little research based on individual panel data, which could allow disentangling the effects of age and gender from cohort and period effects, as well as from other unobservable individual effects. One of the few exceptions is Levin and Stephan (1991), in which the proposition that research activity declines over the life cycle is tested on publication panel data for scientists in six sub-fields of earth science and physics (including condensed matter physics), over the period 1973–1979.

### 4.1 The dataset

The database with which we work is an original panel database that was created from the records of 523 French condensed-matter physicists working at the *Centre National de la Recherche Scientifique* (CNRS) between 1980 and 2002, and born between 1936 and 1960. Condensed-matter (solid-state) physics comprises half of all French academic physics. During the period of study, it was a rapidly growing field with relatively little mobility towards the private sector or the universities, and with well-identified journals.[5] The group of physicists studied here represents a majority of all CNRS researchers in this discipline (they numbered 598 in 2002). The CNRS and universities are the main public research institutions in this domain in France. In 2002, 28.3% of the condensed-matter physicists in France belonged to the CNRS and 70.5% to the academic sector (1489 researchers).

[4] For example, to obtain a nominal size of 5% for the overall test, the individual significance level for each of the three nested tests should be somewhat smaller, equal to $[1 - (1 - .05)^{1/n}] = 0.017$ with $n = 3$ [see Kennedy (1996), p. 92, for a discussion of nested testing].

[5] For further information on the database and its creation, see Turner (2003).

TABLE IV Sample statistics for 465 CNRS physicists.

| *Description* | *Number* | *Share* |
|---|---|---|
| Dummy variables | | |
| Gender (1 = female) | 84 | 18 |
| D (started in Grenoble) | 121 | 26 |
| D (started in Paris) | 167 | 36 |
| D (PhD from a *Grande Ecole*) | 79 | 17 |
| Changed labs one or more times | 205 | 44 |

TABLE V Sample statistics for 465 CNRS physicists.

| *Description* | *Median* | *Mean* | *SD* | *Min* | *Max* |
|---|---|---|---|---|---|
| Variables constant over time | | | | | |
| Date of birth | 1945 | 1946.8 | 7.3 | 1936 | 1960 |
| Average lab productivity† | 2.29 | 2.37 | 0.88 | 0.11 | 7.59 |
| Average lab impact factor† | 3.58 | 3.53 | 0.64 | 1.61 | 7.62 |
| International openness – share articles published international | 0.037 | 0.039 | 0.028 | 0.000 | 0.109 |
| No. of researchers in lab | 43 | 46.4 | 26.3 | 0 | 134 |
| No. of labs in career | 1 | 1.61 | 0.79 | 1 | 4 |
| Time-varying variables (5580 observations for 1986–1997) | | | | | |
| Age of researcher this year | 45 | 44.6 | 8.0 | 20 | 61 |
| No. of articles published in year | 2 | 2.69 | 3.21 | 0 | 62 |
| No. of articles weighted by authors | 0.20 | 0.21 | 0.19 | 0 | 1 |
| Average number of pages | 5.40 | 5.49 | 4.68 | 0 | 58 |
| Impact factor (2 years) | 2.54 | 2.66 | 2.30 | 0 | 21.48 |
| Impact factor (5 years) | 4.36 | 4.15 | 3.18 | 0 | 26.56 |

†Based on only 447 observations.

Our panel database is unbalanced both because the scientists enter at different dates, and because some exit before 2002.[6] We restrict the analysis in this paper to a panel analyzed by Turner and Mairesse (2003) containing 465 physicists, observed from 1986 to 1997, aged 26–60 and with 12 years of data. Tables IV and V contain some simple statistics for our data. Of these researchers 18% are female, rising from 15% in the earliest cohort (those born 1936–1940) to over 20% in the last two cohorts (those born 1951–1960). About the same proportion have a doctorate degree from a *Grande Ecole*, of this number about 16% are female.[7] Over half of them (62%) started their career in either Grenoble or Paris, which are considered the most important centers in this field. Almost half of the researchers changed labs at least once during their career. The average number of researchers in the labs in which they worked was 46, and the sample published at a slightly higher rate than their labs (2.7 papers per year versus 2.3 for the average researcher in the lab).

[6] The identification problem is complicated in our setting by the fact that there is a small amount of variation in the identity given above due to entry at different ages (90% of the researchers enter between age 23 and 32), which yields apparent identification, but where such identification is achieved using only a few of the observations. In this article, we abstract from this complication by defining age to be year-less entry cohort rather than calendar age.

[7] The Grande Ecole degree is a high-level pre-doctoral degree. In the French educational system, after graduation from high school, students can either go directly to university, which does not require any grade or level of achievement in high school, or they can apply to a preparatory class where they spend two years studying the material required to compete for the very selective admission into a *Grande Ecole*. Every student of a *Grande Ecole* has therefore been successful in passing a two-phase selection process: selection on the basis of their grades in high school, and on the basis of the *Grandes Ecoles* entrance exams.

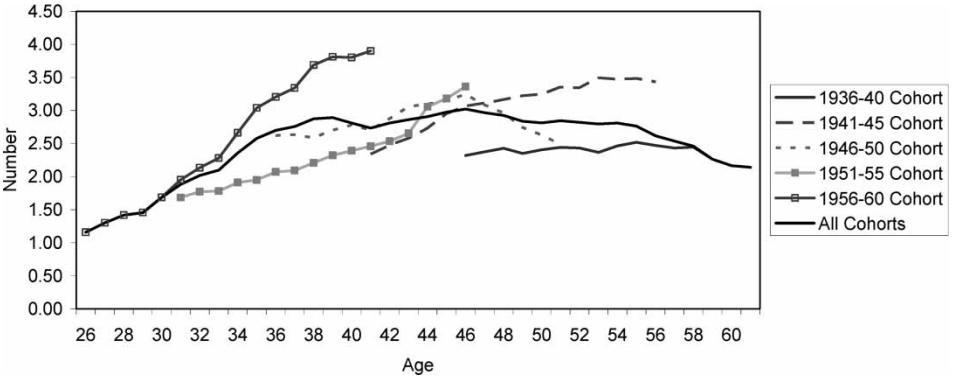


FIGURE 4 Average number of articles published by age (5-year moving average).

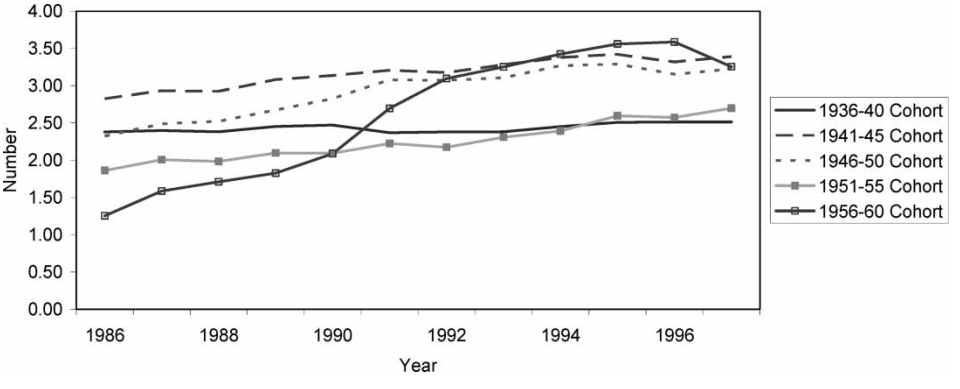


FIGURE 5 Average number of articles published by year (5-year moving average).

As is usual in this literature (Levin and Stephan, 1991), our measure of researcher productivity is the count of articles published during the year.[8] The total number of articles published is about 15,000 (2.7 per person per year), but 25% of the researchers have no publications in a given year, and one individual has 62. Fitting a simple Pareto distribution to these data yields a coefficient of about 0.1, which implies that the distribution from which they are drawn has neither a mean nor a variance.[9] Figures 4 and 5 show the smoothed sample averages of the productivity measure plotted versus age and calendar year, respectively, for five year groupings of the cohorts (year of birth). As expected, the average number of articles published tends to increase over time, although the main differences seem to be by cohort rather than year, with the exception of the most recent cohort. The age distributions for the earlier cohorts suggest a peak somewhere in the late 40s or early 50s, although not very strongly.

[8] We also have several other measures available: articles published weighted by the number of co-authors, the average number of pages in an article, and measures based on citations received in the first two and five years, weighted by the impact factors for the journals in which the citing papers appeared (the average citation rate of its articles). However, in the present article we focus on the article count itself, which is sufficient to illustrate the various identification strategies. See Table V for simple statistics on the other measures.

[9] Obviously, the Pareto properties of the distribution are merely indicative of the level of dispersion in the data. We do not really believe that there is a non-zero probability that an individual researcher will publish an arbitrarily large number of papers per year, so the actual distribution must be bounded, which would imply that the mean and variance exist.

### 4.2 Productivity and Age

In this section of the article, we use the sequence of tests described earlier to ask whether the apparent peak in productivity as a function of age could be due to the confluence of cohort and year effects. In Figures 6(a) and 6(b), we show the results of our tests applied to the actual data on French solid-state physicists. Figure 6(a) considers models with cohort, year, and age effects and Figure 6(b) considers the same models, but this time with individual effects substituted for cohort effects.

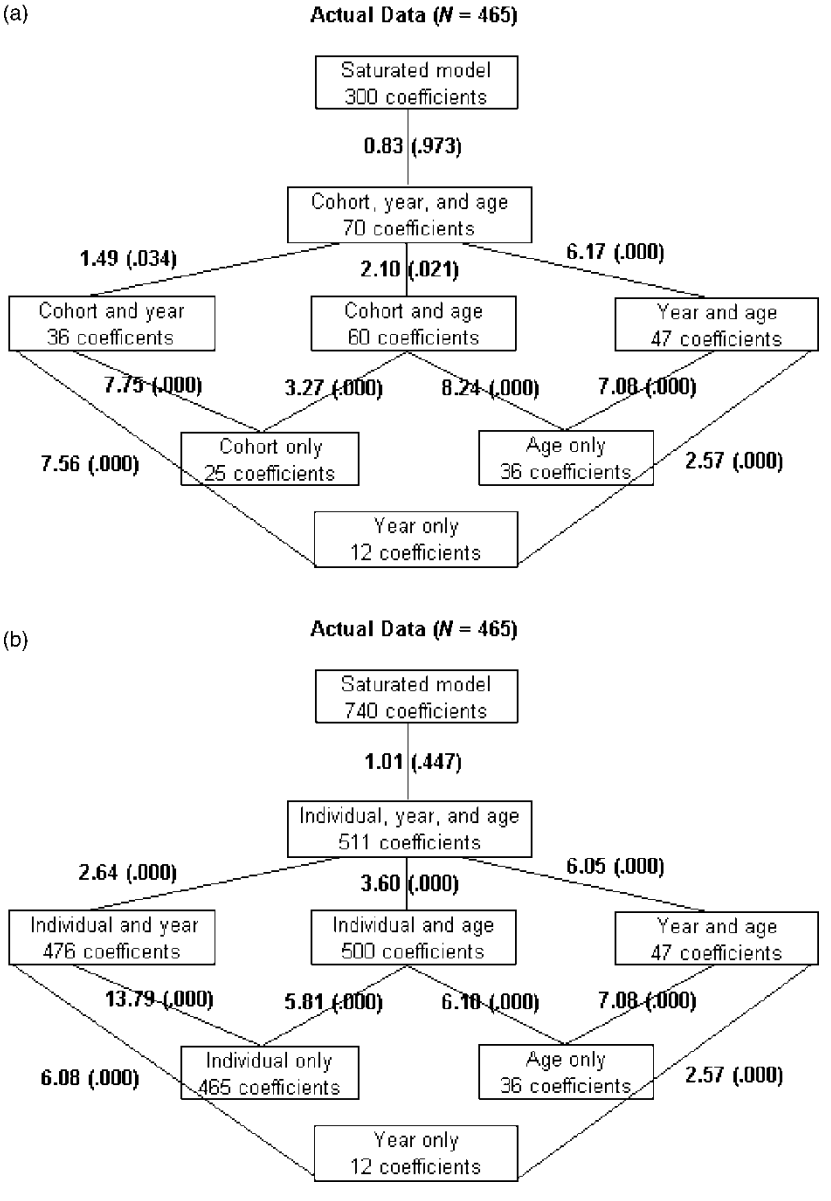


FIGURE 6 (a) $F$-tests for cohort, age, and period models; (b) $F$-tests for age and period models with individual effects actual data ($N = 465$).

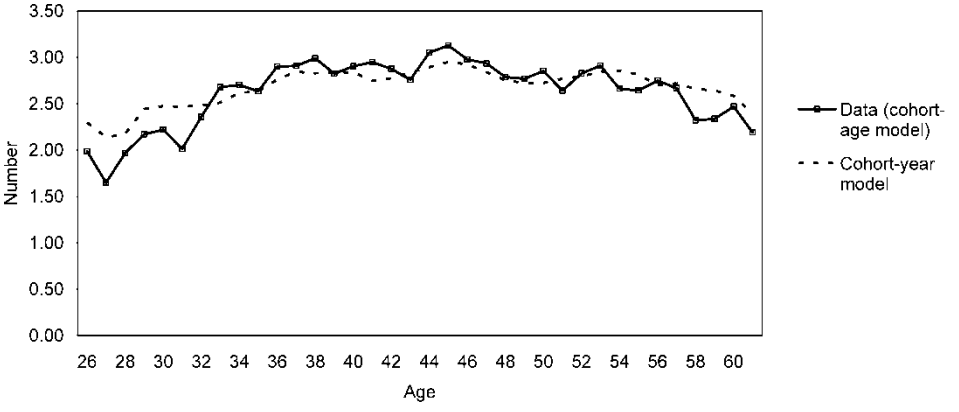


FIGURE 7 Geometric average of number of articles published by age.

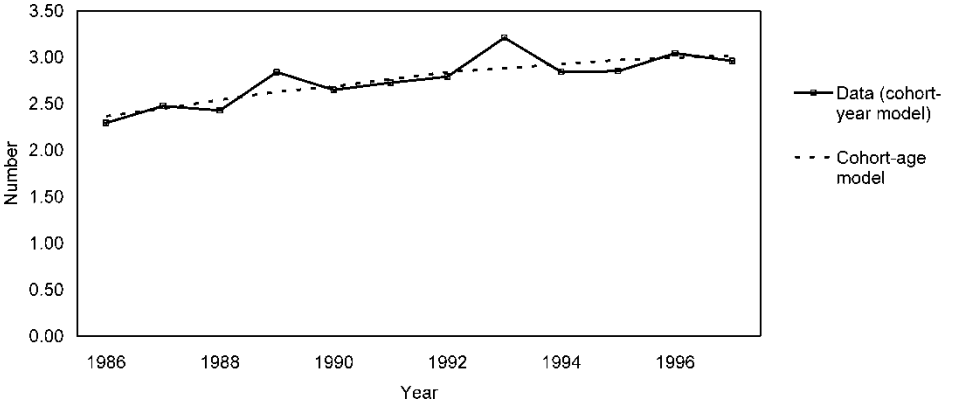


FIGURE 8 Geometric average of number of articles published by year.

The results in Figure 6(a) are similar to those for the simulated data in Figure 3. The preferred specifications with only two sets of dummies are those with cohort and year or cohort and age effects. Although they are both rejected at the 5% level in favor of a specification with all three sets of dummies, using a size adjusted for the fact that the test is nested yields a more equivocal result. Note that the test for a model with only cohort and year effects versus that which includes age effects in addition has a $P$-value of 0.034, which is fairly large given the number of observations (5580) and larger than the adjusted size of 0.025.[10] The conclusion is that the independent effect of researcher age above and beyond that due to the cohort in which he or she entered and the year of publication is at most slight.

Alternatively, if we prefer a specification with cohort and age effects only, that is, a model where calendar time influences only the 'initial condition' for the researcher, such a model would be only marginally less preferred to the cohort–year model. The conclusion is that in order to distinguish these alternatives, it will be necessary to appeal to some prior information, as the data themselves cannot really tell us which is correct.

To underline this point, we show the actual and fitted values from the two models in Figures 7 and 8, plotted first versus age and then versus time. Figure 7 shows the geometric means of the data (publication counts) for each age, and the geometric means of the values predicted

[10] Because we are looking at the combination of two tests here, the correct size is equal to $(1 - 0.95^{0.5}) = 0.025$.

by the cohort–year model (those predicted by the cohort–age model will lie precisely on the actual data).[11] Similarly, Figure 8 shows the same thing by time, with the fitted values from the cohort–age model, since the cohort–year predictions will coincide exactly with the data when it is displayed in this way. Looked at in the age dimension, the cohort–year model appears to miss a bit at the youngest and oldest ages, although it does reproduce the slight peaking. Looked at in the time dimension, the cohort–age model appears to impose an acceptable smoothness on the data. So from this perspective, we might prefer the cohort–age model, even though the fit of the two models is nearly identical ($R^2$ of 0.047 and 0.052, respectively).

Now suppose the research question concerns the age at which publication productivity peaks. In this case, the choice of the model may matter. For example, consider the choice between the three-way model and the cohort–age model, both of which will reproduce the data means when looked at in the cohort–age dimension. Nevertheless, the two models may predict a different productivity peak. A quadratic fit to the two sets of age dummies obtained from these two models using our data yielded the following result: research productivity peaks at 52.2 years of age using the three-way model and at 53.7 years of age using the cohort–age model and ignoring the calendar time effects if they are there. Although this difference is not large, it is significant.[12]

But that is not the end of the problem. Consider the following model which combines a quadratic in age with sets of year and cohort dummies

$$y_{it} = \mu + \alpha_c + \beta_t + \gamma_1 a_{it} + \gamma_2 a_{it}^2 + \varepsilon_{it}. \tag{8}$$

At first glance, this model looks sensible and in fact has often been estimated, sometimes with individual effects rather than cohort effects included (Levin and Stephan, 1991; Turner and Mairesse, 2005). Identification (with an intercept included) requires omission of both of one of the cohort dummies and of one of the year dummies. However, because age ($a_{it}$) is an exact linear function of cohort and period, which identifying assumption you choose (and there are potentially a large number) will affect the estimates of $\gamma_1$ and $\gamma_2$, and therefore, the estimate of the age at which productivity peaks (which is $-\gamma_1/2\gamma_2$).

Figure 9 shows a representative result for our data. The identifying assumptions used were to include a complete set of year dummies, exclude the intercept, and include all but one of the cohort dummies. The excluded cohort dummy was allowed to vary from 1936 to 1960. The figure shows a few representative examples of the resulting age profile (excluding year and cohort effects). Note that all the fits were identical, in the sense that the sum of squared residuals were exactly equal, and they all generated the same age–cohort–year means, but very different age–productivity profiles. The problem is interpretive: the age–cohort–year identity means that it is impossible to identify the productivity curve as a function of age without strong prior restrictions on the year and cohort effects (such as their absence). The age at which productivity peaks also varies significantly for the different normalizations: it is 39.6, 41.3, 42.4, 37.9, and 50.4 when we drop the dummy for entry in 1936, 1940, 1944, 1948, and 1952, respectively.

The situation is even worse for the model with individual effects in place of the cohort effects, along with the year and age effects. For this model, we obtained identification by setting the coefficients of two of the adjacent years equal to each other, again varying the choice of years for which we did this from 1986/87 to 1996/97. In this case, the age at which productivity peaked varied all the way from 0 to 100, of course with large standard errors. Figure 6(b)

[11] Geometric means are used because the model was fit in log-linear form, so these are the unbiased predictions (but without the correction for the residual variance, which is small).

[12] If a quadratic model in age is included directly in the model with cohort–year dummies and that with cohort dummies alone, the difference in peak age is even larger: 50.6 years versus 53.8 years.

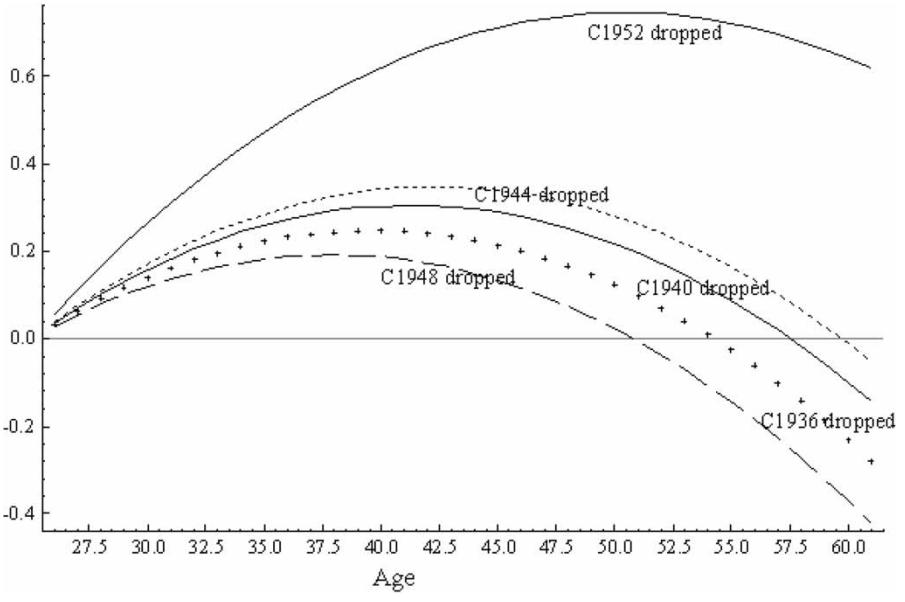


FIGURE 9 Age quadratic estimated with alternative identifying assumptions for cohort effect.

shows the results of conducting our testing methodology on the model with individual, year, and age effects. In contrast to the models with cohort effects, the only model that is accepted is the three-way model. The implication of the results is that we need to include individual, year, and age effects in our model, but that we cannot use the results to infer the age productivity peak, since it is so sensitive to the choice of normalization.

## 5 CONCLUSIONS

This paper has explored a familiar identification problem, that of vintage, age, and time, in a context where it does not seem to have been sufficiently recognized: the identification of cohort, age, and period effects in scientific productivity. We have emphasized the fact that identifying an age-related productivity effect or the presence and location of a productivity peak relies crucially on what we are willing to assume about the variation in the other two dimensions, cohort and time. There is no universal solution to this problem, given the identity that relates the three.

Therefore, we recommend the following: test for the presence of each of the three effects semi-parametrically as we have done in this article. If the tests reveal that one dimension can be ignored, then the most parsimonious specification will include only the other two dimensions. However, the power of such a test clearly goes up with the dimensions of the data: in some unreported experiments, we found that 12 years and 36 ages led to confusion between a cohort–age and a cohort–year model when the former was the true model, whereas 25 years and 44 ages allowed us to distinguish the two.

Alternatively, we return to the original recommendations of Rodgers, who strongly advocated the use of *a priori* information about cohorts or the time periods to help identify the model. We note that this approach was the one taken by Stephan and Levin (1991) when they achieved identification by grouping the cohorts in their sample according to the knowledge base

to which they were exposed in their graduate training.[13] One solution sometimes proposed, grouping cohorts in multi-year intervals seems somewhat less satisfactory in this context. This amounts to achieving identification of the age effect by comparing closely adjacent ages and assuming that they come from the same cohort. In this case, it would seem preferable to use the actual variation in year of entry into the sample (cohort) for individuals of the same age, rather than creating spurious age variation by holding the cohort fixed.

We conclude with a discussion of the impact of the identification problem discussed here on the coefficient estimates for other variables that may be included in the regression. For example, a researcher may be interested in gender differences in scientific productivity, or in the impacts of productivity on the part of other researchers in the lab. Although, clearly these coefficients will be affected by the choice of model when the variables of interest are correlated with cohort, age, or period. In this case, the choice of assumption used to identify the three-way model will not matter for the coefficients of interest. As long as the assumptions used are equivalent in the sense of yielding the same (identical) fit of the model, the estimated coefficients on the variables of interest will be the same. This fact suggests that the safest procedure when the variables of interest are other than age may indeed be to use the saturated or three-way model to estimate scientific productivity, in order to provide maximum control for unknown cohort, year, and age effects. This is the good news. The bad news is that it appears impossible to estimate age productivity effects without strong *a priori* assumptions on the rest of the model.

## *Acknowledgement*

This is a revision of a paper prepared for the SPRU conference in memory of Keith Pavitt at the University of Sussex, 13–15 November, 2003. We are extremely grateful to Serge Bauin and Michele Crance from UNIPS-CNRS, France for their invaluable help in constructing the database of condensed-matter physicists, and to two anonymous referees for helpful comments. We have also benefitted from the work of Ganaes Bascouly, Julia Grenet, Héléne Huber and Lionel Janin in the Panel Data seminar at CREST-ENSAE.

## *References*

Berndt, E.R., Griliches, Z. and Rappaport, N. (1995) Econometric Estimates of Prices Indexes for Personal Computers in the 1990s. *Journal of Econometrics*, **68**, 243–268.

Berndt, E.R. and Griliches, Z. (1991) Price Indices for Microcomputers: An Exploratory Study. In *Price Measurements and their Uses*. Chicago: University of Chicago Press, pp. 63–93.

Bonaccorsi, A. and Daraio, C. (2003) Age Effects in Scientific Productivity: The case of the Italian National Research Council (CNR). *Scientometrics*, **58**(1), 49–90.

Cole, S. (1979) Age and Scientific Performance. *American Journal of Sociology*, **84**(4), 958–977.

Cole, J. and Zuckerman, H. (1984) The Productivity Puzzle: Persistence and Change in Patterns of Publications of Men and Women Scientists, in *Advances in Motivation and Achievement*, Vol. **2**, pp. 217–258.

Cole, J. and Singer, B. (1991) A Theory of Limited Differences: Explaining the Productivity Puzzle in Science. In Zukerman, H., Cole, J. and Bruer, J. (eds.) *The Outer Circle: Women in the Scientific Community*. New York: Norton.

David, P. (1994) Positive Feedbacks and Research Productivity in Science: Reopening Another Black Box. In Granstrand, O., (ed.) *The Economics of Technology*, Amsterdam: Elsevier Science, 65–89.

Diamond, A. (1984) An Economic Model of the Life-Cycle Research Productivity of Scientists. *Scientometrics*, **6**(3), 189–196.

Hall, R.E. (1971) The Measurement of Quality Change from Vintage Price Data, Chapter 8. In Griliches, Z. (ed.) *Price Indexes and Quality Change*, Cambridge, MA: Harvard University Press, 240–271.

[13] However, as we have shown here, and as is also clear from their detailed discussion of the tables in Levin and Stephan, this method of identification breaks down if individual rather than cohort effects are included.

Heckman, J. and Robb, R. (1985) Using Longitudinal Data to Estimate Age, Period, and Cohort Effects in Earnings Equations. In Mason, W. and Fienberg, S. (eds.) *Cohort Analysis in Social Research: Beyond the Identification Problem*. New York: Springer Verlag.

Joreskog, K.G. (2005) The LISREL Program. Available online at: http://www.ssicentral.com/lisrel/

Kennedy, P. (1996) *A Guide to Econometrics*. Oxford, UK: Blackwell Publishers, Ltd., pp. 75–93.

Levin, S. and Stephan, P. (1991) Research Productivity Over the Life Cycle: Evidence For Academic Scientists. *American Economic Review*, **81**(1), 114–132.

Mason, K.O., Mason, W.M., Winsborough, H.H. and Poole, W.K. (1973) Some Methodological Issues in Cohort Analysis of Archival Data. *American Sociological Review*, **38**(2), 242–258.

Mason, W. and Fienberg, S. (eds.) (1985) *Cohort Analysis in Social Research: Beyond the Identification Problem*. New York: Springer Verlag.

Rodgers, W.L. (1982a) Estimable Functions of Age, Period, and Cohort Effects. *American Sociological Review*, **47**(6), 774–787.

Rodgers, W.L. (1982b) Reply to Comment by Smith, Mason and Fienberg. *American Sociological Review*, **47**(6), 787–793.

Smith, H.L., Mason, W.M. and Fienberg, S.E. (1982) Estimable Functions of Age, Period, and Cohort Effects: More Chimeras of the Age-Period-Cohort Accounting Framework: Comment on Rodgers. *American Sociological Review*, **47**(6), 787–793.

Stephan, P.E. (1996) The Economics of Science. *Journal of Economic Literature*, **XXXIV**, 1199–1235.

Stephan, P.E. (1998) Gender Differences in the Rewards to Publishing in Academic Science in the 1970's. *Sex Roles*, **38**(11/12).

Turner, L. and Mairesse, J. (2003) Individual Productivity Differences in Scientific Research: An Econometric Study of the Publication of French Physicists. Paper presented at the Zvi Griliches Memorial Conference, Paris, August 2003.

Turner, L. (2003) La recherche publique dans la production de connaissances: contributions en économie de la science. PhD Dissertation, Université Paris 1. Available Online at: (http://www.crest.fr/pageperso/laure.turner/these.htm).