Supplementary Material to: 'The effect of World War I on naming patterns : a systematic exploration'

1 Aggregate effect of increased paternal name transmission on popularity changes

Previous work has found using microdata that transmission of the father's name increased significantly during the first year of the War (Todd & Coulmont 2021). While about 12% of baby boys were given their father's name at the eve of WW1, this proportion moved to about 18% at the beginning of the War, before returning to its prewar level around May 1915. Since by definition first names that were fashionable about 30 years before the War were frequently chosen to name fathers, we expect a systematic negative correlation between growth before the War and excess expression during the War. In other words, the popularity of old-fashioned first names was temporarily revived simply because they were the names of soldiers sent to the front. Simple assumptions help delineate the specific relationship we expect and test the quantitative importance of this phenomenon[1].

We assume here that increased paternal name transmission is the only change brought about by the War and we focus on males (in females, paternal name transmission is most often achieved using variants, e.g. Paulette for Paul).

The notations are :

- f_0: baseline (prewar) paternal name transmission probability
- f: war-induced increase in probability of paternal name transmission[2]
- p_j: probability of giving name j when the father is not named j and when the father did not transmit his name: $p_j = P(X = j|Z \neq j, X \neq Z)$, with X the son's name and Z the father's name
- C_j: counterfactual number of babies born during the War ('War babies') named j (babies who would

[1] In what follows, paternal name transmission is more broadly defined as the transmission of the first name of someone who belongs to the same generation as the father (father, uncle, etc.).

[2] For instance, a transmission probability moving from 12% to 18% corresponds to f_0=0.12 and f=0.18-0.12=0.06.

have been named j had the War not taken place)

- O_j: observed number of War babies named j
- N: total number of War babies: $N = \sum_j C_j = \sum_j O_j$
- M_j: number of War babies' fathers named j
- T: mean age at reproduction
- r_j: mean growth rate of first name j in the T years preceding WW1, so that $C_j = e^{r_j T} M_j$. [3]

One can write: $C_j = f_0 M_j + p_j(1 - f_0)(N - M_j)$. In words, two subpopulations are given first name j:

- those who inherit first name j from their father (*transmission*)
- the fraction of the $N - M_j$ babies whose father is not named j who are given first name j (*taste*).

Parameter f_0 controls the balance between transmission and taste, and the value for p_j reflects whether fashion favors name j.

Changing f_0 for $f_0 + f$, one can similarly write: $O_j = (f_0 + f)M_j + p_j(1 - f_0 - f)(N - M_j)$.

Then $O_j = C_j + fM_j - fp_j(N - M_j)$.

Using the fact that $p_j(N - M_j) = \frac{C_j - f_0 M_j}{1 - f_0}$ yields:

$$\begin{aligned} O_j &= C_j + fM_j - f\frac{C_j - f_0 M_j}{1 - f_0} \\ &= C_j(1 - \frac{f}{1 - f_0}) + fM_j(1 + \frac{f_0}{1 - f_0}) \\ &= C_j(1 - \frac{f}{1 - f_0}) + \frac{f}{1 - f_0} M_j \\ &= C_j + \frac{f}{1 - f_0}(M_j - C_j) \\ &= C_j + \frac{f}{1 - f_0}(e^{-r_j T} - 1)C_j \end{aligned}$$

Thus, $FC_j = 1 + \frac{f}{1-f_0}(e^{-r_j T} - 1)$ where FC_j $(= O_j/C_j)$ is the fold-change between counterfactual and observed number of babies named j: FC_j, that measures the increase or decrease in popularity of name j due to the war, is a monotonically decreasing function of r_j, and $FC_j > 1$ if and only if $r_j < 0$: excess paternal name transmission creates correlation between increasing popularity during the War and having been in decline in the thirty years or so that separate the fathers' birth from the War.

A rapid investigation indeed reveals that a relationship empirically exists between pre-war growth and

[3] The instantaneous growth rate for name j in year y (net of changes in N_y – see Main Text) can be estimated directly from the model as $\frac{ds_j}{dy}$. We are rather interested in the *mean* growth rate between the fathers and children births. Since the data on which the model is built start only in 1900, we thus compute $(s_j(1916) - s_j(1900))/(1916 - 1900)$ as an estimate of r_j.

wartime overexpression. Removing first names given to fewer than 1,000 babies over the entire 1900-1930 period (restricting our quick analysis to first names where the fold-change and growth rate are estimated with greater accuracy) and those with an FC>1.5 (those clearly influenced by factors other than paternal name transmission) leaves 235 names. A robust linear regression of FC_j on $\exp(-r_j T)$ (assuming $T = 33y$) finds a strong positive association between the two:

Term	Estimate	Std. error	t-value	p-value
Intercept	0.952	0.010	93.2	<0.0001
Slope	0.039	0.004	9.2	<0.0001

2 Total Aggregate Effect of the War on first name choices

2.1 Notations

Using standard potential outcomes notation, we write $Y_i(0)$ the first name of child i had WW1 not taken place. Similarly, let $Y_i(1)$ be the first name actually given to child i.

For each first name j, define A_j the number of children who were given name j *because of the War* and R_j the number of children who were *not* given name j because of the War:

$$\begin{cases} A_j = \sum_i \mathbb{1}\{Y_i(0) \neq j,\ Y_i(1) = j\} \\ R_j = \sum_i \mathbb{1}\{Y_i(0) = j,\ Y_i(1) \neq j\} \end{cases}$$

We observe $O_j = \sum_i \mathbb{1}\{Y_i(1) = j\}$ in the data and use our statistical model (see Main Text) to give an estimate for $C_j = \sum_i \mathbb{1}\{Y_i(0) = j\}$. Let $D_j = O_j - C_j$ be the aggregate additional number of children named j because of the War:

$$\begin{aligned} D_j &= \sum_i \mathbb{1}\{Y_i(1) = j\} - \sum_i \mathbb{1}\{Y(0) = j\} \\ &= A_j - R_j \end{aligned}$$

If $D_j < 0$, more children would have been given name j had the War not happened.

2.2 Relationship between the TAE and the TTE

Our **Total Aggregate Effect (*TAE*)** is defined as $\frac{1}{N}\sum_{j,D_j>0} D_j$ (see Main Text) and can be estimated using observed quantities. This is also equal to $\frac{1}{N}\sum_j D_j^+$, or $\frac{1}{N}\sum_j \frac{|D_j|+D_j}{2}$.[4]

We similarly define here a **Total True Effect (*TTE*)** as the proportion of children whose first name changed because of the War: $TTE = \frac{1}{N}\sum_i \mathbb{1}\{Y_i(0) \neq Y_i(1)\}$. The TTE is the most comprehensive summary measure for the effect of the War on naming behaviors, as an answer to the question: what fraction of babies born during the War was given a different name because of the War? But computing the TTE requires unobserved individual names had the War not happened: the TTE is therefore unknown. What is then the connection between the TTE and the TAE?

To simplify the notations, we write N_{chg} the total *number* of children whose name changed: $N_{\text{chg}} = \sum_i \mathbb{1}\{Y_i(0) \neq Y_i(1)\}$, so that $TTE = \frac{N_{\text{chg}}}{N}$.

We note that $N_{\text{chg}} = \sum_j A_j = \sum_j R_j$, since an individual whose first name changed because of the War is counted exactly once in $\sum_j A_j$ (for his new first name) and once in $\sum_j R_j$ (for the first name he would have had). We thus have $\sum_j D_j = 0$, so that:

$$\begin{aligned}
TAE &= \frac{1}{2N}\sum_j |D_j| = \frac{1}{2N}\Big(\sum_{j,D_j>0} D_j - \sum_{j,D_j<0} D_j\Big) \\
&= \frac{1}{2N}\Big(N_{\text{chg}} - 2\sum_{j,D_j>0} R_j + N_{\text{chg}} - 2\sum_{j,D_j<0} A_j\Big) \\
&= \frac{1}{N}\Big(N_{\text{chg}} - \sum_{j,D_j>0} R_j - \sum_{j,D_j<0} A_j\Big) \\
&= TTE - \frac{1}{N}\Big(\sum_{j,D_j>0} R_j + \sum_{j,D_j<0} A_j\Big) \leq TTE
\end{aligned}$$

We arrive at the intuitevely obvious result that the TAE is equal to the TTE only when $\sum_{j,\ D_j>0} R_j = 0$ (first names with a net popularity gain because of the War didn't "loose" any children because of the War) and $\sum_{j,\ D_j<0} A_j = 0$ (first names with a net popularity loss because of the War didn't "gain" any children). Thus the TAE and TTE are equal if and only if no compensation occurs.

Illustrating the opposite situation is easy. Imagine that only two first names (say, Jean and Emile) would have been given in similar proportions had the War not taken place. Suppose now that for some reason related to the War all children who would have be named Jean are named Emile, and that conversely all children who would have be named Emile are named Jean. Then $TTE = 1$ but $TAE = 0$: aggregate data

[4] Where x^+ is the positive part of x: $x^+ = \max(x, 0)$.

completely fail to capture this overwhelming effect of the War on individual choices. More realistically, compensation may occur in a context of increasing polarization around controversial figures such as, in the case of WW1 France, President Raymond Poincaré. One may speculate that some parents named their son Raymond because of Poincaré's role, while conversely others abandoned this first name for the same reason. The net effect on the TAE would depend on the size of these two subpopulations, but in any case a fraction of war-induced name changes would be missed by the TAE.

Another mechanism leading to compensation during WW1, this time clearly documented, was excess paternal name transmission, that we have already mentioned was quantitatively important at least at the beginning of the conflict.

2.3 Magnitude of the $TTE - TAE$ discrepancy with excess paternal name transmission only

Going back to the idealized situation where excess paternal name transmission is the only mechanism changing naming preferences during WW1, we can further explore the discrepancy between what we can measure – the TAE – and what we would like to measure – the TTE. In this idealized situation, we have:

$$\begin{cases} A_j = fM_j = fe^{-r_j T}C_j \\ R_j = fp_j(N - M_j) = f\frac{C_j - f_0 M_j}{1-f_0} = fC_j\frac{1-f_0 e^{-r_j T}}{1-f_0} \end{cases}$$

Rewriting the general relationship between the TTE and the TAE as:

$$TTE - TAE = \frac{1}{N}(\sum_{j,\ D_j>0} R_j + \sum_{j,\ D_j<0} A_j) = \frac{1}{N}\sum_j \min(R_j, A_j) \quad [\text{since}\min(R_j, A_j) = R_j \ \Leftrightarrow D_j > 0]$$

and plugging in the specific expressions for A_j and R_j, we obtain:

$$\begin{aligned} TTE - TAE &= \frac{1}{N}\sum_j \min(fC_j, fC_j\frac{1-f_0e^{-r_jT}}{1-f_0}) \\ &= \frac{f}{N}\sum_j C_j \min(1, \frac{1-f_0e^{-r_jT}}{1-f_0}) \leq \frac{f}{N}\sum_j C_j = f \end{aligned}$$

The difference between the TTE and the TAE hence ultimately depends on how heterogeneous first names are in terms of rate of growth. In the extreme case where all first names are stable ($r_j = 0$), $TTE - TAE$ is simply equal to f, the excess paternal name transmission rate.

Another telling measure of the discrepancy between the TAE and the TTE is simply $\frac{TTE}{TAE}$. Assuming $f_0 = 12\%$ and $f = 3$ % (and again that $T = 33$ years), the TTE in males is estimated to be 1.5 the TAE. When $f = 6\%$[5], the TTE/TAE ratio reaches 2.1: in this simple, realistic, situation, half of the true effect of the War on naming behavior is missed when focusing on the TAE.

These results suggest that even in the simple case where excess paternal name transmission is the only compensation mechanism at play and is moderately frequent, the TAE significantly underestimates the TTE.

2.4 Breaking down the TAE by sub-periods

What is in general the relationship between the TAE over the entire period of interest and the TAE by sub-periods (for example, by year) ? Considering only two sub-periods 1 and 2 to keep notations simple, and extending those already in use (e.g. considering $C_{j,p}$ the counterfactual numbers babies named j in subperiod $p \in \{1, 2\}$):

$$\begin{aligned} N \times TAE &= \sum_j (O_j - C_j)^+ \\ &= \sum_j (O_{j,1} - C_{j,1} + O_{j,2} - C_{j,2})^+ \\ &= \sum_j (D_{j,1} + D_{j,2})^+ \leq \sum_j (D_{j,1}^+ + D_{j,2}^+) = N_1 \times TAE_1 + N_2 \times TAE_2 \end{aligned}$$

When the effect of the war on first name j is positive on subperiod 1 and negative on subperiod 2 (or the opposite),[6] $D_{j,1}$ and $D_{j,2}$ have opposite signs and $(D_{j,1} + D_{j,2})^+ < D_{j,1}^+ + D_{j,2}^+$. The above inequality is then strict: $N \times TAE < N_1 \times TAE_1 + N_2 \times TAE_2$. This mechanism is indeed observed when breaking down the TAE by year: the weighted average of yearly TAEs is greater than the TAE over the entire period (see Main Text). Since $N \times TTE = N_{chg} = N_{chg\ 1} + N_{chg\ 2} = N_1 \times TTE_1 + N_2 \times TTE_2 \geq N_1 \times TAE_1 + N_2 \times TAE_2$, $\frac{1}{N}(N_1 TAE_1 + N_2 TAE_2)$ is also a lower bound on TTE, and a better one than the overall TAE.

[5] 6% being the excess father-to-son transmission rate of 1914 and early 1915, it is a lower bound for the value of f, the excess paternal name transmission in the broad sense (see note 1), for this period. This analysis is therefore conservative.

[6] This of course cannot happen when no compensation between first names occurs.