

## Additional material of the paper:

*“Does early tracking affect learning inequalities? Revisiting difference-in-difference modeling strategies with international assessments”*

### Extension to other educational institutions

Early tracking is by far the institution that has received the greatest attention in terms of its potential impact on achievement inequalities. The role played by other features of the education systems has been analyzed in a similar perspective, either with single international assessments, or by exploiting two assessments with difference-in-difference strategies. As mentioned above, Ammermueller (2013) uses an extended version of model M2 to analyze the effect on inequalities of the strength of the private sector, the degree of autonomy and time devoted to instruction.

Indeed, the case of tracking is ideal to be analyzed with a difference-in-difference design, because in 4<sup>th</sup> grade children are still in comprehensive school in all countries, whereas in at age 15 tracking has already been enforced in some countries but not in others, allowing studying how inequality evolves in early tracking countries relative to late tracking countries. Nonetheless, models M1-M3 can be adapted to study the effect of other institutional features described by non-binary variables and varying over time and across countries with different patterns. The discussion on the validity of the difference-in-difference individual pooled models developed in the following sections also applies to them.

#### Individual pooled models

Ammermueller (2013) analyzes the effect of tracking and of other educational institutions (share of the private sector, autonomy, instruction time) on family background inequalities. Identification of institutional effects rests on the variability among countries and between assessments of these institutional variables  $I$ . In the general case, the model can be expressed as:

$$Y_{itc} = \alpha_{0tc} + \varphi X_{itc} + \xi_{1c} F_{itc} + \xi_2 F_{itc} T + \lambda_2 F_{itc} T (I_{2c} - I_{1c}) + \varepsilon_{itc}$$

The institution of interest  $I$  here is not necessarily described by a binary variable (consider for example the share of private sector, that can take any value from 0% to 100%), and may vary across stages of schooling in different ways.

Family background coefficients in primary school are unconstrained and estimated as fixed effects (hence  $\gamma_{1c} = \xi_{1c}$ ). Instead, the variation between  $t=1$  and  $t=2$  depends only on institutional changes. Coefficients at  $t=2$  are  $\gamma_{2c} = (\xi_{1c} + \xi_2)$  when  $I_{2c} = I_{1c}$  (i.e. when there are no changes between  $t=1$  and  $t=2$ ) and  $\gamma_{2c} = \xi_{1c} + \xi_2 + \lambda_2(I_{2c} - I_{1c})$  with institutional changes.

The underlying assumptions are that unobserved country characteristics are allowed to affect family background inequalities at  $t=1$ ; instead, the *difference* in family background inequalities between  $t=1$  and  $t=2$  may be influenced only institutional changes occurring in the meantime.

The parameter of main interest is  $\lambda_2$ , corresponding to:

$$(E(\gamma_2|I_2 - I_1 = i + 1) - E(\gamma_2|I_2 - I_1 = i)) - (E(\gamma_1|I_2 - I_1 = i + 1) - E(\gamma_1|I_2 - I_1 = i))$$

representing the effect of a one-unit institutional change on family background inequalities.

### Two-step models

The coefficients of the *individual level model* of test scores  $Y$  are allowed to vary freely across countries and across assessments held at different stages of schooling:

$$Y_{itc} = \alpha_{0tc} + \gamma_{tc}F_{itc} + \varphi_{tc}X_{itc} + \varepsilon_{itc}$$

The (random) regression coefficients of family background at the two assessments depend on institutional characteristics and may be related by a simple *country-level* linear model:

$$\gamma_{2c} = a + b\gamma_{1c} + d(I_{2c} - I_{1c}) + u_c$$

where  $(I_{2c} - I_{1c})$  capture institutional changes occurring between  $t=1$  and  $t=2$  and  $u$  captures country-level unobserved factors affecting inequalities developing in the same time span, assumed to be uncorrelated to institutional changes. The effect of tracking is  $d$ , the average difference in the family-background coefficient at  $t=2$  that can be ascribed to an additional one-unit change in the institutional variable, *given* the coefficient at  $t=1$ :

$$DID = E(\gamma_2|\gamma_1, I_2 - I_1 = i + 1) - E(\gamma_2|\gamma_1, I_2 - I_1 = i)$$

The identifying assumption is that inequality changes between  $t=1$  and  $t=2$  only depend on the institutional variable of interest or on other system-level features not correlated to it.

## DID in pooled individual models: a simulation exercise

To fix ideas, we propose an example showing what is actually estimated with individual pooled models M1 and M2, when test scores are equated and when they are not equated. To illustrate the issue, we impose a rather large difference between parameters in tracked and untracked systems (yet, results are qualitatively the same when using other sets of parameters).

We simulate data for 30 countries, half of which are tracked and the other half untracked. There are two assessments. Each country has 3000 observations per assessment. We first simulate test scores according to equations (6)-(8) in the main manuscript:

$$y_{i1} = \mu_1 + \rho F_i + \varepsilon_{i1} \quad (6)$$

$$y_{i2} = y_{i1} + \delta_i \quad (7)$$

$$\delta_i = \Delta + \beta F_i + \theta y_{i1} + \varepsilon_{i2} \quad (8)$$

as if achievement was expressed in the same metrics at the two assessments, and then transform scores at  $t=1$  into a different metrics:

$$y_{i1} = \varphi + \omega y'_{i1} \quad (10)$$

with scaling factor  $\omega = 0.5$  and  $\varphi = 0$ .

We let  $\rho$  (representing inequalities at  $t=1$ ) vary across countries according to a uniform distribution:  $U(40,60)$  in tracked countries (average=50),  $U(20,40)$  in untracked countries (average=30). Instead,  $\beta$  and  $\theta$  (representing inequalities developing between  $t=1$  and at  $t=2$ ) are fixed within tracking regime. The parameters and the corresponding  $\gamma$  (family background regression coefficients) are described in the upper panel of Table A1. We distinguish between the latent  $\gamma_1$ , – computed as if test scores at  $t=1$  and  $t=2$  were equated, hence expressed on the same scale of test scores at  $t=2$  – and the observed  $\gamma_1$  with non-equated tests (see Section 3).

In the lower panel of Table A1 we compute *DID* according to equation (5) in the manuscript:

$$DID = \left( E(\gamma_2 | I = 1) - E(\gamma_2 | I = 0) \right) - \left( E(\gamma_1 | I = 1) - E(\gamma_1 | I = 0) \right) \quad (5)$$

The true *DID* (based on equated test scores) is 23.5, whereas the observed *DID* (resulting from the differentiation of regression coefficients in the different metrics) is 3.5.

**Table A1. Model parameters and DID with equated and observed scores**

|                       | Tracked | Untracked |
|-----------------------|---------|-----------|
| $\rho$                | 50      | 30        |
| $\beta$               | 40      | 20        |
| $\theta$              | 0.1     | 0.05      |
| $\gamma_1$ (equated)  | 50      | 30        |
| $\gamma_1$ (observed) | 100     | 60        |
| $\gamma_2$            | 95      | 51.5      |

$$\omega = 0.5; \gamma_1 \text{ (equated)} = \rho; \gamma_1 \text{ (observed)} = \rho/\omega; \gamma_2 = \beta + (1 + \theta)\rho$$

|                                   |      |
|-----------------------------------|------|
| <i>DID (equated) – true</i>       | 23.5 |
| <i>DID (observed) – estimated</i> | 3.5  |

$$DID \text{ (equated)} = (95-51.5)-(50-30)=23.5; DID \text{ (observed)} = (95-51.5)-(100-60)=3.5$$

In Table A2 we report the results of 10 simulations: after random generation of the data for each country and assessment according to the parameters described in Table A1, we estimate models M1 and M2 with two different sets of data for  $t=1$ : (i) with (latent) equated scores; (ii) with observed scores. It is clear that, for both models, the analyses with equated scores deliver results in line with the ‘true *DID*’ value and the analyses with observed scores are in line with the computed ‘observed *DID*’, hence, are heavily biased.

**Table A2. Results of 10 replications of the simulation: DID with equated and observed scores**

| Replication | M1             |                 | M2             |                 |
|-------------|----------------|-----------------|----------------|-----------------|
|             | Equated scores | Observed scores | Equated scores | Observed scores |
| 1           | 23.78          | 2.88            | 24.31          | 3.38            |
| 2           | 24.08          | 7.35            | 24.60          | 7.87            |
| 3           | 26.28          | 3.42            | 25.88          | 3.04            |
| 4           | 21.00          | 5.70            | 20.24          | 4.46            |
| 5           | 20.63          | 0.93            | 20.59          | 1.22            |
| 6           | 22.30          | 2.99            | 21.88          | 2.65            |
| 7           | 20.21          | 2.29            | 20.65          | 2.86            |
| 8           | 20.16          | -1.71           | 20.23          | -1.80           |
| 9           | 24.67          | 3.07            | 24.37          | 2.47            |
| 10          | 23.99          | 7.70            | 24.65          | 8.10            |