## Appendix B

The objective is to estimate the effect of moving from the training firm on wages three years after the end of training. The data available is from three graduating cohorts of apprentices who each are observed three years. Omitting the cohort subscript let $D_{ij0}$ be the mover status at the end of training of worker $i$ leaving training firm $j$; let $w_{ijt}$ be log real wages; let $x_{ij0}$ denote observable characteristics of workers, training, and training firm; and let $\phi_{i0}^{j}$ and $\phi_{i0}^{c}$ be training firm and cohort dummies, respectively. The aim is to estimate the effect of moving on wages in the regressions

$$y_{ijt} = D_{ij0}\delta_t + x_{ij0}\beta_t + \sum_{j=1}^{J} \gamma_j \phi_{i0}^{j} + \sum_{c=1}^{3} \gamma_{ct} \phi_{i0}^{c} + \varepsilon_{ijt}, \quad t = 1,3,5,$$

with or without the training firm fixed effects (where $J$ denotes the number of firms). The main parameters of interest are $\delta_1, \delta_3, \delta_5$. To estimate these, the observations are stacked into a panel

$$\begin{bmatrix} w_1 \\ w_3 \\ w_5 \end{bmatrix} = \begin{bmatrix} D_1 \\ 0 \\ 0 \end{bmatrix}\delta_1 + \begin{bmatrix} 0 \\ D_3 \\ 0 \end{bmatrix}\delta_3 + \begin{bmatrix} 0 \\ 0 \\ D_5 \end{bmatrix}\delta_5 + \begin{bmatrix} x_1 \\ 0 \\ 0 \end{bmatrix}\beta_1 + \begin{bmatrix} 0 \\ x_3 \\ 0 \end{bmatrix}\beta_3 + \begin{bmatrix} 0 \\ 0 \\ x_5 \end{bmatrix}\beta_5 + \begin{bmatrix} \Phi_1^j \\ \Phi_3^j \\ \Phi_5^j \end{bmatrix}\gamma + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_3 \\ \varepsilon_5 \end{bmatrix},$$

where $x$ now all control variables with time varying effects. In matrix format this equation can be rewritten as

$$W = \Lambda_1 D\delta_1 + \Lambda_2 D\delta_2 + \Lambda_3 D\delta_3 + \Lambda_1 X\beta_1 + \Lambda_2 X\beta_2 + \Lambda_3 X\beta_3 + \Phi^j\gamma + \vec{\varepsilon}, \qquad \text{(B1)}$$

where $W = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$, $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$, $D = \begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix}$, $\Phi^j = \begin{bmatrix} \Phi_1^j \\ \Phi_2^j \\ \Phi_3^j \end{bmatrix}$, $\vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$, and for ease of notation I have defined $\Lambda_1 = \begin{bmatrix} I_{N_1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $\Lambda_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & I_{N_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $\Lambda_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_{N_3} \end{bmatrix}$. The dimension of each vector $\xi_t$ is $N_t \times 1$, such that the full vectors $\Xi = [\xi_1', \xi_2', \xi_3']'$ are $N \times 1$ with $N = N_1 + N_2 + N_3$ being the total number of observations.

System $(B1)$ is basically a SUR model with cross-equation restrictions. The equations are related because $\text{cov}(\varepsilon_{ijt}, \varepsilon_{ijs}) \neq 0$ and potentially $\text{cov}(\varepsilon_{ijt}, \varepsilon_{ljs}) \neq 0$ for $s \neq t$; the latter correlation could for example arise because of common shocks to the training firm. OLS and OLSFE are not efficient but consistent under these assumptions on the error structure.

To take into account the potential correlations of individuals and trainees within the same training firm, both models are calculated using STATA's cluster procedure to cluster standard errors at the training firm level.

Equation $(B1)$ is also supposed to be estimated with instrumental variables since it is suspected that $E\{D_{ij0}\varepsilon_{ijt}\} \neq 0$, $\forall\ t = 1,3,5$, $i = 1,...,N$. Consider the vector of instruments $Z = [z'_1, z'_2, z'_3]'$ for $D$, for which it is assumed that $E\{z_{ij(i)}\varepsilon_{ijt}\} = 0$, $\forall\ t = 1,3,5$, $i = 1,...,N$. Then IV can be implemented by first regressing $D$ on $Z$ and the other variables of the model in a first stage of the form

$$D = \Lambda_1 Z\pi_1 + \Lambda_2 Z\pi_2 + \Lambda_3 Z\pi_3 + \Lambda_1 X\tilde{\beta}_1 + \Lambda_2 X\tilde{\beta}_2 + \Lambda_3 X\tilde{\beta}_3 + \Phi^j\tilde{\gamma} + \bar{u} \qquad \text{(B2)}$$

and regressing $W$ on $Z$ in the reduced form

$$W = \Lambda_1 Z\phi_1 + \Lambda_2 Z\phi_2 + \Lambda_3 Z\phi_3 + \Lambda_1 X\breve{\beta}_1 + \Lambda_2 X\breve{\beta}_2 + \Lambda_3 X\breve{\beta}_3 + \tilde{\Phi}^j + \tilde{\bar{\varepsilon}}. \qquad \text{(B3)}$$

The IV estimator for $\delta = (\delta_1, \delta_3, \delta_5)$ is then obtained by dividing the first and second stage coefficients, i.e., $\hat{\delta}_t = \dfrac{\hat{\phi}_t}{\hat{\pi}_t}$, $t = 1,3,5$.[52]

To see that this is a consistent estimator for $\delta$, first substitute stage $(B2)$ into the structural equation $(B1)$ to obtain

$$W = \Lambda_1\left[\Lambda_1 Z\pi_1 + \Lambda_2 Z\pi_2 + \Lambda_3 Z\pi_3 + \Lambda_1 X\tilde{\beta}_1 + \Lambda_2 X\tilde{\beta}_2 + \Lambda_3 X\tilde{\beta}_3 + \Phi^j\tilde{\gamma} + \bar{u}\right]\delta_1 + ... + \tilde{\bar{\varepsilon}}$$
$$\Leftrightarrow W = \Lambda_1\left[Z\pi_1 + X\tilde{\beta}_1 + \Phi^j\tilde{\gamma} + \bar{u}\right]\delta_1 + \Lambda_2\left[Z\pi_2 + X\tilde{\beta}_2 + \Phi^j\tilde{\gamma} + \bar{u}\right]\delta_2 + \Lambda_3\left[Z\pi_3 + ... + \bar{u}\right]\delta_3 + ... + \tilde{\bar{\varepsilon}},$$

where I have used the fact that $\Lambda_1\Lambda_2 = \Lambda_1\Lambda_3 = \Lambda_2\Lambda_3 = 0$. Rearranging terms one obtains that $Y = \Lambda_1 Z\pi_1\delta_1 + \Lambda_2 Z\pi_2\delta_2 + \Lambda_3 Z\pi_3\delta_3 + \Lambda_1 F\breve{\beta}_1 + \Lambda_2 F\breve{\beta}_2 + \Lambda_3 F\breve{\beta}_3 + \tilde{\Phi}^j + \tilde{\bar{\varepsilon}}$, where $\breve{\beta}_t \equiv \beta_t + \tilde{\beta}_t\delta_t$, $\tilde{\Phi}^j \equiv \Phi^j\gamma + \Lambda_1\Phi^j\tilde{\gamma}\delta_1 + \Lambda_2\Phi^j\tilde{\gamma}\delta_2 + \Lambda_3\Phi^j\tilde{\gamma}\delta_3$ and $\tilde{\bar{\varepsilon}} \equiv \bar{\varepsilon} + \Lambda_1\bar{u}\delta_1 + \Lambda_2\bar{u}\delta_2 + \Lambda_3\bar{u}\delta_3$. Since we have $E\{z_t u_t\} = 0$ and $E\{z_t\varepsilon_t\} = 0$, it follows that $E\{(\Lambda_t Z)'\tilde{\bar{\varepsilon}}\} = E\{z_t(\varepsilon_t + u_t\delta_t)\} = E\{z_t\varepsilon_t\} + \delta_t E\{z_t u_t\} = 0\ \ \forall\ t = 1,3,5$. Thus, OLS estimation yields consistent estimators in the first stage and the reduced form, i.e.,

---

[52] The IV estimator can also be obtained by estimating the fitted value $\hat{X}$ from the first stage in equation (B2) and substituting it into equation (B1) for $X$ to obtain the second stage. The coefficients on $[\Lambda_1\hat{X}, \Lambda_2\hat{X}, \Lambda_3\hat{X}]$ are then the IV-estimates. This amounts to using $[\Lambda_1\hat{X}, \Lambda_2\hat{X}, \Lambda_3\hat{X}]$ as instruments for $[\Lambda_1 X, \Lambda_2 X, \Lambda_3 X]$ and thus the same proof of consistency applies. Note that this is not the same as treating $[\Lambda_1 X, \Lambda_2 X, \Lambda_3 X]$ as three separate endogenous variables and using $[\Lambda_1 Z, \Lambda_2 Z, \Lambda_3 Z]$ as instruments in three separate first stages.

$p\lim \hat{\phi}_t = \pi_t \delta_t$ and $p\lim \hat{\pi}_t = \pi_t$. Then the Slutsky-Theorem implies

$\hat{\delta}_t^{IV} = \frac{\hat{\phi}_t}{\hat{\pi}_t} \to \delta_t$ as $n \to \infty$.

If we assume that the variance matrix of the error terms is scalar (this is only explained for clarity and <u>not</u> what is done in the paper), by the partitioned inverse formula the correct variance matrix for the IV estimator is $\text{var}\{\hat{\delta}\} = \sigma_{IV}^2 \left(\hat{D}_{[3]}' M_B \hat{D}_{[3]}\right)^{-1}$, where $\hat{D}_{[3]} = \left[\Lambda_1 \hat{D}, \Lambda_2 \hat{D}, \Lambda_3 \hat{D}\right]$ and $B \equiv [X, \Phi]$. The variance term is estimated consistently by $\hat{\sigma}_{IV}^2 = \frac{1}{N-K}\left(\hat{Y} - \hat{D}\hat{\delta}^{IV}\right)'\left(\hat{Y} - \hat{D}\hat{\delta}^{IV}\right)$, where $\hat{D} \equiv M_{[X,\Phi]}\left[\Lambda_1 D, \Lambda_2 D, \Lambda_3 D\right]$ and $\hat{Y} \equiv M_{[X,\Phi]} Y$. In the paper, $\text{var}\{\vec{\varepsilon}\}$ will be assumed to be block diagonal, such that

$$\text{var}\{\hat{\delta}\} = \left(\hat{X}' M_B \hat{X}\right)^{-1} \hat{X}' M_B \, \text{var}\{\vec{\varepsilon}\} M_B \hat{X} \left(\hat{X}' M_B \hat{X}\right)^{-1}.$$

The fitted residuals implied by the IV estimates are again used to calculate the blocks of the variance-covariance matrix. This procedure can be implemented using STATA's cluster sub-routine.