Appendix B: Multivariate Reweighting of a Discrete Distribution Function

For a discrete distribution function, a consistent estimator of the true distribution function is the sample proportion estimator, given by:

$$\hat{f(x)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(x_i = x), \tag{1}$$

where x_i denotes a data point for individual i and $I(x_i = x)$ is an indicator function that equals 1 if $x_i = x$.

Let x_i denote the time to degree for individual i in the sample. Further, let y_i denote an attribute or a set of attributes of individual i in the sample. As discussed previously, respondents come from two different surveys: NLS72 and NELS:88 and the characteristics of those in the sample differ in systematic ways that may affect time to degree. Let the conditional distribution of y for each sample be given by $g(y|72)$ and $g(y|88)$ for the NLS72 and the NELS:88 surveys respectively.

The re-weighting analysis seeks to estimate the counterfactual distribution of time to degree for NELS:88 respondents if they had the same distribution of attributes as in NLS72. More formally, we wish to estimate:

$$f_{72}(x|88) = \int f(x|88, y)g(y|72)dy \tag{2}$$

The difficulty with estimating this counterfactual is that we only observe

$$f(x|88) = \int f(x|88, y)g(y|88)dy \tag{3}$$

By re-weighting the distribution of NELS:88 characteristics, however, we can equate the distribution of characteristics in each sample. In other words, we wish to find a weighting function $\omega(y)$ such that $g(y|72) = \omega(y)g(y|88)$. By Bayes rule, it is straightforward to calculate [1]

$$\omega(y) = \frac{P(72|y)}{P(88|y)}\frac{P(88)}{P(72)}. \tag{4}$$

[1] A more complete derivation is given in Dinardo, Fortin and Lemieux (1996)

Using this weighting function, we can re-weight the density function given by (3) to yield the counterfactual density function given by (2). More formally,

$$
\begin{aligned}
(5) \qquad f_{72}(x|88) &= \int f(x|88,y)g(y|72)dy \\
&= \int f(x|88,y)g(y|88)\omega(y)dy.
\end{aligned}
$$

To implement the multivariate re-weighting technique, we must first estimate the weights for each individual. We run a logit of a dummy variable for being in the 1972 sample on exogenous individual characteristics listed in Table 5. Results of the regression are presented in Table 6. The fitted values of the logit yield estimates of $P(72|y)$ and $P(88|y)$. Denote $P(\hat{72}|y)$ as $\rho(y)$, which implies that $P(\hat{88}|y)$ is denoted $1-\rho(y)$. The weighting function estimator is thus

$$
(6) \qquad \hat{\omega(y_i)} = \frac{\rho(y_i)}{1-\rho(y_i)}\frac{n_{88}}{n_{72}},
$$

where n_{72} and n_{88} are the sample proportions in each survey.

By applying the weighting function estimator to the sample proportion estimator, we construct the following consistent estimator of the counterfactual time to degree distribution

$$
(7) \qquad \hat{f(x)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\omega(y_i)}\mathrm{I}(x_i = x).
$$

Results of the multivariate reweighting given by (7) are presented in Figure 4 of the main text.