Appendix: Missing Data and Multiple Imputation

Upon constructing the variables described in Section 4.1 and combining them into one individual-level data set, we observed that there was a significant amount of missing data. In the NES survey some individuals did not report either occupation, educational attainment, or industry of employment. This prevented us from constructing some of the factor-income trade-exposure variables for these people. The most serious missing-data problem arose from the homeowners' exposure variables. The county-level COM data suppress some information at the two-digit SIC level to prevent disclosure of individual firms. This hampered our construction of *County Exposure 1* and *County Exposure 2*. Recall that for each county these variables require information on 10 and 14 of the 20 two-digit industries, respectively. Suppressing data for just one industry in the county can be sufficient to prevent construction of one or both of the variables for that county. Overall, when we simply dropped observations with any missing data we lost between 4.4% and 73.4% of all observations depending on which model was estimated.

This standard approach for dealing with missing values, known as "listwise deletion," can create two major problems. One is inefficiency suffered from throwing away information relevant to the statistical inferences being made. Furthermore, inferences from listwise-deletion estimation can be biased if the observed data differs systematically from the unobserved data. In our case inefficiency was clearly a problem. We also had little reason to believe our data were missing randomly. Individuals of certain types might tend not to report personal information, and the COM suppression probably hits counties with more concentrated industrial mixes.

Alternatives to listwise deletion for dealing with missing data have been developed in recent years. The most general and extensively researched approach is "multiple imputation" (King et al. (2000), Schafer (1997), Little and Rubin (1987), Rubin (1987)). Multiple imputation makes a much weaker assumption than listwise deletion about the process generating the missing data. Rather than assuming that the unobserved data is

missing completely at random, multiple imputation is consistent and gives correct uncertainty estimates if the data are missing randomly conditional on the data included in the imputation procedures (the method also requires that the parameters describing the data are distinct from the parameters describing the missing mechanism in the data). The approach has several variations but always involves three main steps. First, some algorithm is used to impute values for the missing data. In this step, *m* (*m*>1) "complete" data sets are created consisting of all the observed data and imputations for the missing values. The second step simply involves analyzing each of the *m* data sets using standard complete-data statistical methods. The final step combines the parameter estimates and variances from the *m* complete-data analyses to form a single set of parameter estimates and variances. Importantly, this step systematically accounts for variation across the *m* analyses due to missing data in addition to ordinary sample variation.

The first step in our multiple-imputation procedures was to impute missing observations for *County Exposure 1* and *County Exposure 2*. We based our imputations on 46 county-level variables selected from the COM based on their sample correlation with *County Exposure 1* and *County Exposure 2* (which were two of the 46). For example, we selected county employment in textiles because it had one of the highest sample correlations with our county-exposure variables. In general, our 46 variables were various measures of factor endowments and economic output such as educational attainment and employment by industry. Altogether we imputed 10 complete county data sets.

The exact algorithm used for the imputations is a data augmentation method known by the acronym "IP" because it involves two key steps: the imputation step and the posterior step. The goal of the imputation procedure is to estimate a set of parameters that can be used to create the 10 imputed data sets. In this application, it is assumed that the data have a joint multivariate normal distribution. Consequently, IP employs an iterative sampling scheme where in the first step imputations are drawn from the

multivariate normal conditional predictive distribution of the missing data. This distribution depends on the observed data and the assumed or current value of the complete data parameters. In the second step, a new value of the complete data parameters is drawn from its posterior distribution, which is conditioned on the observed data and the current values of the imputations for the missing data. This posterior step is a simulation from the normal inverted-Wishart distribution. Repeating this iterative sampling scheme produces stochastic subsequences that converge on the stationary predictive distribution for the missing values and the stationary posterior distribution of the complete data parameters.

For the county data set we ran 5000 preliminary iterations of IP and then ran 1000 more to create an imputed data set every 100 iterations of these last 1000. The preliminary iterations ensure that sequences have converged to their stationary distributions. After creating the 10 complete county data sets we merged the NES survey data (including our constructed skill measures and industry measures) with *County Exposure 1* and *County Exposure 2*. The resulting 10 data sets still had substantial amounts of missing individual-level data, however. Consequently, for each of these 10 data sets we ran separate iterations of IP in order to impute values for the missing survey data. We found that 2100 preliminary iterations were more than sufficient for these data sets. An imputation was saved on the last iterations of each of the 10 cases to create our 10 final data sets with no missing data at all. Each of these final data sets contains 1736 observations, equal to the actual number of individuals in the NES survey either supporting or opposing more trade restrictions. Also, each data set contains the exact same non-imputed information (i.e., all observations for the variable *Trade Opinion* plus the non-imputed observations for all the trade-exposure variables). They differ only in their imputed values for missing data. All the main results reported in this paper are qualitatively the same for the case where imputations are also made by treating as missing data the fact that some respondents did not express a trade-policy opinion (i.e.,

they chose the option "haven't thought much about this"). For this analysis the multiple-imputation procedures created 10 data sets of 2485 observations equal to the total number of respondents in the NES survey.

The second step in our multiple-imputation analysis was to run various logit models separately on each of the 10 final data sets. The last multiple-imputation step was to combine the 10 sets of estimation results to obtain a single set of estimated parameter means and variances. The single set of estimated means is simply the arithmetic average of the 10 different estimation results. The single set of estimated variances consists of two parts. The “within” component is simply the arithmetic average of the 10 estimated variances. This accounts for the ordinary within-sample variation. The “between” component is the variance of the estimated parameter means among the imputed data sets. See King et al. (2000) and Schafer (1997) for a complete description of the last multiple-imputation step.