Web Appendix for "Task Specialization, Immigration and Wages"

Giovanni Peri (University of California, Davis and NBER)
Chad Sparber (Colgate University)

January 2009

1 Data and Variables

1.1 General Description

The IPUMS dataset by Ruggles et. al. (2005) provides individual-level data on personal characteristics, employment, wages, immigration status, and occupation choice. As consistent with the literature, we identify immigrants as those who are born outside of the United States and were not citizens at birth. To focus on the period of rising immigration and to use only Census data we consider decennial years from 1960 to 2000. Specifically we use the 1% IPUMS sample for the 1960 and 1970 Census data, and the 5% IPUMS sample for 1980, 1990 and 2000. We include workers who were between 18 and 65 years of age, not residing in group quarters, and who worked at least one week in the year prior to the Census year and at least one hour in the reference week. We do not exclude self-employed as long as they report wage income. The exclusion of self-employed did not noticeably change any result (estimates are available upon request). We also calculate the potential experience of workers assuming that those without a degree started working at age 17, and those with a high school diploma started at 19. We then eliminate workers with less than one year and more than 45 years of experience. Whenever we construct aggregate or average variables, we weight each individual by his/her personal Census weight, multiplied by the number of hours he/she worked in a year. The number of hours worked in a year equals the number of weeks worked in the year (measured by the IPUMS variable *wkswork2* in 1960 and 1970 and *wkswork1* from 1980-2000) times the number of hours usually worked (*hrswork2* in 1960 and 1970 and *uhrswork* subsequently). As *wkswork2* and *hrswork2* are categorical variables we attribute to each category the median value of the interval. Weighting by Hours worked in a year allows us to put less weight on part-time workers, and to create variable values reflecting the amount of hourly labor individuals actually supply.

1.2 Construction of Task Variables

By merging occupation-specific task values with individuals across Census years, we are able to obtain these task supply measures for natives and immigrants by education level in each state over time. The US Department of Labor's *O*NET* abilities survey (we use version 11.0 of the survey, which is publicly available at http://www.onetcenter.org/) provides information on the characteristics of occupations. Initiated in 2000, this dataset assigns numerical values to describe the importance of 52 distinct employee abilities (which we refer to as "tasks" or "skills") required by each SOC (standard occupation classification) occupation. We merge these occupation-specific values to individuals in the 2000 Census using the SOC codes. The arbitrary scale of measurement for the task variables encourages us to convert the values into percentiles. We assume that the 2000 Census is collectively representative of the US workforce, and then re-scale each skill variable so that it equals the percentile score representing the relative importance of that skill among all workers in 2000. For instance, an occupation with a score of 0.02 for a specific skill indicates that only 2% of workers in the US in 2000 were using that skill less often. Since Census occupation codes vary across years, we then assign these *O*NET* percentile scores to individuals from 1960 to 2000 using the IPUMS variable *occ1990*, which provides a crosswalk for occupations over time. The standardization of skill values between zero and one facilitates a more intuitive interpretation of their percentage changes over time.

Table A1 in the Appendix lists each of the 52 *O*NET* variables and organizes them into categories that we use to construct our manual and communication skill supply indices. In our "basic" definition of manual skills we average only the variables capturing an occupation's "Movement and Strength" requirements. As Table A1 shows, those skills can be further divided into "Limb, Hand, and Finger Dexterity," "Body Coordination and Flexibility," and "Strength." Similarly, our basic definition of communication skills includes measures of oral and written expression and comprehension.

In our "extended" definition of manual skills we add "Sensory and Perception" abilities (i.e. those using the five senses) to the physical skill group. In the extended definition of communication skills, used only in the robustness checks performed in the Web Appendix Tables, we introduce "Cognitive and Analytical" and "Vocal" abilities to the communication skill group.

To produce the summary statistics we calculate the aggregate (US or state-level) supply of manual skills for less educated immigrants (M_F), natives (M_D), or both groups of workers (M) by summing the values of m_j across individuals. We also weight m_j by individual sample weights (the IPUMS variable $PERWT$), multiplied by hours worked. We follow an analogous procedure for aggregate communication skills (creating C_F, C_D, and C). Average values for states (or the US) are represented by m or c (with subscripts if appropriate), and are obtained by dividing the aforementioned aggregate variables by the number of hours worked times the sample weight of the

considered population. That is, $m_D = M_D/L_D$, $m_F = M_F/L_F$ and $m = (M_D + M_F)/(L_D + L_F)$. Similarly for average communication skills: $c_D = C_D/L_D$, $c_F = C_F/L_F$ and $c = (C_D + C_F)/(L_D + L_F)$.

2 Regressions to Control for Individual Characteristics

To construct the aggregate state variables we first clean the task and wage data from the effect of individual characteristics. To control for personal characteristics in the task data, we first select workers with at most a high school degree and regress an individual's task supply (derived from her occupation) on a set of dummies capturing experience (44 distinct years of experience indicators), education (an indicator for having obtained a high school diploma), gender (a female indicator), and race/ethnicity (six indicators for Black, Hispanic, Native Americans, Chinese, Japanese, and other races). We do these regressions separately for each Census year and *O*NET* variable and separately for natives and immigrants. Next, we subtract an individual's predicted task supply from his or her observed value. This residual represents an individual's skill "cleaned" of demographic effects. We then compute an individual's total manual and communication task supply by averaging the relevant residuals (i.e., by averaging the cleaned *O*NET* variables belonging to each skill type as defined in Table 1). Finally, we create state-level averages for native workers ($(c_D)_{st}$ and $(m_D)_{st}$) and their ratio $\left(\frac{c_D}{m_D} = \frac{C_D}{M_D}\right)$ for each state s and year t by weighting each individual by his or her personal weight (and hours worked). The cleaning procedure eliminates the cross-state variation of skills due to the demographic features of natives. However it maintains the national average skill intensity for the group of natives and immigrants in each year.

To control for individual characteristics in the wage data and to calculate the unit compensation of communication and manual skills, w_M and w_C, for each state and year we follow two steps. First, we select workers with at most a high school degree and calculate their real weekly wage by dividing the yearly salary income by the number of weeks worked in the year. The nominal figures are then converted into real figures using the CPI-U deflator published by the Bureau of Labor Statistics and available at www.bls.gov/cpi. We regress, by year, the logarithm of individual real weekly wages on a set of dummies capturing experience (44 distinct years of experience indicators), education (an indicator for having obtained a high school diploma), gender (a female indicator), race/ethnicity (six indicators for Black, Hispanic, Native Americans, Chinese, Japanese, and other races) and nativity (US or foreign). These characteristics are identical to those used to clean the individual task data above. These regressions also include occupation by state dummies, whose coefficients represent our estimates for the average log-wage, $\ln(\tilde{w}_{jst})$, for occupation j, state s, and Census year t after removing individual characteristic effects. Regressions weight each individual by their Census sample weight times hours worked.

In the second step, we transform $\ln(\tilde{w}_{jst})$ into levels and regress $\tilde{w}_{jst}$ on the occupation-

specific measures of manual and communication skills (obtained from *O*NET*) using weighted least squares. We do this using the basic as well as the extended definitions of skills described in Table A1. We then allow the coefficients on the skill variables to vary across states so that they capture the compensation (price) of manual and communication tasks in each state. By separately estimating the second stage regression in Equation (1) for each year, we can identify the state and year-specific wages, $(w_M)_{st}$ and $(w_C)_{st}$, received for supplying manual and communication tasks.

$$\tilde{w}_{jst} = (w_M)_{st} \cdot m_j + (w_C)_{st} \cdot c_j + \varepsilon_{jst} \tag{1}$$

Equation (1) implements the theoretical relation to infer the values of w_M and w_C in a market (state) from the occupational wages in that market. In order to obtain coefficients $\widehat{w_{Mst}}$ and $\widehat{w_{Cst}}$ that could be interpreted as the weekly compensation of a skill (and therefore always assuming positive values), and in line with our model, we do not include a constant in (1). We also implement the above regression using the "extended' definition of each skill. The results are very similar to those obtained using the "basic" specification.

3 Instrumental Variables Construction

The "Imputed Mexican share" instrument is constructed as follows. First, we record the actual share of Mexicans in the employment of state s in 1960 ($MEX_{s,t}$), and then assume that the growth rate of the Mexican share of employment between 1960 and year t was equal, across states, to its national average. The figures used to impute the Mexican share of employment are not weighted by hours worked. Equation (2) imputes shares in year t, where $(1 + g_{MEX})_{1960-t}$ is the growth factor of Mexican-born employment nationwide between 1960 and year t, and $(1 + g_{US})_{s,1960-t}$ is the growth factor of US-born workers in state s between 1960 and year t. The identification power of the instrument is based on the fact that some states (such as California and Texas) had a larger share of Mexican immigrants in 1960 relative to others. These states will also have larger imputed shares of Mexicans in 1970 through 2000 and, due to the educational composition of this group, will have a larger immigrant share among less educated workers.

$$\widehat{MEX}_{s,t} = MEX_{s,1960} \frac{(1 + g_{MEX})_{1960-t}}{(1 + g_{US})_{s,1960-t}} \tag{2}$$

Our second set of instruments similarly relies upon the exogenous increase in Mexican immigration but is based upon geography. First, we use the formula for geodesic distance to calculate the distance of each state's population center of gravity (available from the 2000 Census) to its closest section of the Mexican border. We divide the US-Mexico border into 12 sections and calculate the distance between each center of gravity and each section. Then we choose the shortest distance

for each state. Since we already control for state fixed effects in the regressions, we interact the distance variable with four year dummies (from 1970 to 2000). This captures the fact that distance from the border had a larger effect in predicting the inflow of less educated workers in decades with larger Mexican immigration. Second, we also use a Mexican border dummy interacted with decade indicators to capture the fact that border states had larger inflows of Mexican workers due to undocumented border crossings. Since illegal immigrants are less mobile across states, border states have experienced a particularly large exogenous supply-driven increase of less educated immigrant workers. Altogether, our second set of instruments includes both the distance and border variables, each interacted with decade indicators.

4 Computer Usage and Industry-Driven Task Demand

Our period of analysis is associated with large changes in production technologies, particularly in the diffusion of information technologies and computer adoption. Autor, Frank Levy, and Richard Murnane (2003) demonstrate that this change had a large effect in shifting demand from routine to non-routine tasks. Similarly, the increasing importance of advanced services, the demise of manufacturing, and other sector-shifts might have contributed substantially to differences across states in the demand for manual and communication tasks. State-specific technology and/or sector composition could confound the correlation between immigration and task intensity.

We begin to account for these factors by including the share of workers (with at most a high school degree) who use a computer at work to control for the diffusion of technology across states. This data is available in the October CPS Supplements in 1984 and 1997, and in the September CPS Supplement in 2001. We match the 1984 computer data to the 1980 Census data, the 1997 computer-use data to the 1990 Census, and the 2001 computer data to the 2000 Census. We impute a share of zero for all states in 1960 and 1970 since the personal computer was first introduced in 1981.

Our second control accounts more explicitly for the industrial composition of each state in 1960 and its effect on task demand. We create state-specific indices of communication versus manual task demand driven by each state's industrial composition, $\left(\frac{C}{M}\right)^{Tech}$, by assuming that the occupational composition of industries and industry-specific employment shocks are uniform across states. First, we calculate the average physical and language content among all workers for each industry i in year t from national data and record the corresponding ratio $\left(\frac{C}{M}\right)_{i,t}$. Next, we calculate industry-level national employment growth since 1960, $g_{i,t}$. By assuming that industries grew at their national growth rates regardless of the state in which they are located, we can predict the employment share of industries within each state and year, $\widehat{emp}_{i,s,t}$. Finally, we calculate a state's level of relative task demand, $\left(\frac{C}{M}\right)^{Tech}_{s,t}$, as the average value of each industry's $\left(\frac{C}{M}\right)_{i,t}$

weighted by the predicted employment shares.

$$\widehat{emp}_{i,s,t} = \frac{Employment_{i,s,1960} \cdot (1 + g_{i,t})}{\sum_{i=1}^{Ind} Employment_{i,s,1960} \cdot (1 + g_{i,t})} \tag{3}$$

$$\left(\frac{C}{M}\right)_{s,t}^{Tech} = \sum_{i=1}^{Ind} \widehat{emp}_{i,s,t} \cdot \left(\frac{C}{M}\right)_{i,t} \tag{4}$$

5 Estimates of Native Employment Response

Most regional analyses find that immigration generates little to no native employment effect. In a recent note (Peri and Sparber (2008a)), we argue that to obtain an unbiased estimate of the potential displacement effect across states (or any geographical unit), one should perform 2SLS estimation of Equation (5).

$$\left(\frac{\Delta L_D}{L_D + L_F}\right)_{st} = \tau_t + \eta \left(\frac{\Delta L_F}{L_D + L_F}\right)_{st} + \varepsilon_{st} \tag{5}$$

This model regresses the inter-Census change in native employment (ΔL_D) on the change in foreign-born employment (ΔL_F). Effective instruments should avoid "booming region" effects (which would induce positive correlation due to unobserved positive regional shocks). The parameter η then identifies the effect of immigration on native employment. A significantly negative value implies displacement. Using our 1960-2000 state data we estimate a non-significant positive value of η equal to 0.36 (standard error of 0.40) using WLS, and a non-significant positive value of 0.31 (with a standard error of 0.50) using 2SLS and the usual Imputed share of Mexicans and geographical variables as Instruments

Several previous studies that use specifications akin to (5) also tend to find zero or small positive effects. Cortes (2008) uses a variant of (5) in levels to analyze the link between immigration and employment of less educated workers across 25 US metropolitan areas between 1980 and 2000. She finds a positive OLS estimate around 0.20 and an IV value near 0.05. Card (2001), who uses population growth in a city-skill group cell as the dependent variable and the inflow rate of immigrants in the same cell as the explanatory variable, always finds positive and sometimes significant effects on the native population (around 0.10). His subsequent IV estimates (using the shift-share instrument to impute the number of immigrants in a cell) often find results similar to those of his OLS regressions. Ottaviano and Peri (2007) aggregate individuals from all skill levels within a state and estimate an impact of immigration on native employment between -0.3 and 0.3 that is never significant (standard errors around 0.3). Card and Lewis (2007) estimate the effect of low skilled Mexican immigrants on native employment. Their Table 6 results find an

effect between 0 and 0.5 that is rarely significant. Card's (2007) Specification (2) adopts the total (immigrant and native) change in the less educated population (or employment) as the dependent variable. His estimated coefficient implies a value of η slightly larger than zero.