

**Appendix B: Data and Methodological Details for
“Explaining Stagnation in the College Wage Premium,”
by Leila Bengali, Robert G. Valletta, and Cindy Zhao***

**First version: January 27, 2025
This version: June 5, 2025**

* This work was largely completed while Cindy Zhao was a Research Associate at the Federal Reserve Bank of San Francisco. The authors thank David Card, Thomas Lemieux, and David Autor for comments and suggestions on the manuscript. They also thank Daniel Mangrum for discussant comments, along with Rajashri Chakrabarti and other participants at the Federal Reserve System Economic Heterogeneity conference in April 2024, plus session participants and attendees at the March 2025 Midwest Economics Association meetings. The views expressed in this paper are solely those of the authors and are not attributable to the Federal Reserve Bank of San Francisco or the Federal Reserve System.

Appendix B

In this Appendix, we describe our data handling and construction choices for the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) microdata (Flood et al. 2024). Our sample is for the years 1962–2024 (reference years 1961–2023 for the variables in the CPS ASEC, such as hours and earnings, that ask about the year prior to the year in which the respondent is surveyed).¹ We largely follow the methods outlined in Card and Lemieux (2001), with some deviations, as described in this Appendix.

I. Data Cleaning and Processing

- We drop individuals meeting any of these criteria: no or unknown weeks worked in the reference year, no or unknown annual wage and salary income (“earnings” or “wages”) in the reference year, weekly reference year earnings (defined below) of less than \$50 in 1989 dollars (using the CPI-U), not aged 25–64 in the reference year (we define reference year age as survey year age minus one), unknown race/ethnicity, zero or negative ASEC survey weight, individuals part of the 3/8 sample redesign in the 2014 ASEC (see Flood et al. 2024), unknown full- or part-time status in the reference year, unknown educational attainment, imputed values for annual earning amounts.
- For (survey) years prior to 1976, only a binned count of weeks worked in the reference year is available. To transform these bins into actual weeks of work, we use mean weeks worked in each of the bins based on data from 1976–1980 (bins and means used provided below). After 1976, the exact weeks worked variable is used.
 - 8 weeks for 1–13 weeks category
 - 18 weeks for 14–26 weeks category
 - 34 weeks for 27–39 weeks category
 - 43 weeks for 40–47 weeks category
 - 48.5 weeks for 48–49 weeks category
 - 52 weeks for 50–52 weeks category
- We follow the methodology from the Center for Economic and Policy Research (CEPR) to adjust top-coded annual earnings. This method uses non-top-coded earnings to create a lognormal approximation of the distribution of earnings and then replaces each top-coded earnings value with the mean above the top code from the distribution.² We apply this procedure prior to implementing any restrictions based on demographics (such as age,

¹ See Flood et al. (2024). Reference year 1962 (survey year 1963) is dropped because the educational attainment variable is blank for all observations, despite the CPS ASEC codebook indicating that this variable should be populated. This issue has been documented elsewhere (see IPUMS documentation of HIGRADE).

² See <https://ceprdata.org/cps-uniform-data-extracts/cps-outgoing-rotation-group/cps-org-faq/>.

gender, or educational attainment). This is to ensure consistency in replaced values across analysis runs.

- We construct weekly earnings as annual earnings in the reference year (adjusted as noted) divided by weeks worked (defined as noted) in the reference year and calculate real earnings using the CPI-U.
- We largely follow Valletta (2018) in forming educational attainment groups. Specifically, prior to 1992, workers with 12 years of completed education are classified as high school graduates, while workers with at least 16 years of completed education and at most 17 years of completed education are classified as college graduates. Those with 18 or more years of education (regardless of completion) are coded as having a graduate degree. (This marks a departure from Valletta (2018). The categories created as described produce the smoothest time series of education shares across the 1992 redesign.) After 1992, workers who have a high school diploma or equivalent are classified as high school graduates, while workers with a bachelor's degree are classified as college graduates.

II. Wage Premium Construction

- We estimate the college wage premium using both men and women when running analysis either by age group or by age-by-gender groups. (This is a departure from Card and Lemieux (2001), who use only men in estimating the wage premium.) In both cases, to estimate the wage premium, we use respondents who meet all of the following criteria (above and beyond the data restrictions noted above): exactly a high school graduate or exactly a college graduate, full-time worker in the reference year, wage and salary worker in the reference year.
- We group individuals into 5-year age bins (or 5-year age-by-gender bins).
- For the estimated wage premia, we regress the natural log of real weekly earnings on a dummy for being a college graduate, a linear age term, and an indicator for nonwhite race. Regressions are weighted using ASEC survey weights. The regressions are run separately for each year and age-by-gender group.
- We save the standard error on the college dummy coefficient from each year and group's regression for use in subsequent analysis.

III. Group-Specific Supply Construction

- Yearly hours of work are calculated by multiplying weeks of work by 40 for full-time workers and by 20 for part-time workers.
- We define "high school labor" (H_{jt} , for year t and age or age-by-gender group j) following Card and Lemieux (2001) as the total annual hours worked by high school graduates plus the total hours of those with less than a high school degree (weighted by their earnings relative to high school graduates) plus a share of the annual hours worked

by workers with some college. This share is determined by finding the earnings of “some college” workers as a weighted average of high school and college graduates’ earnings. Similarly, “college labor” (C_{jt}) is defined as the total annual hours worked by college graduates plus the total hours of those with more than a college degree (weighted by their earnings relative to college graduates) plus a share of the annual hours worked by workers with some college. The share is determined through the calculation mentioned previously. We use ASEC survey weights when summing annual hours.

- We create these labor supplies for each year and age group (or year and age-by-gender group) using individuals who meet all of these criteria (above and beyond the data restrictions noted above): any level of educational attainment, full- and part-time workers in the reference year, any class (i.e. not just wage and salary) of worker in the reference year, both men and women.

IV. Aggregate Relative Supply Construction

- As noted in the main text, C_t and H_t are the aggregate quantities of college-educated and high school-educated labor. As explained in the main text, these are defined as

$$H_t = \left[\sum_j (\alpha_j H_{jt}^\eta) \right]^{\frac{1}{\eta}} \quad (A1)$$

$$C_t = \left[\sum_j (\beta_j C_{jt}^\eta) \right]^{\frac{1}{\eta}} \quad (A2)$$

- To create the aggregate supply measures (C_t and H_t), we need estimates of η and of α_j and β_j for each of our age or age-by-gender groups, j .
- The estimation of η ($= 1 - 1/\sigma_A$) is described in the main text.
- Card and Lemieux (2001) show that α_j and β_j can be estimated as the exponentiated coefficients from a complete set of group effects (the b_j ’s below) in regressions that also include unrestricted year effects (the d_t ’s below; equations in natural log terms):³

$$\log(w_{jt}^h) + \left(\frac{1}{\sigma_A}\right) * \log(H_{jt}) = d_t^h + b_j^h + \epsilon_{jt}^h \quad (A3)$$

$$\log(w_{jt}^c) + \left(\frac{1}{\sigma_A}\right) * \log(C_{jt}) = d_t^c + b_j^c + \epsilon_{jt}^c \quad (A4)$$

³ We adjusted for a typo in Card and Lemieux (2001; equations 12a and 12b) that omitted the log operator from the second terms on the left-hand side of these two equations.

References

- Card, David, and Thomas Lemieux. 2001. "Can falling supply explain the rising return to college for younger men? A cohort-based analysis." *The Quarterly Journal of Economics*: 116(2), pp. 705-746.
- Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Megan Schouweiler, and Michael Westberry. 2024. IPUMS CPS: Version 12.0 [dataset]. Minneapolis, MN: IPUMS. <https://doi.org/10.18128/D030.V12.0>.
- Valletta, Robert G. 2018. "Recent Flattening in the Higher Education Wage Premium: Polarization, Skill Downgrading, or Both?" In Charles Hulten and Valerie Ramey, eds., *Education, Skills, and Technical Change: Implications for Future U.S. GDP Growth*, NBER-CRIW conference volume 313-3: University of Chicago Press.