Web Appendix: The Phantom Gender Difference in the College Wage Premium

William H.J. Hubbard

whubbard@uchicago.edu

Summer 2011

1 Robustness to Sample Composition and Estimation Specification

1.1 Census and ACS Data

A falsification test for my methodology is to employ the same methodology using Census and ACS data for years 1970-2007.[1] For all Census years since 1980, the Census wage income topcodes have been much higher than the corresponding topcodes for CPS data. Further, the ACS wage data, which covers years 2000-2007, is topcoded at 200,000, higher than the CPS topcode through 2001 and equal thereafter.

Because of this, fewer observations in these samples are topcoded; in all years, fewer than 2 percent of all Census and ACS observations are topcoded.

Because fewer observations are topcoded, bias from topcoding—or from any adjustment to

[1]Specifically, the Census samples are the 1 percent unweighted extracts for the 1970-2000 Censuses, including the 1970 Form 2 State sample and 1980 Metro sample within IPUMS. The ACS samples are the individual year data sets for 2000-2004 and the three-year data set covering 2005-2007.

topcodes—should be smaller than in the CPS data. Further, in the 2000 Census and in all ACS surveys, the wage and salary income for topcoded observation is replaced by the mean of all topcoded wages in the same state. For these years, therefore, there is no need to rely on fitting a Pareto distribution to the wage data in order to generate adjustment factors—the year- and sex-specific adjustment factors can be inferred directly from means of topcoded wages. If the adjustments used in the text for the CPS topcoded wage data are valid, then estimates using Census and ACS data should reach essentially the same results.

This is the case. Figure 1 superimposes the results from identical OLS regressions using CPS, Census, and ACS data. Figure 2 does the same for median regressions.

1.2 Non-Parametric Estimates

In the spirit of Katz and Murphy (1992) and more recently Aguiar and Hurst (2007), I also estimate college wage premiums non-parametrically by dividing the universe of observations into cells based on year and (potential) experience, and then computing the college wage premium on a cell-by-cell basis. For each cell, I compute the mean wage among high school graduates and the mean wage among college graduates; the college wage premium is the log of the ratio of these averages. I use these cell-specific college wage premiums to estimate a yearly college wage premium for each sex. Following Katz and Murphy (1992), I compute each year's premium as a weighted average of the cell-specific premiums, using fixed weights for each cell that hold the demographic composition of the sample constant over time. I derive these weights by calculating each cell's share of the FTFY working population in each year, and then computing the average share for each cell over the entire sample period. This average share is the fixed weight assigned to that cell for all periods. Thus, for n demographic cells and t periods of data, I construct a $1 \times n$ vector N of fixed demographic weights, and a $n \times t$ matrix of cell-specific college wage premiums W. The $1 \times t$ vector of yearly college wage premiums is then NW.

Figure 3 displays fixed-weight mean estimates using Census, ACS, and CPS data. Figure 4 presents similar results based on a fixed-weight median estimates.[2] The same consistent pattern appears across all these estimates.

[2]The fixed weights for the Census sample correspond to average shares of the FTFY labor force (by sex) for the 1970-2000 Censuses; the fixed weights for the ACS sample correspond to the average shares for the 2000-2007 ACS samples.

Figure 1: College Wage Premium, OLS Regressions, by Data Set

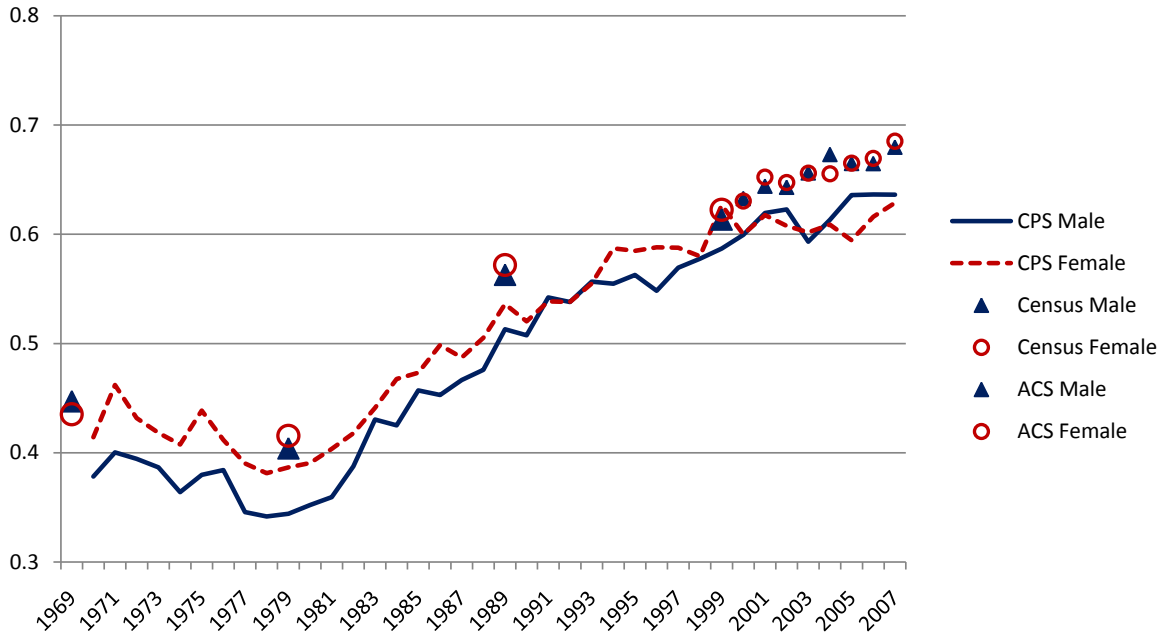


Figure 2: College Wage Premium, Quantile Regressions, by Data Set

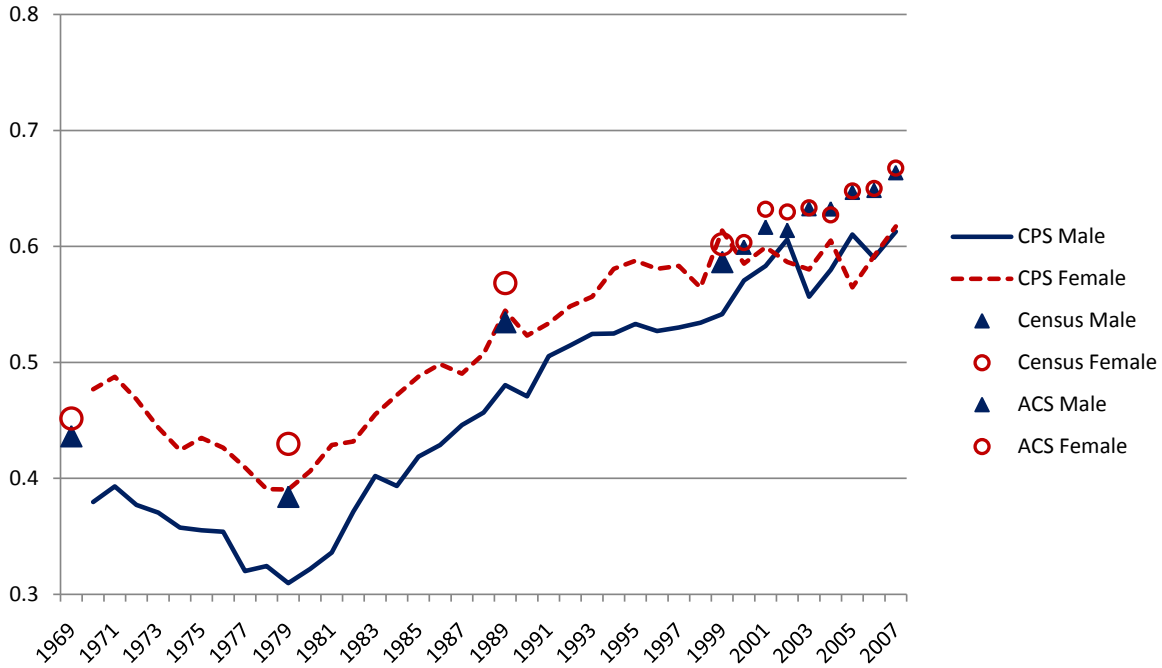


Figure 3: College Wage Premium, Fixed-Weight Mean Estimates, by Data Set

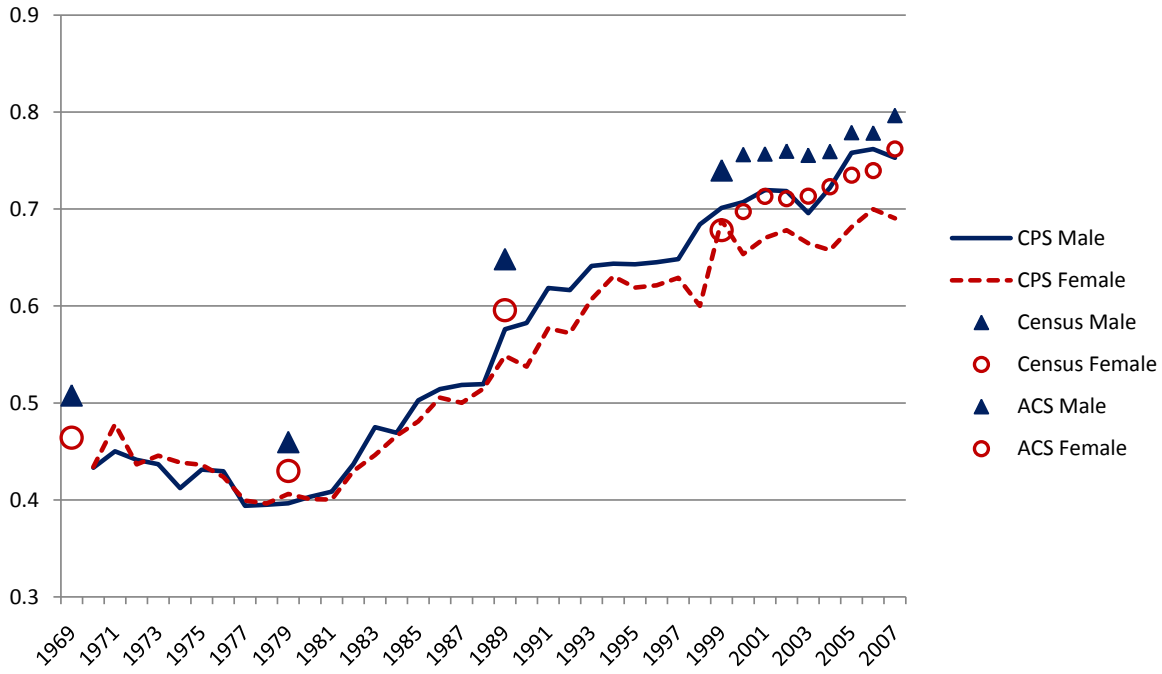
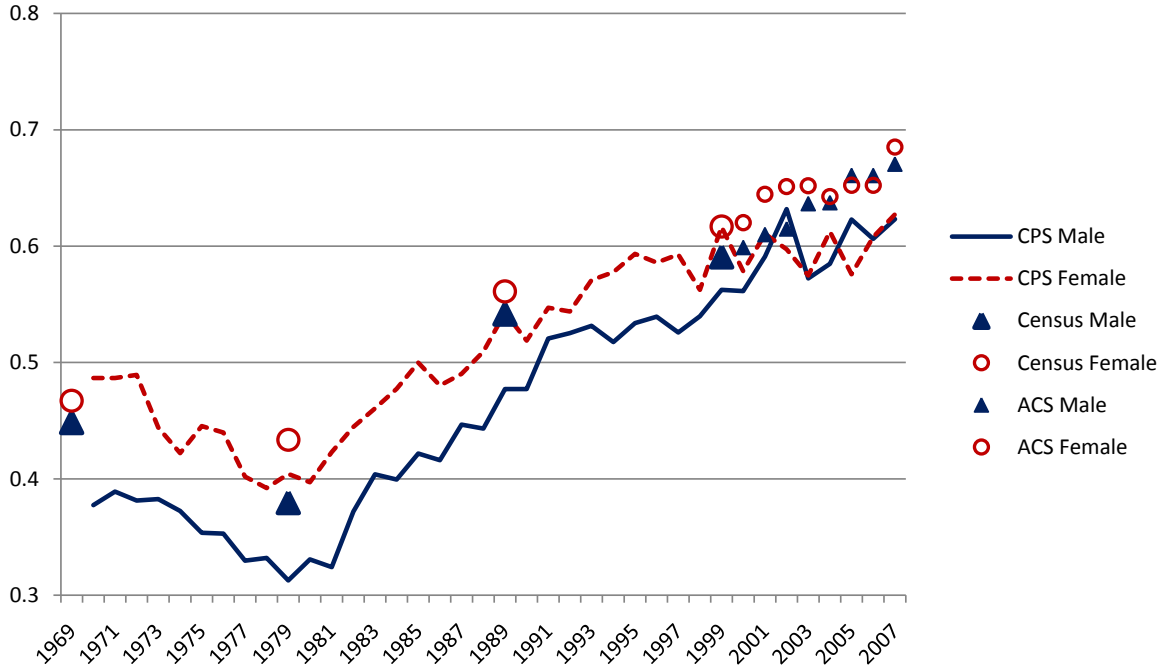


Figure 4: College Wage Premium, Fixed-Weight Median Estimates, by Data Set



1.3 CPS Sample Scope and Estimation Specification

In addition to the specifications I describe in the text, I run OLS regressions and fixed-weights estimates that vary both the CPS sample scope and regressors. These regressions serve two purposes. First, they replicate estimates from the existing literature, allowing a direct comparison of the effect of topcodes and censoring on results reported in other studies. Second, these alternate specifications allow me to test the robustness of my results. The results, which I report in more detail below, vary the scope and specification of the estimation in the following ways:

- using annual, weekly, or hourly wages;
- including all education levels or comparing only college and high school graduates;
- including or excluding Census region dummies;
- pooling all workers, all white workers, or only white, non-Hispanic workers;
- including all workers, only full-time workers, or only FTFY workers;
- weighting all workers equally or weighting by weeks worked or by hours worked (in all cases observations were weighted, at a minimum, by their CPS person weights).

The results are robust across all specifications. I now describe the replications of results in the literature.

1.4 Replication of Existing Results

First, I replicate the results reported in Chiappori, Iyigun, and Weiss (2009) ("CIW") with and without adjustments to topcodes.

Figure 5 approximates Figure 7a from CIW. To create this figure, I use the sample from the main text (herein, "Main Sample"), further limited to white workers age 25-54, with non-negative potential experience, for years 1975-2007. I recensor wages for 1995-2007 at 100,000, then convert all wages to hourly wages in 2005 dollars using the CPI-U, and trim wages below $2/hour or above $200/hour. I regress log wages on a dummy for female sex; education dummies for less than high school graduate, some college, bachelor's degree, and

Figure 5: CIW Wage Premium, by Degree Type, Recensored Wages

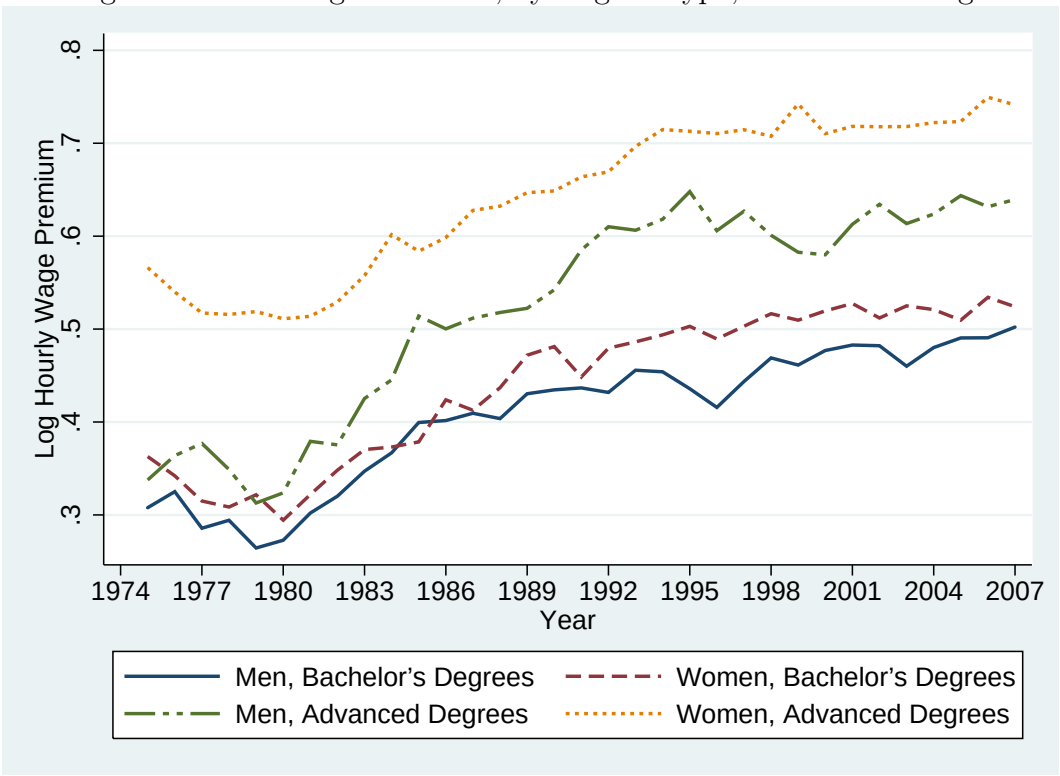
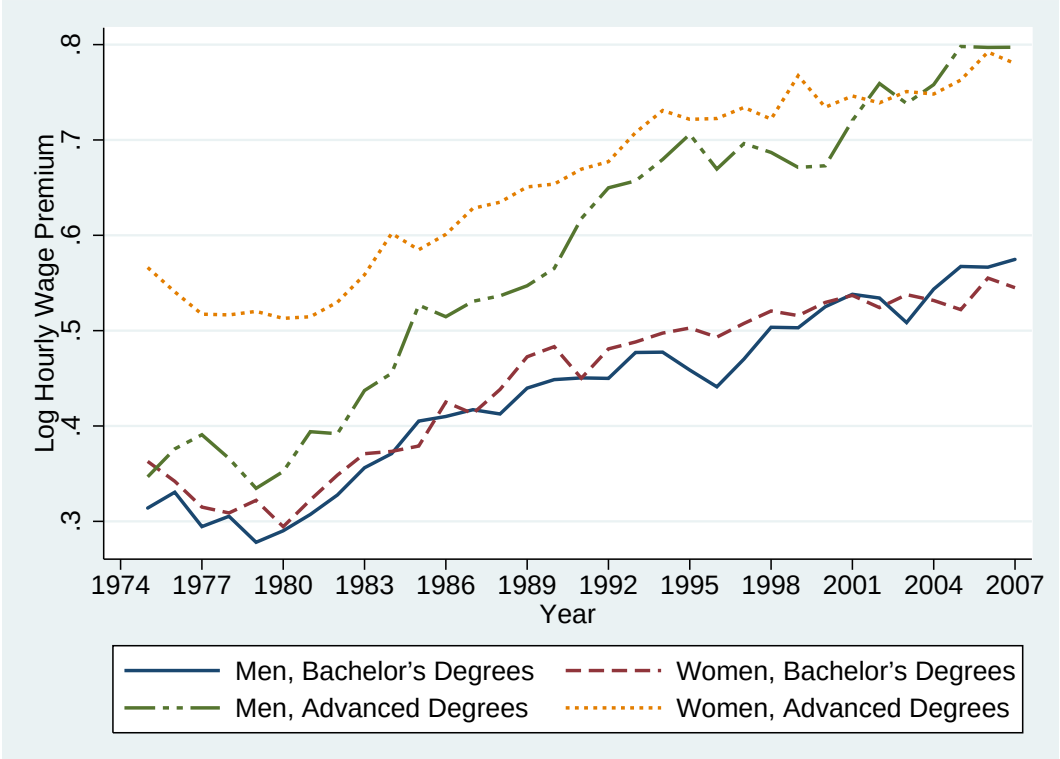


Figure 6: CIW Wage Premium, by Degree Type, Topcodes Adjusted



advanced degree; potential experience and potential experience squared; Census region dummies; and female interactions on all terms. The regression includes a Heckman (MLE) selection correction using marital status and child status dummies as identifying variables. I weight all workers by their CPS sample weights. Figure 6 follows the same procedure, except that I adjust for topcodes.[3] Although the CIW sample, unlike the sample in the text, includes part-year and part-time workers and weights them equally with FTFY workers, this sample yields the same conclusions. Figures 7 and 8 likewise replicate Figure 7b from CIW, with and without topcode correction. The only change in specification is that a single dummy for college graduate (and a corresponding female sex interaction term) replaces the separate dummies and interactions for bachelor's degree and advanced degree.

Next, I consider Card and DiNardo (2000). Figure 9 approximates their Figure 5. To generate this figure, I use my Main Sample, limited to high school graduates and bachelor's degree (only) holders with non-negative experience, for years 1975-2007. I rencensor wages for 1995-2007 at 100,000, then convert all wages to real 1979 hourly wages using the CPI-U-X1, and trim wages below $1/hour or above $100/hour. I regress log wages on a dummy for female sex; a dummy for college graduate; potential experience, its square, and its cube; a dummy for non-white race; and female interactions on all terms. I weight each observation by its CPS sample weight times hours worked (divided by 2000). Figure 10 introduces adjustments for topcoding. While this sample includes non-white workers as well as part-time and part-year workers, the effect of the topcode correction remains the same.

Finally, I consider DiPrete and Buchmann (2006) ("DPB"). DPB look at FTFY white workers age 25-29 and age 30-34 and calculate the log of the ratio of college graduate annual earnings (wage and salary income plus non-farm business income plus farm income) to high school graduate annual earnings. It is important to note here that the estimated college wage premiums for particular age groups may not be representative of premiums estimated using the whole working-age population. As I show here, while the results reported in DPB look appropriate for 25-34 year-olds, they are sensitive not only to topcoding but to the age range of the sample population.

Figure 11 is based on Figure 3 from DPB. For this figure, I used my Main Sample for years 1967-2007, further limited to FTFY white workers age 30-34, who are either high school graduates or college graduates. I calculate earnings as the sum of wage income, non-farm

[3] As with all of the estimations reported in this Part, I recompute adjustment factors using the given sample rather than applying the adjustment factors I derived for sample used in the text.

Figure 7: CIW College Wage Premium, Recensored Wages

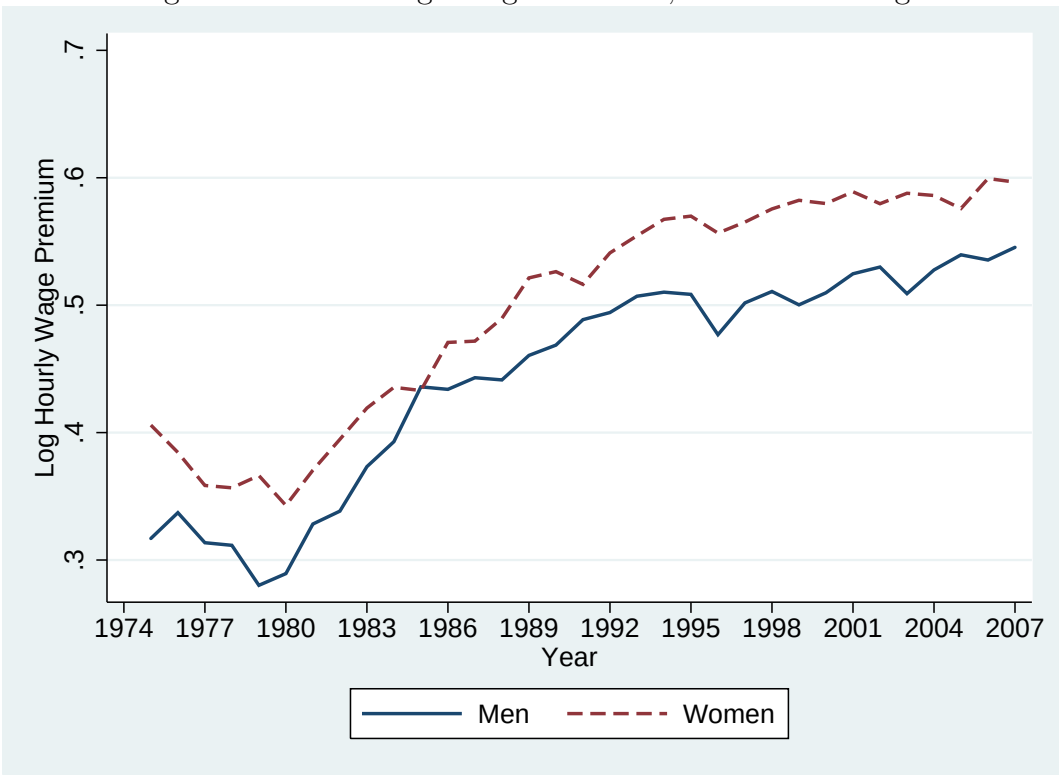


Figure 8: CIW College Wage Premium, Topcodes Adjusted

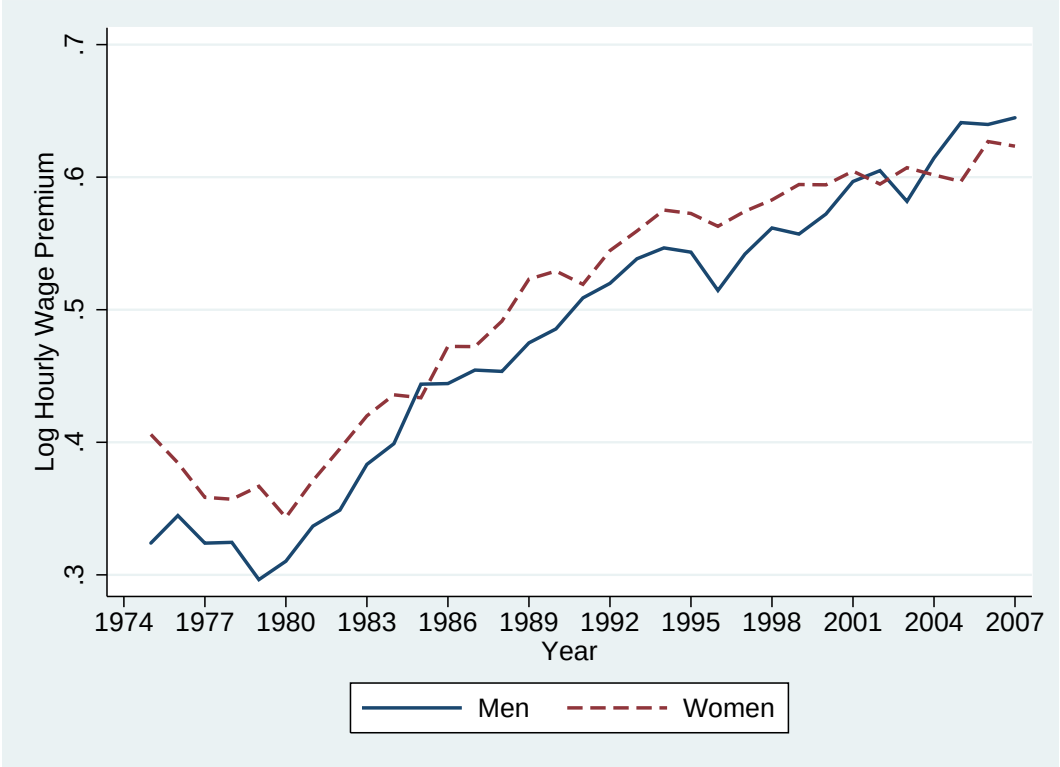


Figure 9: CDN College Wage Premium, Recensored Wages

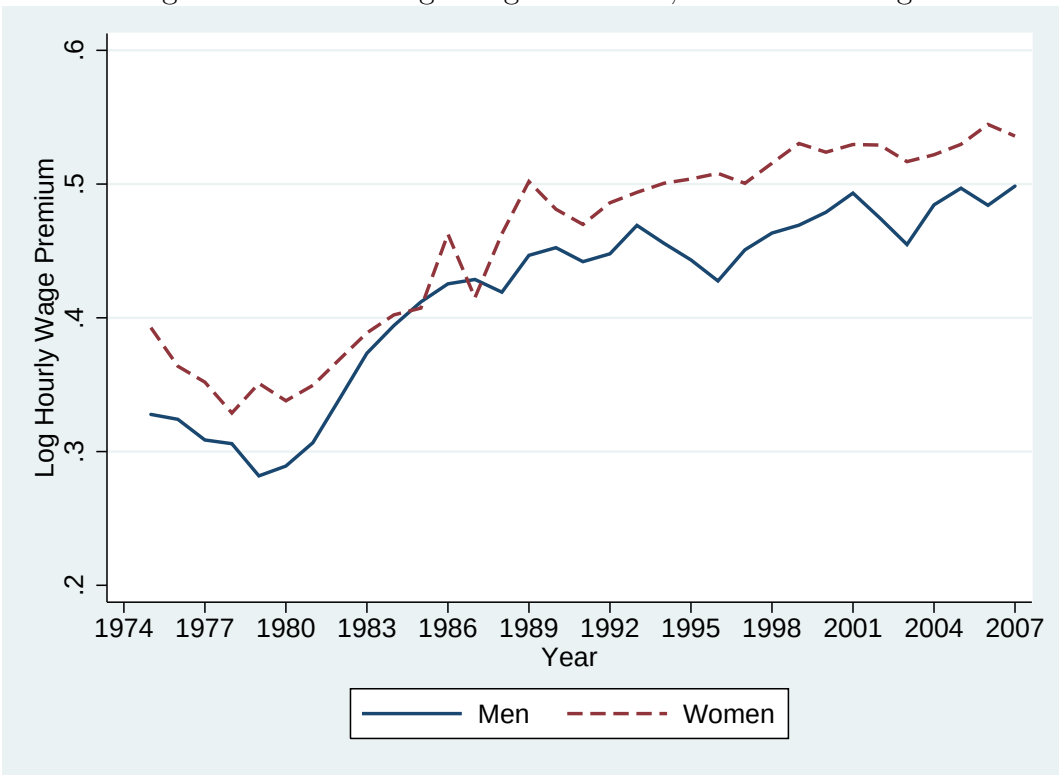
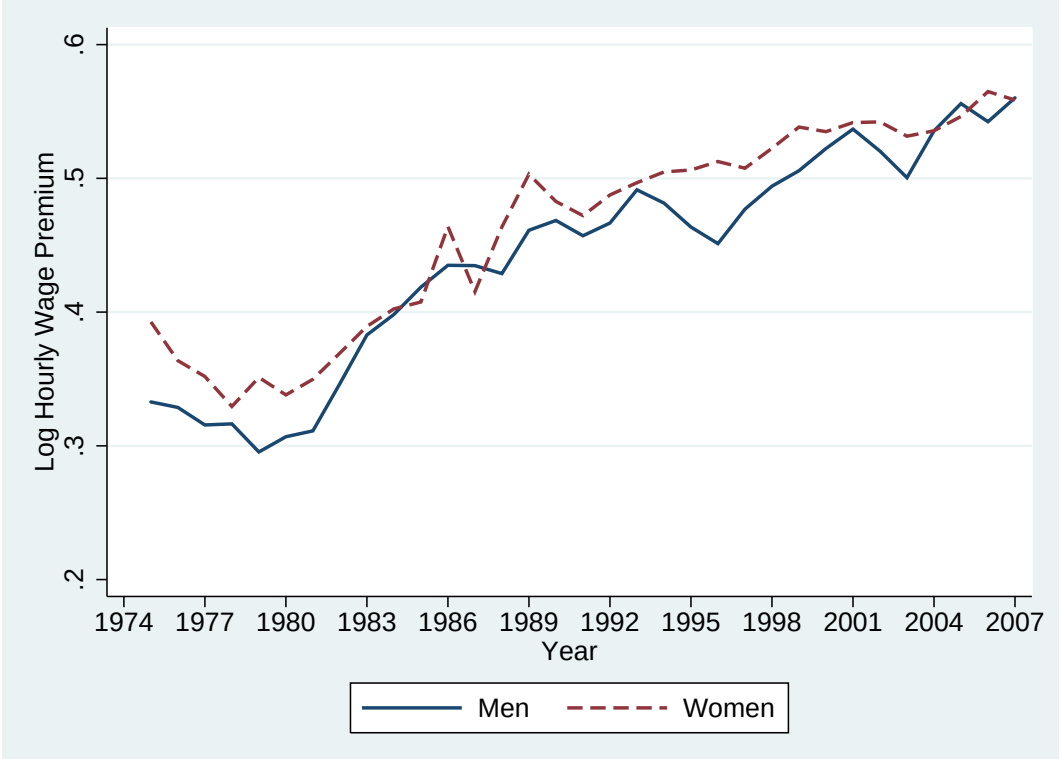


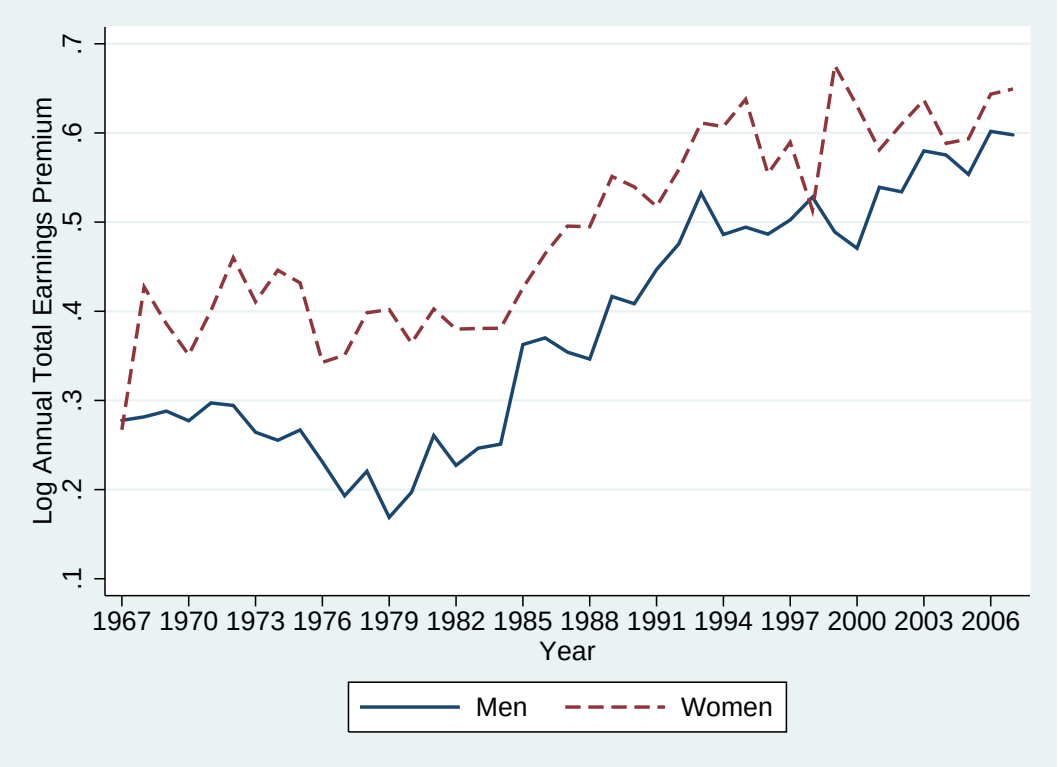
Figure 10: CDN College Wage Premium, Topcodes Adjusted



business income and farm income. I then recensor wages for 1995-2007 at the topcode and regress log earnings on a dummy for female sex, an education dummy for college graduate, and a female interaction term. I weight all workers by their CPS sample weights.

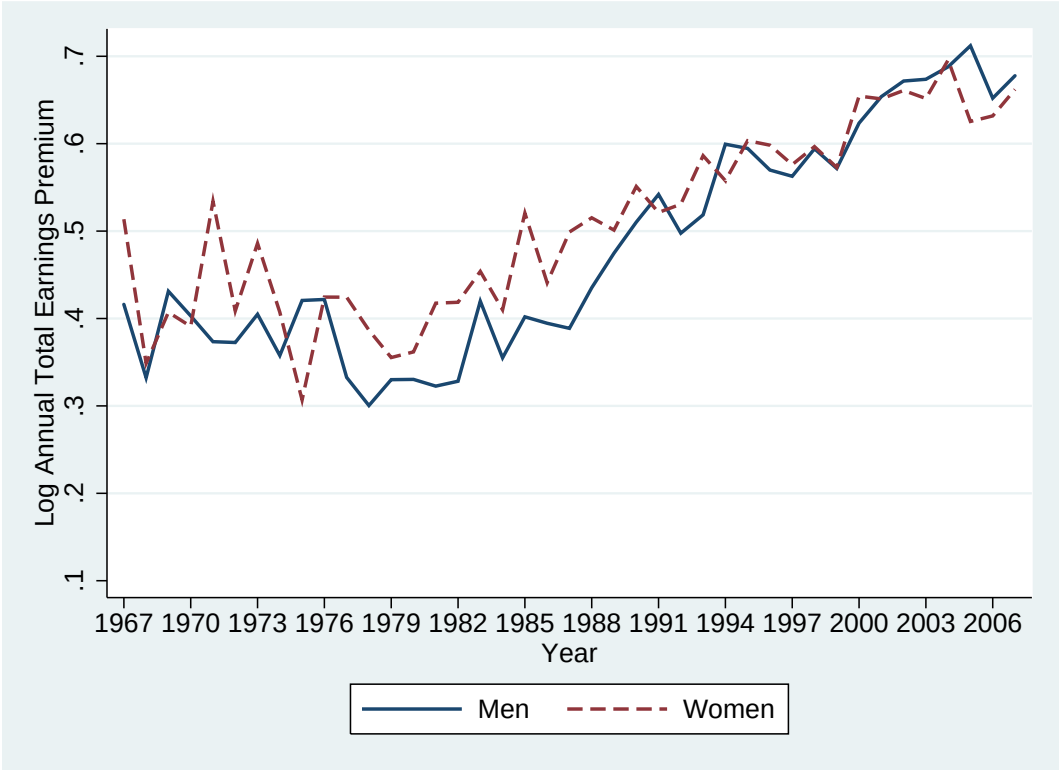
In Figure 11, even after adjusting for topcodes we find an apparent gap in favor of women. As it turns out, however, there is a strong age gradient in the college wage premium, such that there is a positive gender gap in favor of women at younger ages and a negative gap (i.e., a gap in favor of men) at older ages. If we look at a group just 10 years older, the apparent gap disappears. See Figure 12.[4] We see that the presence of a gender difference in the college wage premium depends on the age group considered.

Figure 11: DPB College Wage Premium, Workers 30-34, Topcodes Adjusted



[4]I do not correct for topcoding or bottomcoding of business or farm income. Topcodes and bottomcodes of those categories of earnings represent a miniscule fraction of all observations, and information on conditional means of topcoded observations is extremely sparse.

Figure 12: DPB College Wage Premium, Workers 40-44, Topcodes Adjusted



2 Results for Black and Hispanic Workers

While the results in the text focus on white, non-Hispanic workers, looking separately at only black workers and only Hispanic workers reveals somewhat different patterns. Figure 13 presents results from yearly regressions for black, non-Hispanic workers. Figure 14 presents corresponding results for Hispanic workers. Other than the choice of race and Hispanic status, the regressions are identical to those in the text. (Note that for these demographic groups, adjusting topcodes has relatively little effect, because of the small fraction of topcoded observations in these samples.) Interestingly, for black workers, it appears that the female advantage in the college wage premium has *not* disappeared in recent years. Hispanic workers, however, have a pattern of gender difference similar to white, non-Hispanic workers.

Figure 13: College Wage Premium, Black, Non-Hispanic Workers, Topcodes Adjusted

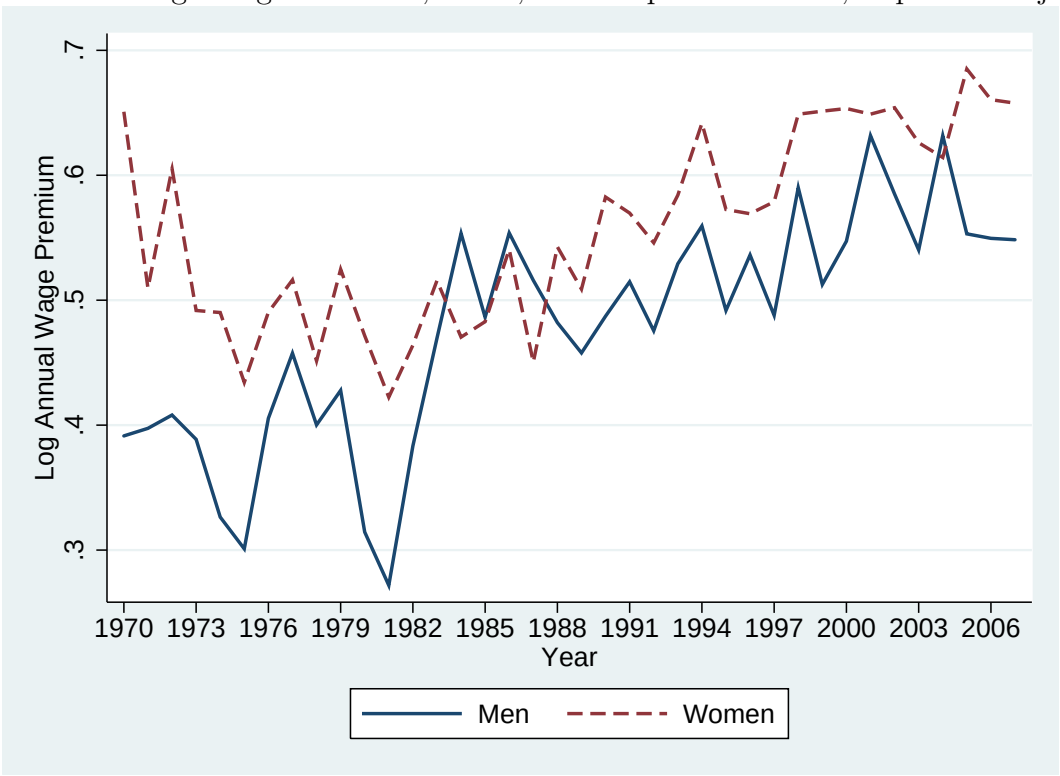
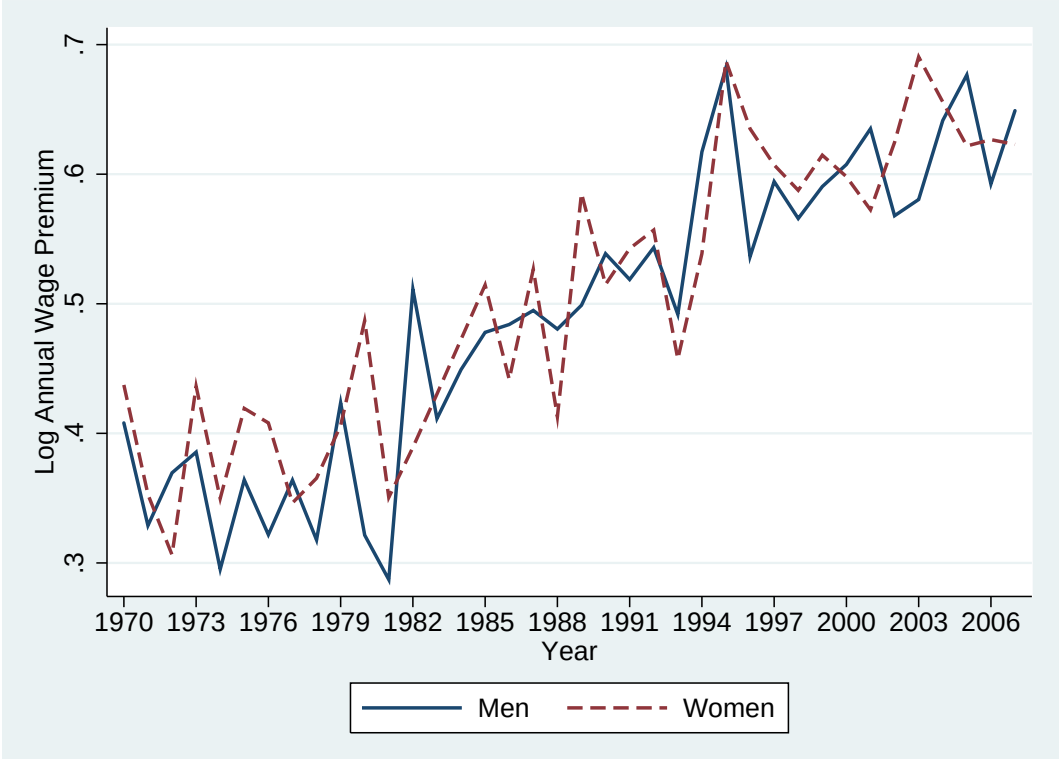


Figure 14: College Wage Premium, Hispanic Workers, Topcodes Adjusted



3 Census and ACS Samples

For the Census and ACS results reported in this appendix, I use the 1970-2000 Census 1 percent samples, and the 2000-2007 American Community Survey ("ACS") data series. See Ruggles, Steven, Matthew Sobek, Trent Alexander, Catherine A. Fitch, Ronald Goeken, Patricia Kelly Hall, Miriam King, and Chad Ronnander. 2008. Integrated Public Use Microdata Series: Version 4.0. Minneapolis, MN: Minnesota Population Center. http://usa.ipums.org/usa/. Specifically, the Census samples are the 1 percent unweighted extracts for each year, including the 1970 Form 2 State sample and 1980 Metro sample within IPUMS. The ACS samples are the individual year data sets for 2000-2004 and the three-year data set covering 2005-2007.

For the analysis of Census and ACS data, I attempt to make the sample definition for these sources as close to that used for the CPS Text Sample, subject to differences in the variables in the IPUMS-USA and IPUMS CPS series. I limit the 1970-2007 IPUMS-USA data as follows: I include non-institutionalized persons age 18 to 65 who are not in school, who are private or government employees for a wage or salary, and who worked last year.[5] I recode demographic information and education categories as in the Main Sample. For the 1990 and 2000 Censuses, and for the ACS surveys, I generate potential experience using the 1988-1990 CPS data as described above.

I further restrict the Census and ACS samples to FTFY, white, non-Hispanic workers with 1-40 years of experience. I deflate all wage values using the PCE index to 1982 dollars, and drop all observations with annual wages less than $3484, or one-half minimum wage in 1982. I also exclude observations flagged as containing "allocated" (i.e., imputed) values for education. Observations flagged as containing allocated values for either the amount or source of wage and salary income are dropped from all calculations. No results are sensitive to the inclusion or exclusion of either type of imputed data.

[5] In IPUMS-USA, the relevant variables are AGE, GQ, SCHOOL, WORKEDYR, and CLASSWKD.

References

Card, David, and John E. DiNardo. 2002. "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles." *J Labor Econ* 20(4):733-83.

Chiappori, Pierre-Andre, Murat Iyigun, and Yoram Weiss. 2009. "Investment in Schooling and the Marriage Market." *AER* 99(5):1689-1713.

DiPrete, Thomas A., and Claudia Buchmann. 2006. "Gender-Specific Trends in the Value of Education and the Emerging Gender Gap in College Completion." *Demography* 43(1):1-24.

Katz, Lawrence F., and Kevin M. Murphy. 1992. "Changes in Relative Wages, 1963-1987: Supply and Demand Factors." *QJE* 107:35-78.