

NOT FOR PUBLICATION

Appendix A. Data and Specifications

This appendix summarizes the creation of the variables used in the analysis as well as the construction of the alternative specifications used for figures 3 and 4. The independent variables, including the key *ELA* measure, are described first, followed by the sequence of dependent or outcome variables. (The dependent variables are available in every wave of the survey unless otherwise stated.) Finally, each alternative specification is discussed.

Age and year of birth

Determining the age of the respondents at each survey is crucial, both in identifying early legal access, which is age dependent, and because the effects of early legal access are likely to vary over the lifecycle. Both age at time of interview and date of birth (month and year) are asked in various waves of the survey; however, they are not always consistent. Date of birth was asked in 1968, 1977, 1978, 1982, 1988 and 1991 and confirmed or corrected in 1995, 1997, 1999, 2001, and 2003. Of the 5,159 women in the sample, 94 (1.8 percent) had conflicting birth date reports, and another 818 (15.9 percent) had only a single report. For the conflicting cases, all available data were used to check birth reports, but, in most cases, the modal reported year and month of birth was used.¹ From the date of birth information, age at the end of each survey year (not at the time of interview) was constructed for consistency between early and later waves.²

State of residence

The geocoded version of the *NLS-YW*, available at Census Research Data Centers, contains the state of residence of each respondent for each wave of the survey. Using respondents' age information and variables pertaining to mover status in the public-use data, one can construct variables for the state of residence at key ages (such as 18, 19, 20, and 21) for most but not all respondents. In some cases, women exit the sample before they reach the key ages; in others, women in the older cohorts who move

¹ The exact code is available from the authors upon request.

² The early waves sampled respondents in the early months of the year but later waves sampled respondents in later months.

frequently during the key ages are not observed until they are older. Nonetheless, for each of the key ages (18 through 21), between 80 and 90 percent of the respondents were successfully matched to a state of residence.

Early Legal Access to the Pill (ELA)

By researching state laws, the authors compiled a list of the years in which each state legally allowed unmarried women (of age 20 or younger) to have access to the birth control pill (see Appendix B: Legal Variables). Using the restricted version of the *NLS-YW*, state of residence at each survey is observed and the respondents' state of residence at age 21 is used to generate the *ELA* variable. A respondent's *ELA* status was coded 1 if her year of birth plus 20 was greater than or equal to the year in which her residence state at age 21 first allowed legal access. State of residence at age 21 rather than age 20 was used because it was identifiable for more women (4,419 versus 4,398) and the correlation between the two was high ($r = 0.94$).

Early Abortion Access (EAA)

Five states (Alaska, California, Hawaii, Washington, and New York) and the District of Columbia legalized abortion in 1970, three years before *Roe v. Wade*. We code a respondent as having EAA if she lived in one of the above areas at age 21 *and* was born in 1950 or later; these are the cohorts of women who had legal abortion access in their states of residence before the age of 21. To address the possibility that women crossed state lines to obtain an abortion, we also constructed a measure of the distance in miles between each state's population centroid in 1970 and the closest major location providing abortions in the pre-*Roe* period (District of Columbia, Los Angeles, San Francisco, Buffalo, and New York City). This distance was then transformed into its natural logarithm.

Age at first marriage

Although age at first marriage is directly asked in 1968, this is useful only for women who had been married prior to the first interview. To determine marital ages for the rest of the sample, three additional sources are used: (a) marital histories, (b) changes in current marital status, and (c) timing of changes in marital status. Marital history questions are asked in 1978, 1983, 1997, 1999, 2001, and 2003.

In 1978 and 1983, the questions ask about up to the three most recent marriages (including the current one); in the latter years, only the date of the most recent marriage is asked. Current marital status is asked in every survey year. Changes in marital status are reported in 1969 and 1970 and every survey year from 1985 onwards. We observe no first marriage date for 809 women. This outcome is used only in Appendix C: Additional Estimates and Sensitivity Checks.

Wages and salary earnings

Hourly rates of pay for the current or most recent job (measured in cents) and annual wage and salary earnings from the previous calendar year are available for years 1968 through 1993. For 1995 through 2003, the hourly rate of pay variable is for the first (main) job, and annual wage and salary earnings are for the previous 12 months rather than the previous calendar year. Information on wages and salary earnings excludes farm, business, or self-employment income. Each of the wage, earnings, and income variables is converted from nominal to 2000 dollars using the PCE deflator and then converted to natural logarithms. Although there is no effective top code to hourly wages, annual earnings are subject to censoring from above, with the top code varying across years. (Generally, fewer than 2 percent of women have top-coded earnings in any year.) In the analysis, hourly wage outliers (less than 2 or more than 100 real dollars) are excluded.

Cumulative experience

We measure cumulative work hours at the start of each calendar year as the sum of hours of work reported since 1967. We approximate hours of work with the product of usual weekly hours and our best estimate for the number of weeks worked each year.

We rely on three sets of questions to compute number of weeks worked. In 1968, 1969, 1975, 1977, 1980, 1982, 1985 and 1987, respondents were asked to report the number of weeks they worked in the previous calendar year. In 1970, 1971, 1972, 1973, 1978, 1983, 1988, 1991 and 1993, the survey asked the number of weeks worked since the last eligible interview, regardless of whether or not that interview took place. In 1970, 1971, 1972, 1973, 1995, 1997, 1999, 2001 and 2003, the survey asked weeks worked since the last actual interview. We combine these measures as available, being careful to

avoid double-counting. (This procedure is complicated and idiosyncratic to each survey wave; the code used is available upon request.)

Despite our best efforts, we note that it is not possible to create a truly comprehensive measure of weeks worked for several reasons. First, there are some gaps in coverage for which no weeks worked questions were asked: The initial shift from calendar year to survey period leads to a small time period (generally under 6 weeks) for which we have no measure of weeks worked. The size of this coverage gap increases over time. For example, we miss nine to eleven months between the 1973 interview and January 1, 1974, and the entire calendar year of 1975. Second, item non-response for a question regarding weeks worked poses a significant problem because cumulative experience is dependent on all past responses. It is only possible to recover cumulative experience for women who miss an interview and are subsequently re-interviewed *if* the later interview asks about weeks worked since the last actual interview.

Our main measures address these concerns with additional sample restrictions or assumptions. We address the coverage issue by rescaling the experience measure to a base of full coverage. We effectively assume that the fraction of weeks *observed* working is the same as the fraction of weeks *elapsed* spent working; that is, we scale the cumulative weeks worked measure by the ratio of total weeks elapsed to total weeks for which there is coverage. For the second problem, we exclude women once they have an episode of an item non-response for the weeks worked question. For the third problem, we restrict estimation to women who have a valid weeks report in every survey wave (no missed interviews and no item non-response). None of these alternate measures, whether used individually or all together, changes the qualitative pattern of results we find of *ELA* on cumulative experience. The numbers and estimates reported in table 4 apply the first and second measures but exclude the third in the interest of maintaining a larger sample size.

College enrollment

Using questions that asked about current enrollment in an academic program of study, as well as the highest grade completed, a respondent was coded as enrolled in college (a binary variable) if she was enrolled and the highest grade completed was at least 12. As a result, “college enrollment” includes all

forms of academic post-secondary education but excludes vocational/occupational training. Note that women who did not graduate from high school are excluded (coded as missing).

Highest grade completed

The basis of these variables is the set of revised highest grade completed questions. Although the “revised” set has supposedly been cleaned and corrected of errors found in the original highest grade completed questions, an inspection revealed that several problems remained, and these were often some form of non-monotonic progression. Five hundred thirteen women (10.0 percent) had at least one discrepancy, but in most cases these were minor, such as a jump up or down of one grade in a single survey wave before returning to trend. The “revised” variables were cleaned further of likely misreports using responses from previous and later years. Specifically, “jump” deviations that last only a single wave (in some cases, two waves) are smoothed by replacing these values with those that occur both before and after the deviation. For example, a woman whose highest reported grade is 12 in 1975 and 1977, 10 in 1978, and 12 in 1980 and 1982, would have the 1978 value recoded to 12. This procedure leaves 205 women (4.0 percent) with a non-correctable discrepancy, such as multiple, non-monotonic jumps; these respondents are flagged and excluded from the analysis. Including these women alters the results very little.

Labor-force participation

Labor-force participation (LFP) is based on the employment status recode (1968 through 1993) or monthly labor recode (1995 through 2003) variables. The LFP dummy variable takes the value of 1 if the respondent is employed at the time of the survey (whether at work or not) or unemployed, and 0 otherwise. Note that choice of specific activities in the survey for non-labor-force participants changed between 1993 and 1995, when the *NLS-YW* adopted the new *CPS* definitions. Results using this measure are reported in footnote 27.

Usual weekly hours

These variables are based on a question asking about the usual hours worked per week at the respondent’s job. For most years, the job is defined to be either the one currently held or the job most

recently held since the last interview; however, in 1970, 1971, 1972, 1973, 1978, and 1983, the question pertains to the current job only. In these cases, another question specifically referring to the usual hours worked at the most recent job is used to supplement the current job question to maintain comparability: Respondents with missing values for the current job only question are replaced with the usual hours worked from the most recent job question. Finally, because responses in some years are top-coded at 99 hours while some are not, values above 99 are recoded to exactly 99. This affects no more than 1 to 3 women in any year and has a negligible impact on the estimates.

Occupational training

Although the *NLS-YW* asks several questions throughout the survey waves about occupational training, the questions are not completely consistent across waves. In 1968 and again from 1980 through 2003, the survey asked whether respondents had undergone (a) any on-the-job training since the last interview, and (b) any other occupational or vocational training. From 1969 to 1978, however, these two different types of training were co-mingled in a single training question. For consistency, both training types are combined into a single (binary) indicator that captures whether the respondent underwent any form of vocational or occupational training, on-the-job or otherwise, since the last interview. The estimation sample for training includes only respondents who were not currently attending an academic program, because training questions were asked only of respondents not enrolled in an academic program until 1975.

Occupation

For each wave of the survey, there is a variable containing the 3-digit Census code of the respondent's current or most recent job. Through 1993 the variable is for current or most recent job; for 1995 through 2003, when the new (circa 1994) *CPS* definitions were used, the variable for job 1 (the main job) is used. Unfortunately, a consistent coding is not available in the data. The coding at the beginning of the survey is based on the 1960 scheme, and it is available through 1993. Coding based on the 1980 scheme begins in 1980 and runs through 1999; the 1990 scheme runs from 1993 through 2001; and the 2000 scheme runs from 1995 through 2003. Thus, there is significant overlap for several years. In the

interest of creating a longer series, we aggregate the different coding schemes by collapsing the 3-digit job codes into four groups that can be made consistent over the entire time period. We use a coding scheme as soon as it becomes available, so we use the 1960 scheme for data years 1968 through 1978, the 1980 scheme for years 1980 through 1991, the 1990 scheme in 1993, and the 2000 scheme for years 1995 through 2003. The four groups are: all professional and managerial jobs, non-traditionally female professional and managerial jobs, clerical and sales jobs, and all other jobs. “All professional and managerial jobs” generally includes any 3-digit code that falls under the “professional, technical and kindred workers” or “managers, officials, and proprietors except farm” categories (or their equivalent) from any of the coding schemes. “Non-traditionally female professional and managerial jobs” is a subset of the first category that excludes the traditionally female occupations of nurses and elementary, secondary, and not elsewhere classified (n.e.c.) teachers. “Clerical and sales jobs” includes 3-digit codes listed under the clerical or sales categories, and “all other jobs” includes all 3-digit codes not in one the previous groups, including craftspeople, operatives, agricultural workers, and service jobs. The complete list of 3-digit Census job codes to our four groups by coding scheme is available by request. For the analysis in table 4, a woman must be currently employed to be counted in one of the four job groups; if she reported a 3-digit code in the survey but also reports not being currently employed, we code her as a zero in all four job categories.

IQ and Childhood Family Socioeconomic Status

The 1968 wave of the *NLS-YW* included a questionnaire for the high schools of the respondents, which in addition to asking about school characteristics also asked for the most recent intelligence or aptitude test of the respondent. Scores were reported for 3,530 of the respondents (though almost none for respondents born in 1953). See Griliches, Hall and Hausman (1978) for an assessment of whether scores are missing at random in the *National Longitudinal Survey of Young Men*. The agency that processed the *NLS-YW*, the Center for Human Resource Research (CHRR), converted these scores from various tests composites to a unified “IQ score” based on a normally-distributed national population with mean 100 and standard deviation 15. (More information on this procedure can be found at

<http://jenni.uchicago.edu/evo-earn/IQ.pdf>.) Based on this distribution and the unified score, a respondent was also classified into an IQ quantile and stanine. Using information from the initial survey wave on father's occupation and education, mother's education, eldest sibling's education, and availability of reading material at home, CHRR also constructed a summary family socioeconomic status variable to follow a normal distribution with mean 100 and standard deviation 30. Our analysis breaks these measures into tertiles.

Attrition

In most cases, the empirical analysis has made no attempt to restrict the sample to non-attriters. The decision to exploit every person-year observation was made in order to maximize sample size. One of our sensitivity checks, reported in figures 3 and 4 and in tabular form in appendix C, shows that findings based upon a balanced panel of individuals are very similar to those reported in the paper. In addition, regressions, available upon request, show no correlation between each year's interview status and *ELA*.

Variables Used in Table 1 Balancing Tests

- (1) **Father worked for pay:** binary variable equal to one if a respondent's father worked for pay when respondent was 14. About 93 percent of the sample had a father working for pay at age 14. (Note: This is *not* conditional on having a father in the HH).
- (2) **Father held professional job:** binary variable equal to one if a respondent's father had a "professional" job when respondent was 14. "Professional" has the same coding as in the main results, based on 1960 occupational definitions. About 20 percent of the sample had a father working in a professional job. (Note: This is conditional on having had a father working at age 14).
- (3) **Mother worked for pay:** binary variable equal to one if a respondent's mother worked for pay when respondent was 14. This was *not* asked of respondents who lived with their mother as the sole parent. About 39 percent of the effective sample had a mother working for pay at age 14. (Note: This is *conditional* on having a father (or other male adult) in the HH).

- (4) **Mother held professional job:** binary variable equal to one if a respondent's mother had a "professional" job when respondent was 14. "Professional" has the same coding as in the main results, based on 1960 occupational definitions. About 13 percent of the sample had a mother working in a professional job. (Note: This is conditional on having had a mother working at age 14).
- (5) **Duncan index of household head:** Duncan index socioeconomic job score of head of household when respondent was age 14, as created by CHRR in the data. Values are conditional on the head (not necessarily father) working when respondent was 14. (The scale runs from 3 to 97).
- (6) **Socio-economic status:** socioeconomic index of respondent's parents in 1968, as provided in the data. Based on father's occupation and education, mother's education, eldest sibling's education, and availability of reading material at home. By construction, $SES \sim N(100,900)$.
- (7) **Magazines in home:** binary variable equal to one if a respondent had magazines available at home when she was age 14. About 64 percent of the sample did.
- (8) **Newspapers in home:** binary variable equal to one if a respondent had newspapers available at home when she was age 14. About 83 percent of the sample did.
- (9) **Respondent held library card:** binary variable equal to one if a respondent had a library card when she was age 14. About 70 percent of the sample did.
- (10) **Two-parent household:** binary variable equal to one if a respondent lived in a household with two parents (including step-parents) at age 14. About 80 percent of the sample lived with two parents at age 14.
- (11) **Number of siblings:** number of siblings of respondent in 1968 (not necessarily in the household); we can't reliably determine whether this includes step- and half-siblings.
- (12) **Father born in U.S.:** binary variable equal to one if a respondent's father was born in U.S./Canada. About 96 percent of sample had father born in U.S./Canada.

- (13) **Highest grade completed by father:** highest grade completed by father, in 1968. Conditional on having a father in household. Item non-response is relatively high; *ELA*, however, is uncorrelated with whether father's HGC is observed.
- (14) **Highest grade completed by mother:** highest grade completed by mother, in 1968. Conditional on having a mother in household. Item non-response is relatively high; *ELA*, however, is uncorrelated with whether mother's HGC is observed.
- (15) **Parents' education goals for respondent:** number of years of schooling respondent's parents want respondent to obtain, when respondent was 14.
- (16) **Atypicality index of mother's job:** atypicality index of respondent's mother's job when respondent was 14, conditional on respondent's mother working then. Atypicality index is the female percentage of an occupation minus the percent of the experienced civilian labor force that was female in 1970; negative numbers indicate more atypical occupations.
- (17) **Respondent's IQ score:** continuous IQ score of respondent. Reference distribution is independent national norm, not empirical sample. Only two-thirds of the entire sample had an IQ or achievement test administered; while these two-thirds were slightly above national norms, the presence of an IQ score is uncorrelated with *ELA*.
- (18) **Rural residence:** binary variable equal to one if a respondent resided on a farm/ranch or in another rural area at age 14. About 26 percent of the sample lived in a rural area at age 14.

Alternative Specifications

Figures 3 and 4 include six specifications: one following equation (1) called our baseline specification, one following equation (1') that augments our baseline specification with abortion controls, and four alternative specifications of (1') described below. Tabular presentation of estimates from equation (1') are presented as the main tables of the paper. Tabular presentations from all other specifications can be found in Appendix C.

Linear state-specific time trends: The specification in equation (1') is augmented with the interactions of each state of residence dummy with the year of observation.

Vietnam casualties: Using data from the National Archives on the Vietnam Conflict (<http://www.archives.gov/research/military/vietnam-war/electronic-records.html>), the specification in equation (1') is augmented with controls for state-level casualties. These controls include state-specific annual death rates lagged one, two, and three years; and cohort-specific, state-level death rates within two years of a woman's date of birth.

Balanced panel: The specification in equation (1') is estimated on a sample that is restricted to women who are interviewed in every survey wave from 1968 through 2003 and successfully answer all relevant questions (no item non-response).

High school state: This specification uses state of residence during high school (rather than at age 21) for all state-based variables. Like state of residence at age 21, this variable is created using each wave's state of residence, move histories, and tenure at current residence. Because older cohorts are farther removed from high school age, they are less likely to be successfully matched, particularly if they moved frequently. (While this problem exists for state of residence at age 21, it is more pronounced for high school state.)