Developing Synthetic Worklife Earnings Estimates

Robert Kominski
Tiffany Julian
Housing & Household Economics Statistics Division
U.S. Census Bureau

INTRODUCTION

One well-studied topic of social science is the relationship between education and earnings (Card, 1998). Particularly in an achievement-based society, examining and understanding this basic relationship is one of the most fundamental research and policy issues there is desire to understand. In the universe of analyses conducted over numerous decades, researchers have attempted to identify and isolate various unique characteristics of individuals which may in some way mediate this basic relationship. These items include not only demographic factors such as gender, race and age (Day and Newburger, 2002), but other individual qualities, such as family characteristics, ethnicity, language ability, quality of schooling, and physical disability (Kerkhoff, 1995; Loprest and Maag, 2003). Extending forward, elements such as the specific field of training and industry of employment, geographic location and other economic characteristics (Hecker, 1998; National Science Foundation, 2005) have all been examined at some level, to determine their effects on the basic education-earnings relationship.

This project represents the first steps in developing a set of synthetic estimates of the earnings one might derive over a typical working life, given the level of education the individual has. This is not a new idea – other analyses precede this one, including a fairly comprehensive Census Bureau report (Day and Newburger, 2002). In this analysis, however, we examine a variety of extensions of the basic analysis, looking for factors of potential mediation, and trying to find ways to estimate their impact.

There are a variety of technical, measurement and methodological issues we must explore and resolve on the way to what we hope will be a final set of estimates. This work is the focus of this paper, and some of the main issues to explore are outlined in the next sections.

DATA

The data for this research comes from the Multiyear American Community Survey (ACS) data file for the period 2006-2008. The ACS represents a part of the Census Bureau's revised approach in how it conducts the federally-mandated Decennial Census of the Population of the United States. Prior to implementation of the ACS in 2005, a growing

2

array of topic-oriented data has been a part of each census since the first U.S. census in 1790. Over time, the once-a-decade Decennial Census data collection effort has increasingly obtained information not only on the basic demographic age-race-sex variables necessary for the apportionment activity specified by the Constitution, but on a wide array of other topics. Since the 1960's this collection of additional information has been accommodated through the use of 'short' and 'long' form questionnaires, with all households answering the items on the short form, and just a sample of household answering the more detailed long form questionnaire.

In 1997, the Census Bureau began the process of moving all of the detailed (long form) questionnaire content beyond that needed for apportionment into a continuous ACS program. The ACS is a large monthly national survey of the U.S. population (including Puerto Rico), that works to get data from about a quarter million households each month. (Note: Beginning in 2006, persons in group quarters residences also became a routine part of the sample.) All data collected in a given calendar year are brought together and then weighted to the independent population estimates for July 1 of that year, in order to provide nationally-representative data on the full long form content on a yearly basis (instead of once every ten years.) In order to provide estimates for very small pieces of geography and subpopulations, the Census Bureau takes sequential yearly files, combines them, and produces 'multi-year files' with much larger samples. This analysis uses a multiyear file for the 2006-2008 period in order to provide sufficient characteristic detail for the analysis.

All earnings estimates have been adjusted into the calendar year 2008 value, and all variables used in the analysis have been reviewed to ensure they all use comparable analytic categories across the three years of data, 2006-2008[1].

ANALYSIS ISSUES AND METHODS

There are two basic methods undertaken in the analysis. First, a large multi-way cross-tabulation is constructed for the relevant education and demographic variables. Then, the yearly average earnings are computed for each 'cell' of this table. A second analysis

---

[1] For more information see the ACS Design and Methodology report available on the web at : http://www.census.gov/acs/www/SBasics/desgn_meth.htm

uses these same variables to estimate yearly mean earnings via a regression approach. In both cases, each cell, or 'predicted value' is only appropriate for the specific age group for which it is computed. The 'synthetic' part of the analysis is to take the age-specific values and sum them up across the period of time we define as a 'worklife.' The resultant sum defines the overall value which that particular combination of factors would yield, were the relationship between the demographic variables and earnings (as well as inflation itself) to remain constant over time.

In order to proceed with the analysis, there are several general questions/issues that need to be addressed and resolved. These are discussed below:

Question One: What is the basic relationship and how are the items defined?

We have chosen to use the educational attainment question from the ACS, as well as the earnings items on the survey; these questions are reproduced in Figure 1. Education is assessed by means of a sixteen category item, where the respondent chooses the 'highest degree or level of school completed.' This item provides completion levels ranging from no formal education to completion of a Doctorate degree. This question does not attempt to measure ALL education, just the self-reported highest level. In addition, the item does not assess education outside the 'regular' schooling system, so levels of attainment/ competency such as certifications and licensures are not directly measured. Even this 'limited' variable, however, has sixteen distinct values. Preliminary analysis of yearly earnings across levels of education showed small levels of variation between several of the categories, especially in the range below high school completion. Therefore, this analysis uses nine distinct education categories.

Earnings are defined as the wages and salaries received from all jobs worked in the twelve months prior to interview. This is determined through the sum of two questions— one on "wages, salary, commissions, bonuses or tips from all jobs"; the second on "self-employment income". We make no attempt in this paper to estimate income from other sources such as investment or public assistance. Likewise, we do not measure forms of wealth such as property ownership or savings.

Question Two: What is the appropriate analytic universe?

Analyses of education and earnings are often limited to those persons thought to have the 'purest' form of the relationship, defined as persons who are engaged in 'full-time, year-round' work. In other words, only those persons who held a job the entire year and also worked in a full-time capacity are the basis of those analyses. In some regards, this is a reasonable restriction. However, many persons do not enjoy this status throughout their career. People routinely change jobs, leave the labor force for some period of time, or engage in less than full-time employment—either by choice or because of other labor market issues. In fact, the likelihood of being a 'full-time, year-round' worker is correlated with level of education.

In this analysis we have decided to use three different universes for computation: persons employed 'full-time, year-round' (during previous twelve months); persons with some employment, but less than 'full-time, year-round' (in the past twelve months); and all persons, regardless of employment status. The value in examining all three groups is to provide an analysis where the employment context is at its strongest ('full-time, year-round'), contrasted with a group less tightly attached to the labor force, or to a complete accounting of the education/earnings relationship for the full population.

A different dimension in universe determination focuses on the age range of concern. This analysis focuses on the earnings over one's "work life". In practical reality the age range associated with work status varies considerably across the population. So, some choice must be made to define a standard 'size' work life that accounts for the age range that is most reasonable for the population. We choose the age range of 25-64. This effectively eliminates younger persons, many of whom are still in school, or are at the beginning stages of developing their career and maximizing wages. It also includes at the upper end some individuals who stop work before the end of the range (age 64) and effectively have no or very low earnings.

Question Three: Do tabular and regression methods produce similar results?

Analyses such as this require large amounts of data to develop proper estimates, especially if earnings are to be examined across multiple dimensions of disaggregation. One traditional method of computing lifetime earnings uses a large cross-tabulation to

define average earnings within each cell, then 'sums up' these values across the lifetime. A different method is to construct a regression model that parameterizes each of the variables and categories of interest, and then compute lifetime earnings by simply computing the estimated values for different combinations of the model specification. The tabular method has the advantage of an exhaustive combination of interactions between all demographic factors, but as variables are added the table can become cumbersome and cell sizes can become small. The regression method allows more factors to be included in the estimate at the expense of exhaustive interactions.

Since we will conduct the analysis using both a tabular method and a regression approach, one key is to use comparably-coded variables in each part. These variables are as follows:

EDUCATION (nine categories): None-8th grade; 9th-12th grade (no diploma); High School Diploma; Some College; Associate's Degree; Bachelor's Degree; Master's Degree; Professional Degree; Doctorate Degree.
As discussed earlier, the full sixteen category item produces both a sizable overall table, plus the differences in earnings at lower education levels are small. For these reasons we have collapsed to a more efficient set of education levels.

RACE/ETHNICITY (five categories): Hispanic; White Not Hispanic; Black Not Hispanic; Asian Not Hispanic, Other Not Hispanic. (NOTE: Referred to hereafter as: Hispanic, White, Black, Asian, Other.)
The Census Bureau collects Race and Hispanic origin as two separate data items. In the race question, respondents may choose as many races as they wish. Over time, analyses have demonstrated that many persons choosing a Hispanic origin also choose 'white' as their race. Consequently, in order to optimize cell considerations, we use a single race/Hispanic ethnicity variable which focuses on three single races, non-Hispanic choices (White, Black and Asian) a single Hispanic category (superseding all race choices) and a fifth 'Other' category (including all multi-race choices). The tradeoff here again is to optimize sample sizes within the cells of a very large cross-tabulation.

GENDER (two categories): Male; Female

EMPLOYMENT STATUS (three categories): 'Full-time, year-round' during the previous 12 months; any employment less than 'full-time, year-round' in the previous twelve months; All Persons.

Figure 2 gives a good idea of the size of these three groups across education levels, as well as an estimate of the median earnings within each group. As can be inferred from the size of the bars and relative yearly earnings, the choice of work status universe can have a sizable impact on the estimated values.

AGE (eight categories): 25-29 yrs; 30-34 yrs; 35-39 yrs; 40-44 yrs; 45-49 yrs; 50-54 yrs; 55-59 yrs; 60-64 yrs.

The term "worklife" may connote different things for different people. We chose this 40-year range as being best representative for the estimation purpose here. Obviously many people begin work before age 24; others end work before age 64; still others work well beyond age 64. Starting at age 24 effectively accommodates the large portion of the population whose education is complete by that point in time. The choice of a standard 40-year worklife provides consistent comparability but may also exclude some variation in how long people work in their lives.

In computing the Synthetic Worklife Earnings (SWE) estimates using the tabular method, we construct a table of median (or mean) earnings for the table of Age (8) by Gender (2) by Race/Ethnicity (5) by Education (9) for each of the three Employment Status groups. This yields three 720-cell tables which can be summed across the eight Age categories (each multiplied by 5) for any specific Gender—Race/Ethnicity—Education combination to yield the SWE for that group.

In the regression context these same variables are used in an OLS estimation of mean annual earnings. The regression utilizes dummy (values 0 and 1) variables to proceed with the estimation in the same manner as the tabular approach. While variables such as

Age and even Education are routinely used as continuous-level variables, we use a dummy variable approach to maintain comparability with the tabular method. As with the tabular approach, separate models are estimated for each of the three Employment Status universes. The SWE for any given Gender—Race/Ethnicity—Education combination is then accomplished by plugging in the appropriate codes for each group and the eight relevant Age levels, multiplying by 5, and summing the components.

Question Four: Does SWE estimation using median earnings yield different results than mean earnings?

Figure 3 shows the annual mean and median earnings for nine different education levels. Mean earnings estimates are always higher, due to the influence of the extreme upper tail of the distribution, which the median effectively accounts for. For professional degree holders, this difference is quite sizable – in the range of $40,000. The tabular approach affords us a clear method for estimating the effect of using median or mean earnings. The large detailed data described above will be constructed using both medians and means, and the corresponding SWE are derived. These results will give us a better idea of the impact of using mean or medians.

Question Five: What is the additional impact of demographic, personal and other factors?

While the basic relationship considered in this analysis is that of education and earnings, there is little doubt that this relationship is routinely mediated through other factors. Standard demographic factors such as gender and race are demonstrated variables which have an impact on earnings, independent of education (Bianchi 1996, Siegel 1965). However, beyond these basic demographic components, are there other factors which may result in a significant impact on synthetic worklife earnings?

Adding additional factors to the tabular method creates tables that are ever larger, spreading sample farther, and potentially weakening the results. For this analysis, we limit the inclusion of additional factors to the regression-based analysis. Three additional factors are considered: Citizenship Status (three categories- noncitizen; naturalized; US by birth); Language (three categories – English only; Language Other Than English and

speaks English "very well"; Language Other Than English and speaks English less than "very well"); and Census Regions (four categories – North; South; Midwest, West). As with the initial regression analysis, these other factors are parameterized as a series of dummy variables.

ANALYSIS

As outlined above, the analysis uses the 2006-2008 Multiyear ACS data file. This is a sizable data file – a total of 5,837,976 households and 13,676,996 persons are in this dataset. All analyses are conducted using the Census Bureau internal datasets – these afford us the advantage of a much larger sample than the public-use data file (which is about 66% of housing units sampled) as well as providing earnings data which are not top-coded (as they are on the public-use file). Both of these advantages provide us with a much stronger and powerful dataset to work with. Table 1 provides the sample (unweighted) frequency distributions for each of main variables in the analysis, for the three different work universes in the study. Even the smallest of the three groups (persons not working at all) has nearly a half-million sample cases.

The next step was to construct the basic five-way cross-classification table for which the medians and means will be estimated. In doing this we were sensitive to identify cells with small sample counts (less than 50 cases) which might lead to unstable mean/median estimates. Our examination of this table led us to conclude that some collapsing needed to occur over the sixteen category education variable. Most of this collapsing occurs in levels of education below a high school degree, with no schooling through 8th grade brought together, and grades 9 through 12 (no diploma) combined, as well as combining persons with some college and one or more years but no degree. This resulted in a nine category education variable, which when combined with our other factors yields a 9 (education) by 2 (gender) by 5 (race/ethnicity) by 8 (age) cross-classification of 720 cells for each of the three work universes. Within the three 720 cell tables, there are very few cells which have fewer than the desired 50 sample cases – 40 such cells in the 'less than full-time, year-round' subgroup, only 14 cells in the 'full-time, year-round' subgroup,

and a mere 2 cells in the total population table. Of these 56 cells, 25 are in the "Other race, Doctorate" category.

Computation of means and medians were done using the SAS program, primarily using the PROC MEANS procedure. Standard errors for these estimates utilized the 80 replicate weight factors provided in the ACS dataset. A simple explanation of this method is that the replicate weights are used to compute 80 different estimates (both of the mean/medians as well as their standard errors) with slightly different weights each time (reflecting the complex sampling design of the survey). The average of these 80 estimates constitutes a better, less biased estimate than conventional direct estimation provides[2].

The regression modeling of earnings used the SAS SURVEYREG procedure to take into account the replicate weights and complex sampling design of the ACS. These models were estimated using the dummy variables described earlier for each of the three Employment Status populations.

FINDINGS

Tabular-Based Results

Tables 2 and 3 show, respectively, the median and mean synthetic worklife earnings (SWE) estimates for each of the 90 demographic subgroups for the three employment status universes. For 'full-time, year-round' workers (the 'a' panel of each table), the median SWE ranges from a low of $704,005 for Hispanic females with education of None-8th grade, to a high of $4,754,930 for White males with a Professional degree (not statistically different from Asian or Other with a Professional degree).

Mean earnings, because they may include some persons with exceptionally high earnings, distort the average values in an upward direction. While the same two groups mentioned above still occupy the low and high points in the SWE estimates using means, Hispanic

---

[2] Please see Chapter 12 of the Design and Methodology Report for more information: http://www.census.gov/acs/www/SBasics/desgn_meth.htm

females with education of None-8th grade have a mean SWE value of $818,906, while White males with a Professional degree have a mean SWE value of $6,699,037.

Figure 4 shows the comparison of mean and median SWE values across education groups for the 'full-time, year-round' worker population. For each of these groups mean SWE is always higher. Figure 5 extends the analysis to all 90 Gender—Race/Ethnicity—Education groups in the study. While these points correlate at a level of .99, all 90 data points are below the 'equivalence line' (where mean and median values would be the same). (NOTE: Of the 90 mean-median comparisons, only Other males and females with a doctorate were not statistically different.) The pattern of the values suggests that at higher earnings levels the discrepancy becomes more pronounced.

Estimated values for the second employment status group, those who were employed but not 'full-time, year-round' (the 'b' panel of each table), show much lower median SWE values at the lower end of the education spectrum, with several groups (all females with education of None-8th grade) having median SWE of between $326,209 and 368,881 (with no clear statistically significant pattern among these values). At the high end, White male and Asian male Professional degree holders have estimated SWE's of $3,217,658 and $3,445,297, respectively (which are not statistically different). SWE estimates using means show similar higher valuations, as with the first subgroup analysis.

The analysis becomes more difficult when the group of study is the entire adult population (the 'c' panel of each table). For some demographic subgroups, the majority of people in specific cells actually have no earnings at all. Thus, in several cases, as Table 2c shows, median SWE in the total population group with for a person who is say, a White Female with education of None-8th grades. This is because more than 50% of the people in each of the single-age category groupings in the base table had no earnings. The corresponding part of Table 3c, however, shows that the mean SWE for these groups is a non-zero value. In this case the mean SWE values give a much better notion of "average" earnings, while the median SWE's are overwhelmed by the presence of large segments of the population with zero earnings values.

Regression-Based Results

Table 4 provides the results of the regression-based estimation. Because the regression uses a different computational approach, attempting to fit parameters to variations from cell-specific means derived from the independent variables, the possibility for bad fit in any single cell in the cross-classification is always present, if the earnings values within that cell are highly inconsistent. Overall, however, the regression results are highly similar to the tabular method. Figure 6 shows the tabular mean and regression mean SWE estimates for the 90 various education/gender/race-ethnicity combinations. These points correlate at a level of .96. As the figure shows, at lower levels of SWE tabular based estimate are lower than regression-based estimates, but as the values go up, the regression-based means become larger.

In the regression method, some patterns across groups seen in the other method are still evident. For example, the highest estimated SWE from this approach is for White males with a Professional degree ($6,259,261), just as with the tabular approaches. (The Asian male Professionals value of 6,035,220 is statistically different.) At the low end of values, Black females with education of None-8th grade produce an SWE of $472,344.

Because this method uses a regression parameter approach, it is possible to generate negative cell estimates, and by extension, negative SWE estimates. This can be seen in the panel C of Table 4, where for all Black females with education of None-8th grade or 9th-12th grade, the SWE's are negative (-$87,891 and -$14,213, respectively). Table 5c, which shows the actual regression parameters for the models for the total population, demonstrates how this is possible. Looking at the parameter values from the "Basic Demographics" model, an education level of None-8th grade has a value of -$9,962; the race/ethnicity category 'Black' has a value of -$8,548. Even though there are positive dollar values associated with every age category, none of them exceed the combined education/race value of None-8th grade/Black of -$18,510. Thus, in summing these values repeatedly across the forty years of worklife age range, negative net values are compounded, resulting in an ultimate regression-based SWE that is negative, an unlikely real-world result.

Note that this situation is a function of the estimation method used here. With fuller parameterization, including all of the cross-cell interaction effects included (and

sufficient sample to accommodate this), a more highly-parameterized model would more accurately depict the data and the variability therein. But the results demonstrated here act to provide sufficient caution that a regression-based approach to computing SWE's must be considered carefully. Nevertheless, the regression approach may have value for certain other aspects of estimation that the tabular approach cannot provide. These extensions are discussed in the next two sections.

Variation across Demographic Factors: Gender, Race/Ethnicity and Age

One of the main questions raised in an analysis such as this is the extent to which factors other than education play a role in the earnings of individuals. In both the tabular and regression analyses, we have included gender, race/ethnicity, and age to build the SWE. The regression results help to show the level of impact attributable to each of the three demographic factors of gender, race/ethnicity and age.

Since the parameter values in these models represent dollars, one simple way to understand the overall impact of a given dimension is to look at the range of variability the categories of a given factor provide in the estimate. So, for example, in looking at the results for the 'full-time, year-round' population (Table 5a, Model 2), it can be seen that the range of effect for gender is nearly $20,000 a year, since the 'male' value is $19,886, holding all else constant. The range for race/ethnicity is somewhat smaller, around $11,000, since the largest single parameter effect is for Blacks at $10,732, holding all else constant. The range impact for age is the most interesting. The lowest age group of 25-29 has been used as the omitted category in the regressions; all age categories from 40-44 and higher have a range of over $20,000—larger than the gender effect. However, the actual variability in the age categories from 40-44 to 60-64 is relatively small, with a total range of just over $2,000 a year ($22,385 minus $20,073). Thus, while age appears to have the greatest range of the three demographic variables, most of it occurs across the front half of the range. Beyond age 40, there is no substantial change in the impact of age.

Returning to the main relationship of this analysis, however, none of these demographic variables demonstrate the kind of variability in range that the education levels demonstrate. The parameters range from a low of -$8,887 (None-8th grade) to a high of

$105,168 (professional degrees), holding all else constant. The range across the education variable is about $114,000 – over FIVE times the range exhibited by the demographic factor of age. Thus, from this simple evaluation of relative impact, it is clear that the demographic factors supplement, but do not displace education as a critical component in understanding variation in earnings.

Figure 7 summarizes the variation in median SWE for various gender—race/ethnicity groups by education level.

Estimating the Impact of Other Factors

Apart from the basic education/earnings relationship we have estimated, and the contribution of demographic context factors such as gender, age, and race/ethnicity, there are other additional factors that might mediate the earnings of individuals. For example, the occupation of an individual certainly is one dimension that could have an impact on earnings (Weinberg, 2004). In this last section we look at three additional factors for their possible impact on earnings—citizenship status, English language ability, and geographic region of the country.

In recent decades, a large increase in foreign-born workers has worked to change the landscape of labor in the United States (Mosisa, 2002). Many of these workers occupy lower-paying and intermittent jobs. While some workers from abroad are well-educated and have high skill sets, others do not. For some immigrants, the English language is itself an obstacle to success in the labor force. Finally, labor markets are not completely uniform across the entire country, and relative local demands, to some extent, also mediate the earnings workers receive.

The third model of Table 5a shows the results of inclusion of these three factors, Citizenship, English language ability and Region, and their relative impact on estimated earnings for the 'full-time, year-round' worker population. These results are graphically depicted in Figure 8, showing both the relative effect of various variables, and the changes that result in overall impact, as demographic and other characteristics are added to the basic education/earnings model.

While all of these factors are highly significant, they add only a small amount to the explained variance in the model –these results are likely a function of the extremely large

sample size of the model. Naturalized citizens actually experience a small yearly increase over native born persons—$3,032, holding other factors constant, but persons who are not citizens show a negative impact (-$1,177, holding all else constant). The impact of English ability is quite sizable. All persons who speak a language other than English show a negative effect. Even those who report speaking another language but speak English 'very well' have a yearly impact of -$1,925, holding other factors constant. However, persons who report English ability below this level show one of the largest single-category effects, with a yearly loss in earnings of -$12,735, holding other factors constant, relative to persons who are English-only speakers.

Region effects are also evident, with a range of about $8,000 across the four census-defined regions, holding other factors constant. In general, earnings are highest in the East, and lowest in the Midwest, controlling for the other factors in the model.

SUMMARY

This project has focused on a variety of technical and substantive issues associated with the creation of Synthetic Worklife Earnings Estimates (SWE's). A variety of alternative universes, measures and computational strategies have been examined. It is clear that there is no 'one size fits all' approach to the problem.

We conclude by returning to several of the basic issues raised and addressed in this research. By no means are these answers conclusive or definitive; the exercise has, we believe, illuminated some of the things that need to be considered in estimating SWE's. Choice of analytic universe clearly impacts the overall SWE values estimated, but it does not appear to markedly change the relative positions of various socio-demographic groups in overall SWE. Estimation of SWE's using means or medians also has an effect, with the mean-based SWE's being larger, in some cases sizably so. In some cases, because of large numbers of persons in some subgroups with no earnings at all, median SWE values may be biased. Tabular and regression methods yield similar patterns in results; however, the regression method provides a more efficient estimation approach for the inclusion of a variety of other factors.

Future work on this topic will extend work on some of these technical issues, as well as engaging other possible dimensions that may impact synthetic worklife earnings estimates.

REFERENCES

Bianchi, Suzanne M., and Daphne Spain. 1996. "Women, Work, and Family in America." *Population Bulletin* Vol. 51, No. 3. Washington DC: Population Reference Bureau.

Card, David. 1998. "The Causal Effect of Education on Earnings Handbook of Labor Economics," in: O. Ashenfelter & D. Card (Ed.), *Handbook of Labor Economics* pp. 67-86.

Day, Jennifer Cheeseman and Eric C. Newburger. 2002. "The Big Payoff: Educational Attainment and Synthetic Estimates of Work-Life Earnings." U.S. Census Bureau, Current Population Reports, P23-210. Washington, DC: U.S. Government Printing Office.

Hecker, Daniel E. 1998. "Earnings of College Graduates: Women Compared with Men" *Monthly Labor Review*. pp 62-72.

Kerckhoff, Alan C.. 1995. "Institutional Arrangements and Stratification Processes in Industrial Societies" *Annual Review of Sociology* 21:323-347.

Loprest, Pamela and Elaine Maag. 2003. *The Relationship between Early Disability Onset and Education and Employment.* The Urban Institute. 2100 M Street NW, Washington, DC 20036.

Mosisa, Abraham T. 2002. The Role of Foreign-Born Workers in the U.S. Economy. *Monthly Labor Review*. Pp 3-14.

National Science Foundation, Division of Science Resources Statistics (NSF/SRS). 2005. *2003 College Graduates in the U.S. Workforce: A Profile*. NSF 06–304. Arlington, VA.

Siegel, Paul M. 1965. "On the Cost of Being a Negro" *Sociological Inquiry* 30:41-57.

Weinberg, Daniel H.. 2004. "Evidence from Census 2000 About Earnings by Detailed Occupation for Men and Women." U.S. Census Bureau. Census 2000 Special Reports,CENSR-15. Washington, DC: U.S. Government Printing Office.