Wage Gap Estimation with Proxies and Nonresponse

Barry Hirsch
Department of Economics
Andrew Young School of Policy Studies
Georgia State University, Atlanta

Chris Bollinger
Department of Economics
University of Kentucky, Lexington

2009 North American Summer Meetings of the Econometric Society
Boston
June 6, 2009
Overview

Household surveys include earnings nonrespondents. Approximately 30% of workers in the CPS-ORG and 20% in the March CPS-ASES (Annual Social and Economic Supplement) have earnings imputed by the Census.

This paper examines the issue of response bias in the CPS, plus the related issue of proxy respondents.

For our sample years 1998-2008 the CPS ORG has a 29.8% imputation rate based on use of usual weekly earnings among all employed wage and salary workers, unweighted. The imputation rate increases with weighting, when using hourly as well as weekly earnings to measure wage, and when restricting the sample to full-time workers ages 18-65. Our weighted full-time ORG sample has an imputation rate of 34.1%.

Same pattern holds true for the 1999-2008 March CPS (years 1998-2007), with an imputation rate of annual earnings the previous calendar year being 18.1% for the total unweighted sample. Restricting this to a weighted, full-time/full-year, age 18-65 increases the imputation rate to 19.6%.
**Nonresponse/Imputations for Earnings in CPS, Weighted Primary Sample**

<table>
<thead>
<tr>
<th>Year</th>
<th>ORG %Non-Respondent</th>
<th>ASES %Non-Respondent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>27.2%</td>
<td>17.2%</td>
</tr>
<tr>
<td>1999</td>
<td>31.0%</td>
<td>16.6%</td>
</tr>
<tr>
<td>2000</td>
<td>33.3%</td>
<td>19.6%</td>
</tr>
<tr>
<td>2001</td>
<td>35.0%</td>
<td>20.2%</td>
</tr>
<tr>
<td>2002</td>
<td>35.0%</td>
<td>21.4%</td>
</tr>
<tr>
<td>2003</td>
<td>36.4%</td>
<td>20.7%</td>
</tr>
<tr>
<td>2004</td>
<td>36.1%</td>
<td>20.8%</td>
</tr>
<tr>
<td>2005</td>
<td>35.7%</td>
<td>19.1%</td>
</tr>
<tr>
<td>2006</td>
<td>35.7%</td>
<td>20.0%</td>
</tr>
<tr>
<td>2007</td>
<td>34.9%</td>
<td>20.1%</td>
</tr>
<tr>
<td>2008</td>
<td>34.6%</td>
<td>n.a</td>
</tr>
<tr>
<td>Overall</td>
<td>34.1%</td>
<td>19.6%</td>
</tr>
</tbody>
</table>

Nonresponse represents refusals or simply “I don’t really know”. Census uses Hot-Deck procedures to impute these earnings.
Conventional wisdom (Angrist and Krueger, *Handbook of Labor Economics*, 1999) was that estimation of wage gaps is largely unaffected by inclusion of imputed earners or proxy respondents.

Hirsch and Schumacher (*Journal of Labor Economics*, July 2004) show that wage gap estimates associated with non-match criteria (e.g., union wage gaps) are severely biased, coefficient attenuation they refer to as “match bias.” Match bias affects large numbers of papers across multiple literatures in labor economics.

Match bias exists even if nonresponse is completely Missing at Random.

Bollinger and Hirsch (*Journal of Labor Economics*, July 2006), examine the case of imperfect matching (e.g., returns to schooling) as well as non-match criteria.

Procedures to correct for match bias assume conditional missing at random (CMAR).

In this paper we address two questions:

1. **How serious is nonignorable response bias in the CPS?** Is CMAR approximately correct? If so, then simple correction methods for match bias are possible (e.g., omit imputed earners from the estimation sample).

2. **Are earnings responses affected by proxy respondents?** If so, how, and how does proxy response interact with nonresponse?
Previous Literature on Response Bias and Proxy Response

Greenlees et al. (1982) & Herriot and Spiers (1975)
Validation study. 1972-73 CPS/IRS match.
Negative Selection into nonresponse

David et al. (1986)
Validation study. 1980 CPS-IRS match.
Small negative selection, most severe for married.

Korinek, Mistiaen, and Ravallion (2005)

Unit nonresponse negatively correlated with income across states.
Response is not random within states (as Census weights assume)

Validation studies. Small or non-existent bias from proxy reports.
Census Hot Deck Imputation Methods

**CPS-ORG (Monthly Earnings Files): Cell Hot Deck Method**

Cell Hot Deck used in ORG since 1979. It requires exact match with thousands of possible combinations or cells.

- 1979-1993: 11,232 cells
- 1994-2002: 14,976 cells
- 2003-present: 11,520 cells

For 1994-2002, seven categories:
  - gender (2 cells)
  - age (6)
  - race (2)
  - education (3)
  - occupation (13)
  - hours worked (8)
  - receipt of tips, commissions or overtime (2)

Beginning 2003: 10 broad occupation categories
Prior to 1994: 6 hours worked cells (no variable hours cells)

Location not used explicitly.
Match bias intuition is straightforward.

The Census uses imputation methods in which earnings nonrespondents are assigned earnings of a donor with an identical set of characteristics.

Match criteria include occupation, schooling, age, gender, etc. as shown.

Industry, union status, etc. are not criteria used in the matching.

Wage differential estimates with respect to non-match criteria (e.g., union status) are biased toward zero, the match bias or attenuation being approximately equal to the proportion of workers with imputed earnings.

Take the case of the union wage gap. Most union nonrespondents matched to and assigned earnings of nonunion donors. Some small proportion of nonunion nonrespondents is assigned earnings of union donors.

Among the 30% nonrespondent sample, there is little or no observed (donor) wage differential between union and nonunion workers. Standard OLS coefficients attenuated by roughly 25% on non-match characteristics.

This match bias occurs for numerous non-match criteria – union, industry, public, ethnicity, foreign born, veteran, marital status, region, city size, tenure, computer use, employer size, etc.

There also more subtle forms of match bias associated with imperfect matching (e.g., education, age, occupation), longitudinal analysis, and dated donors.
Fig. 1: Wage Gap Estimates from Male Full Sample, Imputed Earners, and Respondents

Source: Bollinger and Hirsch (JOLE, 2006).
I. Match Bias for Imperfectly Matched Attributes

**Education:**

Matching is based on 3 broad education groups:

- **Low** not high school graduate
- **Middle** high school graduate through less than B.A.
- **High** B.A. and above

The schooling match creates an interesting form of match bias:

- Flattens estimated earnings-schooling profiles *within* the low, middle, and high education groups
- Creating large jumps in returns *across* groups

Use CPS beginning in 1998 when Degree information is supplemented with information on GED and years spent in post-secondary education, with and without degree completion (e.g., some college, no degree, M.A. program length, etc.)

Small downward bias to *linear years of schooling* coefficient.

In **sheepskin** models (mixed years and degrees), imputations biases downward years of schooling and biases upward degree coefficients.
Estimates are from a pooled wage equation of respondents and imputed earners using the Current Population Survey monthly earnings files (CPS-ORG) for 1998-2002. The female sample size is 369,762 – 270,537 respondents and 99,225 with earnings allocated (imputed) by the Census. The sample includes all non-student wage and salary workers, ages 18 and over. Shown are log wage differentials for each schooling group relative to earnings respondents with no schooling. In addition to the education variables, control variables include potential experience (defined as the minimum of age minus years schooling minus 6 or years since age 16) in quartic form, race-ethnicity (4 dummy variables for 5 categories), foreign-born, marital status (2), part-time, labor market size (6), region (8), and year (4).
CPS Data

**Outgoing Rotation Groups (ORG):**
- Earnings supplement administered to ¼ of sample in 4\(^{th}\) or 8\(^{th}\) interview month
- Earnings and hours worked in previous week and wage when appropriate.
- Use all months January 1998 – December 2008
- Cross section and 2 period panel (4\(^{th}\) month to 8\(^{th}\) month)
- \(N = 1,499,630\) (full-time wage and salary workers, ages 18-65)

**Annual Social and Economic Survey (ASES or March CPS):**
- Earnings supplement administered to all rotation groups in March.
- Earnings, weeks worked and hours worked for previous calendar year.
- Cross section and 2 period panel (two years of earnings)
- \(N = 564,722\) (FT/FY wage and salary workers, ages 18-65)
Response Bias

Concern is nonignorable response bias. If nonresponse is not CMAR, then earnings may differ in ways not captured by measurable controls.

We use two principal methods to analyze response bias:

1. Selection models, which requires appropriate instruments to identify response
2. Longitudinal analysis
Nonresponse Rates by Predicted Wage Ventile

Percent Imputed by Predicted Wage Ventile, CPS-ORG, 1998-2006

Percent Imputed

Percentile of Predicted Wage

Full Sample
Men
Women
Determinants of Nonresponse

Probit Models of Nonresponse (Marginal Effects):

Response rates decline with age
Lower response rates for minorities except Hispanics
Response rates decline with metropolitan area size
Response rates differ across regions (low in East, higher in Mountain & West)
Men have lower response than women.
HS graduates and those with professional degrees have lowest response rates

Interview Structure / Possible Instruments to Identify Selection

Proxy rather than self reports lower response rates.
   Spouse proxies lower (6-8%)
   Other (non-spouse) proxies far lower (25%)
ORG: February or March 4-5% higher response
March: rotation group month-in-sample, MIS 1 or 5 2% higher response
# Nonresponse/Imputation Rates by Interview Month and Proxy Status

<table>
<thead>
<tr>
<th>Proxy Status</th>
<th>CPS ORG Sample</th>
<th>CPS ASES Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self Report</td>
<td>756,693</td>
<td>281,887</td>
</tr>
<tr>
<td>Any Proxy Report</td>
<td>742,937</td>
<td>282,835</td>
</tr>
<tr>
<td>Spouse Proxy</td>
<td>452,234</td>
<td>185,813</td>
</tr>
<tr>
<td>Nonspouse Proxy</td>
<td>290,703</td>
<td>97,022</td>
</tr>
<tr>
<td>MIS 1 or 5</td>
<td>n.a</td>
<td>142,330</td>
</tr>
<tr>
<td>MIS 2-4, or 6-8</td>
<td>n.a</td>
<td>422,392</td>
</tr>
<tr>
<td>February</td>
<td>123,985</td>
<td>n.a</td>
</tr>
<tr>
<td>March</td>
<td>122,831</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Total %Imputed</strong></td>
<td><strong>1,499,630</strong></td>
<td><strong>564,722</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CPS ORG Sample</th>
<th>27.8%</th>
<th>24.0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPS ASES Sample</td>
<td>15.2%</td>
<td>18.6%</td>
</tr>
</tbody>
</table>
Approach 1 to Nonresponse: 
Selection Wage Models

Focus on Full Time and Full Year (ASES) Workers

Separate Models for Men and Women

Compare Three Models:
   - Linear Model estimated with OLS
   - Selection Corrected for Nonresponse using 2-step Heckman Procedure
     *With Proxy* as Selection Instruments (*plus month instruments*)
   - Selection Corrected for Nonresponse using 2-step Heckman Procedure
     *Without Proxy* as Selection Instruments (*only month instruments*)
## Selection Coefficient (Mills Ratio) Results

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ORG w/o proxy variables</strong></td>
<td>-0.167**</td>
<td>-0.114**</td>
</tr>
<tr>
<td><strong>ORG with proxy variables</strong></td>
<td>-0.166**</td>
<td>-0.034</td>
</tr>
<tr>
<td><strong>ASES w/o proxy variables</strong></td>
<td>-0.267**</td>
<td>-0.142**</td>
</tr>
<tr>
<td><strong>ASES with proxy variables</strong></td>
<td>-0.276**</td>
<td>-0.062</td>
</tr>
</tbody>
</table>

**Men:**
Negative Selection into response is a significant issue for men. Proxy (spouse and other) appears to be a valid instrument for men (no statistical difference in selection coefficient estimates between two models).

**Women:**
Selection does not appear to be significant among women. Proxy does not appear to be a valid instrument for women (significant change in selection coefficient).

Limiting this model to heads and co-heads sharply reduces negative selection among men, while not affecting women. Strong negative selection among men is driven heavily by non-heads (employed male children, partners, other household members).
How does response bias affect wage regression coefficients?

We show by comparing OLS and selection predicted log wages

Predicted Mean Log Earnings

<table>
<thead>
<tr>
<th></th>
<th>ORG Sample</th>
<th>ASES Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Men</td>
<td>Women</td>
</tr>
<tr>
<td>OLS Predicted Earnings</td>
<td>2.981</td>
<td>2.790</td>
</tr>
<tr>
<td>Selection Corrected Predicted</td>
<td>3.069</td>
<td>2.806</td>
</tr>
<tr>
<td>(Corrected – OLS) Differential</td>
<td>0.088</td>
<td>0.016</td>
</tr>
<tr>
<td>Bias in gender gap estimates</td>
<td>-0.072</td>
<td></td>
</tr>
</tbody>
</table>

Virtually all the difference in coefficients is in the intercepts and not slopes. Selection is nearly a worker fixed effect. Sample selection appears to be nearly parallel to the regression line for both men and women, but “chopping off” the highest earners for men.

**Implication for gender wage gap: 0.07 or 7% understatement of $W_M/W_F$**
Approach 2 to nonresponse: Longitudinal analysis

Log Wage and Earnings Growth Rates by Imputation Status
(Full-time workers both years, no controls)

<table>
<thead>
<tr>
<th>Year1, Year 2</th>
<th>ORG Sample</th>
<th>ASES Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>percent</td>
<td>Men</td>
</tr>
<tr>
<td>Respond, Respond</td>
<td>58.9</td>
<td>0.028</td>
</tr>
<tr>
<td>Respond, Impute</td>
<td>12.6</td>
<td>0.004</td>
</tr>
<tr>
<td>Impute, Respond</td>
<td>12.8</td>
<td>0.037</td>
</tr>
<tr>
<td>Impute, Impute</td>
<td>15.7</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Rationale is that the **imputed value** is the Census estimate of the average earnings of respondents, conditional on X’s.

Mean wage changes among those switching between reporting and not reporting earnings reveal that **imputed earnings understate** the earnings of non-reporting workers. This analysis controls for worker heterogeneity; supports finding of **negative selection** into response.
Puzzle
  CS selection models: negative selection for men, weak selection for women
  But panel data show negative selection as strong (or stronger) for women

Can CS selection and panel models be reconciled?
  In part – but not fully
    a. The panel and CS compositions differ, the panel having more co-
       heads of households and fewer working children, partners,
       roommates, and other household members
    b. Co-heads less likely to drop out of a sample
    c. Estimating the CS and panel models from the same panel samples
       shows much weaker selection for men, which narrows the puzzle.
    d. But it also drives it to zero for women.
    e. So we cannot explain the similar male/female panel results.
Both models rely on not fully verified assumptions. Whether it’s the
selection model or panel model assumptions that are most likely to be
violated, we cannot know with confidence.
Is Nonresponse a “Fixed Effect” in panel models?

Selection wage change equation, with nonresponse defined as not reporting in either or both of the two years. Selection identified by Proxy and Month variables

**Selection Coefficients for Nonresponse in both periods**

<table>
<thead>
<tr>
<th></th>
<th>ORG Sample</th>
<th>ASES Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Men</td>
<td>Women</td>
</tr>
<tr>
<td>Mills lambda</td>
<td>0.006</td>
<td>-0.008</td>
</tr>
</tbody>
</table>

Selection coefficients are small and largely insignificant for men and women in both ORG and ASES. These panel results include X’s in levels, but we get the same basic result excluding these.

**Conclusion:** Selection into response is a fixed effect, with little to no response bias in longitudinal wage change coefficients.

Longitudinal analysis well suited for analyzing proxy effect on earnings.
Proxy Responses

Proxy responses are an important determinant of selection into response. Does proxy reporting affect earnings apart from selection into response?

In analyzing the effects of proxy response, we distinguish between proxy reports from spouses and non-spouses.

Descriptive data on proxy response:

**Self Reports and Proxy Earnings Responses, by Gender and Marital Status**

<table>
<thead>
<tr>
<th></th>
<th>ORG Sample</th>
<th></th>
<th></th>
<th>March ASES Sample</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Men</td>
<td>Women</td>
<td>All</td>
<td>Men</td>
<td>Women</td>
</tr>
<tr>
<td>Self Reports</td>
<td>50.5%</td>
<td>42.9%</td>
<td>59.8%</td>
<td>49.9%</td>
<td>42.8%</td>
<td>59.1%</td>
</tr>
<tr>
<td>Proxy Spouse</td>
<td>49.5%</td>
<td>57.1%</td>
<td>40.2%</td>
<td>50.1%</td>
<td>57.2%</td>
<td>40.9%</td>
</tr>
<tr>
<td>Non-spouse</td>
<td>30.2%</td>
<td>36.6%</td>
<td>22.2%</td>
<td>32.9%</td>
<td>39.5%</td>
<td>24.5%</td>
</tr>
<tr>
<td></td>
<td>18.4%</td>
<td>20.5%</td>
<td>18.0%</td>
<td>17.2%</td>
<td>17.8%</td>
<td>16.4%</td>
</tr>
<tr>
<td>%Spouse Prox</td>
<td>60.9%</td>
<td>64.1%</td>
<td>55.3%</td>
<td>65.7%</td>
<td>68.9%</td>
<td>59.9%</td>
</tr>
</tbody>
</table>
Using standard OLS wage equations:

1. Proxy earnings, conditional on $X$, 2-3 percent lower, more for men than women
2. Separate effect of having a non-spouse proxy and a spouse proxy
   a. Non-spouse proxy: large negative effect of about -6%
   b. Spouse proxy: zero effect on reported earnings

<table>
<thead>
<tr>
<th></th>
<th>ORG Sample</th>
<th>ASES Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Men</td>
<td>Women</td>
</tr>
<tr>
<td>All Proxy Respondents</td>
<td>-.028**</td>
<td>-.019**</td>
</tr>
<tr>
<td>Nonspouse Proxy</td>
<td>-.068**</td>
<td>-.086**</td>
</tr>
<tr>
<td>Spouse Proxy</td>
<td>-.014**</td>
<td>-.008</td>
</tr>
</tbody>
</table>

It turns out that these results are highly misleading, the difference between spouse and non-spouse proxies being due to worker heterogeneity. But this does point up a problem for standard wage equation analysis.
Longitudinal analysis well suited to examine proxy effects
Panel data wage and earnings growth for proxy switchers

<table>
<thead>
<tr>
<th>Indicated Category</th>
<th>ORG Sample</th>
<th>ASES Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Men</td>
<td>Women</td>
</tr>
<tr>
<td>Nonspouse Proxy</td>
<td>-.026**</td>
<td>-.016**</td>
</tr>
<tr>
<td>Spouse Proxy</td>
<td>-.024**</td>
<td>-.020**</td>
</tr>
</tbody>
</table>

In OLS cross sections, it is non-spouse proxies who report substantially lower earnings, but this appears to be a fixed effect (heterogeneity). Longitudinal evidence is that proxy effects are negative but small for both men and women and for spouse and non-spouse proxies, on the order of 1.5% to 2.5%.

Implications: Workers who have earnings reported by a non-spouse household proxy tend to have lower unmeasured skills. Women understate men’s earnings more than men understate women’s.

Good news: Non-spouse and spouse proxies equally proficient at reporting earnings and bias is small, about 1.5% to 2.5%.

Bad news: Omitted variable bias in a standard cross-sections. Proxy variables contain omitted information, negative for nonspouse proxy.
Conclusions

- Earnings nonresponse is a critically important issue in the CPS.
- First-order problem is match bias from imputation. Wage analyses should omit imputed earners (possibly use IPW) or must account for selection.
- Response bias is second order issue, potentially important, not widely studied.
- Appears to be negative selection into response based on CS selection models.
- Negative CS selection substantive for men (≈10%). Near zero for women.
- Selection into response less an issue for household co-heads.
- Panel evidence shows similar negative selection for women as well as men.
- Response bias largely a fixed effect (affects intercept). Panel coefficients largely unaffected by selection. Imputed earners must be omitted.
- Proxy response highly correlated with nonresponse; must consider together.
- Spouse and non-spouse proxy respondents differ in relation to wages and nonresponse in cross section analysis. Apparent large non-spouse and zero spouse proxy effects on CS wages reflect heterogeneity.
- Based on longitudinal analysis, proxy effect on wages for both spouse and non-spouse proxies is likely small, on the order of minus 2 percent.