

Is It What You Do Or Where You Work That Matters? Gender Composition and the Gender Wage Gap Revisited[↑]

Mahmood Arai, Lena Nekby, Peter Skogman Thoursie*

Abstract:

This study uses matched employer-employee private sector data for Sweden to analyze if single-sex job cells, where job is defined as an occupation within an establishment, pose a potential problem for estimation of within job gender wage differentials. In order to assess a within job gender wage gap, it is essential to control for wage differentials that exist between job cells. This can be done either by controlling for gender segregation via a gender composition measure or by controlling for all between job differences with job fixed effects. We argue that if a substantial amount of job cells are single-sex, the estimate for gender wage differentials may be underestimated in estimation controlling for gender composition due to collinearity between the gender composition measure and the female dummy variable.

Keywords: Gender wage differentials, gender segregation, discrimination

JEL: J16, J31, J62, J71

[↑] We have benefited from comments from Juhana Vartiainen and Ryszard Szulkin as well as seminar participants at the Trade Union Institute for Economic Research and the European Association for Labor Economists.

* Mahmood Arai; Department of Economics, Stockholm University and Stockholm Linnaeus Center for Integrations Studies (SULCIS), ma@ne.su.se. Lena Nekby (**corresponding author**); Department of Economics, Stockholm University, S-10691 Stockholm, Sweden, SULCIS and IZA Research Fellow, lana.nekby@ne.su.se. tel: +46(0)8164481, fax: +46(0)8159482. Peter Skogman Thoursie; Department of Economics, Stockholm University, peter.thoursie@ne.su.se.

1. Introduction

The gender wage gap has been a subject of study in economics for decades. One of the main questions in this literature is how much of the gender wage gap can be attributed to segregation into jobs and how much is due to wage differentials within jobs. Until recently, only individual level data have been available to study this question and the only information on job affiliation has been through occupational codes. Using this type of data, many studies have analyzed the role of occupational segregation in explaining the gender wage gap.¹ The recent availability of matched employer-employee data have made it possible to examine the role of segregation departing from other empirical measures of job. Job has then typically been defined as an occupation within an establishment, or more specifically, as an interaction between establishment and occupation.² The aim of this paper is to discuss how the empirical definition of job, which is contingent on how data is aggregated, may influence estimation of within wage differentials.

In order to assess a within occupation-establishment (job) gender wage gap, it is essential to control for wage differentials that exist between job cells. Previous studies have relied on two methods to do this, either by including in wage estimation a control for the share of female workers or by fixed effects estimation aimed at removing all between cell differentials through the introduction of dummy variables. Theoretically, these two methods are equivalent, i.e., yield an identical coefficient for the female dummy variable which measures the gender wage gap in wage estimation (see discussion in Datta Gupta & Rothstein, 2005). If however a substantial amount of job cells are single-sex, i.e., all female or all male, the estimate of the impact of gender job segregation on gender wage differentials may be underestimated. The appropriate strategy in such a case is to control for segregation by means of job-cell fixed effects.

2. Single-sex job cells

Do men and women working within the same occupation and establishment experience different wages? Theoretically, research aims to capture within “job” gender wage differentials, i.e., wage differences between men and women doing the same *work* within the same establishment. However as well-defined job codes do not exist, data restrictions limit how job is defined empirically to an interaction between establishment and occupation. The

¹ See for example Macpherson & Hirsch, 1995.

² See Groshen, 1991; Datta Gupta & Rothstein, 2005; Bayard et al., 2003 and Hellerstein & Neumark, 2005.

size of defined job cells will then vary depending on the level of aggregation of occupation codes. This is due to the fact that establishment is a more strictly defined entity. Hence, a more detailed classification of occupation leads to more narrowly defined job cells.³

We argue that using female shares in estimation in order to assess the role of segregation has potential drawbacks stemming from how job cells are defined. While a narrower cell definition more precisely controls for between-cell variation, there are problems with using a “too” narrow job cell definition.

Our concern is that as job becomes more narrowly defined, the probability of obtaining occupation-establishment cells with only men or only women increases. By definition, there are no within job-cell wage differentials for such single-sex job cells. The implication is that in a standard wage regression with a female dummy variable aimed at measuring the gender wage gap and a control for segregation, i.e., the share of females in a defined job cell, these two right hand side variables will have identical values (0 or 1) for a proportion of observations. The obvious example of this is a man working in an all male job or a woman working in an all female job. For these observations, the gender dummy variable partially captures between cell differentials which is not intended.

How big of a problem is this? Previous studies using matched employer-employee data to study within job wage differentials do not explicitly show the distribution of female shares across jobs or across individuals and therefore do not account for whether or not single-sex job cells are problematic for estimation. Using matched employer-employee private-sector data from Sweden, we can show the distribution of female shares in order to illustrate not only the dimension of single-sex jobs but also its dependence on level of aggregation of occupation. Ultimately, we can also show how this potential problem may affect estimation of the impact of segregation on the gender wage gap.

The data used in this paper are based on a sample of employees from matched registers for all workers employed in establishments included in the 1995 Swedish Establishment Survey (*Super APU*).⁴ In order to avoid observations on young employees in the entry phase of labor market participation, the sample is limited to those workers between 30 to 64 years of age in 1995.⁵ In terms of establishment restrictions, the sample is limited to

³ How closely the empirical definition of job matches the theoretical definition of job is debatable. Consider two cashiers working in different supermarkets within the same chain, theoretically these two occupation-establishment cells are likely be considered as the same job, empirically they are defined as different.

⁴ See le Grand *et al.* (1996) and the Introduction in Vilhelmsson (2002) for more information on *Super APU*.

⁵ In addition, part time workers, i.e., employees working less than 50 percent, are excluded from the sample.

those 340 establishments with more than 100 employees. This is done in order to allow for sufficient variation in the gender composition measure.

The final sample consists of 159,511 employees. In order to compare the consequences of using more or less aggregated occupation codes, we use two aggregation levels for occupation, one based on 8 broad categories (one-digit level) and the other on 67 categories (two-digit level). Using 8 (67) occupational levels together with 340 establishments, 1,592 (4,949) occupation-establishment cells are defined implying an expected 100 (32) employees per job.⁶ In comparison, previous studies have, based on the information available, only 0.04 - 17.5 employees per job.⁷

Figure 1a shows the distribution of employees across occupation-establishment female shares based on the broader categorization of occupation (8). As shown by the columns for female shares equal to zero and one, 3,742 men work within all male jobs and 664 women in all female jobs. In other words, approximately 2.8 percent of the individual sample are found in single-sex occupation-establishment cells. Figure 1b shows the same distribution using the finer categorization of occupation (67 occupations). The number of men working in all male jobs grows to 9,922 and the number of women in all female jobs to 5,307. This implies that when occupation-establishments cells are more narrowly defined, 9.5 percent of the individual sample are found in single-sex jobs. Note that these results are found despite the broad restrictions made concerning firm size. With no restrictions on firm size, 12 percent of the individual sample are found in single-sex cells with the broader definition of job (based on 8 occupations) and 19.2 percent with the finer definition (67 occupations).

- Figure 1a and Figure 1b here -

A large number of individuals in single-sex job cells may be due to one or few large employers that sort men and women into different jobs. Figure 2a and 2b show the distribution of jobs over job female shares indicating that a substantial number of jobs are single-sex. Using 8 (67) occupation categories, 22.5 (53.6) percent of jobs are single-sex.

- Figure 2a and Figure 2b here -

⁶ Note that the number of defined job cells is less than the maximum possible as not all occupations exist in all establishments.

⁷ See Groshen, 1991; Datta Gupta & Rothstein, 2005 and Bayard *et al.*, 2003.

The implication of single-sex job cells for wage estimation is that for all individuals belonging to occupation-establishment cells with female shares equal either to zero or one, the segregation measure becomes equal to the female dummy variable stealing potential variation from the female dummy variable. One suggested rule of thumb is that multicollinearity is a concern if the overall coefficient of determination (R^2 's) in the regression is less than any of the individual R^2 's from regressions of each independent variable on the other regressors (see for example discussion in Greene, pp 421-422). Table 1 shows the R^2 's for the wage equation and for the regression of each independent variable (female dummy and each female share variable) on the other regressors. Estimation is based on the two specifications controlling either for female shares at four levels, industry, occupation, establishment and occupation-establishment (Model 1), or at the occupation-establishment level only (Model 2). The coefficient of determination for the wage equation is approximately 0.4 and the R^2 's when female and female shares are used as dependent variables respectively are nearly as high (gender dummy) or higher (female share in Model 1). Note that the R^2 's in estimation with the gender dummy as the dependent variable increases with the finer occupation categorization implying a larger potential multicollinearity problem with more narrowly defined occupation-establishment cells. This type of multicollinearity problem has serious implications for the interpretation of results, the coefficient measuring gender wage differentials (the female dummy variable) will be underestimated and therefore the effect of segregation on the gender wage gap overestimated.

- Table 1 -

How big of a problem this is in estimation is generally difficult to ascertain since previous studies do not explicitly account for the distribution of female shares across individuals or across jobs. Given that multicollinearity is a problem, and our data certainly indicates that it may be, a “safer” estimation practice is to use fixed effects. Not only will fixed effects estimation clearly control for all between-cell wage differentials but also, and in contrast to controlling for female shares at various levels, fixed effects estimation will not suffer from the multicollinearity problem described above.

One potential drawback of fixed effects estimation is that there is no direct interpretation of the effect of segregation on wages. On the other hand, wage estimation that controls simultaneously for different levels of segregation, common in the literature, also yield non-robust estimates of the effect of segregation. As shown in Table 1 (Model 1),

female shares at various levels are highly correlated to each other. In fact, R^2 's from regressions using female share as the dependent variables are in most cases substantially higher than the R^2 's from the wage regression. Not only does this multicollinearity problem lead to non-robust estimates of the effect of different types of segregation on wages but it also confounds any analysis of the relative importance of the different types of segregation. In this respect, a more appropriate method is to control in separate estimations for each type of segregation (see Datta Gupta & Rothstein, 2005).

In addition, fixed effect estimation is the most flexible functional form. The effect of segregation may vary across the distribution of female shares, i.e., may vary according to whether female shares are at low or high levels. Fixed effects estimation does not restrict a numeric distance between cells. It is however possible, in wage estimation using female shares, to use higher order polynomials of female shares in order to capture non-linearities. Note that in our example, including higher degrees of polynomials did not alter the impact of segregation on the gender wage gap.

To illustrate these issues further and to compare with previously reported results on gender wage differentials based on matched employer-employee data, the following wage segregation models are estimated: (1) log wages regressed on a gender dummy variable, aimed at measuring the gender wage gap, controlling simultaneously for gender composition (share female) at four levels: industry, occupation, establishment and occupation-establishment, (2) log wages regressed on a gender dummy variable controlling for gender composition at the occupation-establishment level only and (3) log wages regressed on a gender dummy variable controlling for the full set of occupation-establishment fixed effects. All three specifications include controls for a basic set of human capital variables, experience (quadratic), seniority, education (6 dummy variables) and immigrant status and are estimated on the two afore-mentioned aggregation levels for occupation (see Table 2 for descriptive statistics).

- Table 2 here -

3. Wage regressions

Table 3 shows results for the female dummy variable in three different wage estimates and for the two different levels of occupation (8 and 67 occupations respectively).⁸ Model 1 replicates studies that control for segregation by including measures for gender composition at four levels (industry, occupation, establishment and occupation-establishment). Similar to these studies, segregation together with human capital differences is found to explain approximately 50 percent of the gender wage gap. Results when controlling only for female share at the occupation-establishment level (model 2) are very similar to results obtained when female shares at all levels are included in estimation (Model 1).⁹ This is not surprising since female shares at various levels are highly correlated.¹⁰

- Table 3 here -

Model 3 shows the results for fixed effects estimation where all between job differences are controlled for by including a full set of job dummy variables in estimation. Results from fixed effects estimation indicate that the gender wage gap increases from the previous two estimates using controls for female shares. Fixed effects estimation yields results indicating that between job wage differences account for only 35-36 percent of the gender wage differentials implying a substantial within occupation-establishment gender wage gap.

These results suggest that wage estimates controlling for female shares may overestimate the importance of gender segregation on gender wage differentials due to multicollinearity issues stemming from single-sex job cells. Using female shares in estimation implies that for the group of observations belonging to single-sex cells, female share does not capture all between cell differences. Some of the between differences are instead captured by the female dummy variable. When between differences in single-sex cells are lower on average than within differences in mixed-sex cells, the gender variable will render a lower estimate than it should due to the fact that this variable is partially capturing between cell differences.

⁸ Results from separate estimation shows that human capital differences between men and women alone explain 25 percent of the raw gender wage gap of 18 percent.

⁹ In separate estimations for each model, standard errors were clustered at the occupation, establishment and occupation-establishment level. Significance levels were unaffected by choice of clustering level.

¹⁰ Coefficient estimates for the segregation measures are not shown but are available upon request. As noted, these segregation measures are highly correlated. Indeed, coefficient estimates are non-robust and vary in strength and sign not only between the two levels of aggregation but also between model 1 and model 2 as well as for separate wage estimations where controls for female shares at each level are included one at a time.

Although there is a small increase in the proportion of the gender wage gap explained by gender segregation with the finer job classification (Model 1 and 2), the more notable difference in results is between using a segregation measure and fixed effects. This disparity in results, in our data, is found for both aggregation levels.

4. Conclusions

The availability of matched employer-employee data have made it possible to examine the role of segregation departing from a definition of job rather than occupation as done in the previous literature. Job is then typically defined as an occupation within an establishment. Results presented in this paper suggest that a too narrow definition of job, stemming from finely defined occupation categories, leads to a substantial amount of single-sex job cells. This implies that in a typical wage regression aimed at measuring the gender wage gap, the measure for female shares becomes equal to the female dummy variable. The role of gender segregation on the gender wage gap may then be overestimated as the measure for female share steals an unjustifiable amount of variation from the female gender dummy variable.

If a substantial amount of job cells are single-sex, as is the case in any labor market characterized by gender segregation (and exacerbated by finer definitions of occupation), choice of estimation method is no longer trivial. Using job fixed effects rather than female shares to control for between group differences, avoids the problem of single-sex jobs being perfectly correlated with the female dummy variable. Our results indicate that when fixed effect estimations are performed, segregation together with human capital differences explain only 35 percent of the gender wage gap compared to 50 - 53 percent when female shares are used in estimation.

Our study suggests that it is essential, in wage estimation controlling for segregation, to report the distribution of female shares across jobs and workers in order to determine if multicollinearity is a potential problem. The example used in this paper has been restricted to studying the impact of gender segregation on gender wage differentials. Our arguments are however equally viable to any type of empirical work dealing with the problem of estimating the relationship between two variables through the control of some type of category affiliation.

References

Bayard, K., Hellerstein, J., Neumark, D., and Troske, K. (2003), "New Evidence on Sex Segregation and Sex Differences in Wages from Matched Employee-Employer Data", *Journal of Labor Economics*, 2003, vol. 21, no. 4: 887-922.

Datta Gupta, N. and Rothstein, D. (2005), "The Impact of Worker and Establishment-level Characteristics on Male-Female Wage Differentials: Evidence from Danish Matched Employer-Employee Data", *Labour: Review of Labour Economics and Industrial Relations*, 19 (1): 1-34.

Greene, W. H. (1997), *Econometric Analysis*, Third edition, Prentice Hall. New Jersey.

Groshen, E. L. (1991), "The Structure of the Female/Male Wage Differential: Is It Who You Are, What You Do, or Where You Work?", *Journal of Human Resources*, 26 (3): 475-472.

Hellerstein, J. and Neumark, D. (2005), "Using Matched Employer-Employee Data to Study Labor Market Discrimination", in William Rodgers, ed., *Handbook on the Economics of Discrimination 2006* (Great Britain: Edgar Elgar Publishing).

le Grand, C., Szulkin, R. And Tåhlin, M. (1996), "Sveriges arbetsplatser - organisation, personalutveckling, styrning", Second Edition, SNS Förlag, Stockholm.

Macpherson, D. A., and Hirsch, B. T. (1995), "Wages and Gender Composition: Why Do Women's Jobs Pay Less?", *Journal of Labor Economics*, 13(3), 426-471.

Vilhelmsson, R. (2002), "Wages and Unemployment of Immigrants and Natives in Sweden", Ph.D. Thesis, Swedish Institute for Social Research, Stockholm University, Stockholm.

Tables

Table 1: Coefficients of Determination (control variables in parentheses).

Dependent Variable (control variables in parenthesis)	8 Occupations		67 occupation	
	(1)	(2)	(1)	(2)
Log Wage (F + HC + SF)	0.43	0.42	0.43	0.42
Female (HC + SF)	0.32	0.32	0.41	0.41
Share female: Occupation-establishment Model 1: (F + HC + SF); Model 2: (F + HC)	0.75	0.30	0.78	0.39
Share female: Establishment (F + HC + SF)	0.78	--	0.73	--
Share female: Occupation (F + HC + SF)	0.42	--	0.62	--
Share female: Industry (F + HC + SF)	0.64	--	0.64	--

Notes: Model 1 controls for human capital variables (HC), gender (where relevant (F)) and share female (SF) at four levels: occupation-establishment (job), establishment, occupation and industry (3 levels when share female is dependent variable). Model 2 controls for human capital variables, gender (where relevant) and share female at the establishment-occupation level only (controls for gender and human capital when share female is dependent variable).

Table 2: Means and Frequencies, By Gender.

Variables	Men	Women
Monthly full-time wage,	20.2	16.5
SEK 1000	(8.1)	(5.1)
Experience	26.1	23.6
	(8.7)	(7.6)
Seniority	7.0	6.5
	(3.0)	(3.1)
Immigrant (1/0)	0.13	0.16
Education levels:		
Elementary	0.16	0.16
Compulsory	0.10	0.14
Upper secondary , <12 years	0.28	0.35
Upper secondary	0.20	0.14
University, <3 years	0.12	0.11
University	0.14	0.11
No. of observations	114,133	45,378

Note: Information based on the 1995 *Super APU* sample of employees aged 30-64 years employed in establishments with at least 100 employees. Standard deviations in parentheses.

Table 3: Wage Regressions (log monthly full-time equivalent wages). Unadjusted gender wage differential = -0.18

	8 occupations		67 occupations	
	Wage Gap	% Explained	Wage Gap	% Explained
1. Human Capital & Female Shares (4 levels):	-0.090*** (0.002)	50 %	-0.085*** (0.002)	53 %
2. Human Capital & Female Shares (job level only):	-0.090*** (0.002)	50 %	-0.083*** (0.002)	54 %
3. Fixed Job Effects	-0.116*** (0.001)	35 %	-0.113*** (0.001)	36 %
No. of observations	159,511		159,511	

Notes: Robust standard errors in parentheses. Standard errors were clustered, in separate estimation for each model, at the occupation, establishment and occupation-establishment level. Significance levels are unaffected by choice of clustering level. Human capital controls, included in all specifications, are experience (quadratic), seniority, education (6 dummy variables) and immigrant status. Specification 1 controls simultaneously for female shares at the industry, occupation, establishment and job-level (occupation-establishment level).



