"Does it get better?"
A critical response to Carpenter and Eppink (2017)

Duc Hien Nguyen
University of Massachusetts Amherst

May 3, 2020

1 Introduction

Questions of whether labor market discrimination exists, and to what extent it may explain inferior economic outcomes for distinct social groups have been an important topic of research since Gary Becker's seminal book (Becker (1957)). Discrimination is a matter of importance for policy makers and researchers because of intrinsic concern over fairness and equality. For instance, although the male-female gender wage gap in the US has narrowed over time, in 2010 women still earns only 77 cents for every dollar of men's earnings, on average. Close to 40 percent of this gap cannot be explained by conventional productivity-related characteristics e.g. experience, education, race, industry, or occupation (Blau and Kahn (2017)). Such persistent inequality is indicative, although not conclusive, of gender discrimination in the labor market. Moreover, discrimination also entails a loss of efficiency, as the most qualified and productive workers may be prevented from matching with the appropriate jobs due to discrimination. Last but not least, discrimination can be a self-fulfilling prophecy, such as when both women and men share the belief that women are more likely to agree to do low-promotability tasks (Babcock et al. (2017)). Because of this belief, more women than men are asked to do such tasks, and they are also more likely to agree to those requests. This slows women's career advancement relative to men, generates gender disparity, and unwittingly fulfills the stereotype. Overall, there is an on-going need to investigate empirically the existence of discrimination, its potential sources, and any appropriate policy intervention.

Starting with Badgett (1995), a strand of the literature has been devoted to the experience of lesbian, gay, bisexual, and to a smaller but growing extent, transgender workers. Early studies often find a consistent earnings penalty for gay men relative to straight men, and an earning premium for lesbian relative to straight women, and some suggestive evidence that both earnings gaps seem to narrow over time (Klawitter (2015), Mueller (2014), Dilmaghabi (2018)). More recent studies are split on whether the gay earnings penalty still persists or has it been closed (Waite and Denier (2015), Cerf (2016), Jepsen and Jepsen

(2017), Martell (2018)), but a paper by Carpenter and Eppink (2017) published in the Southern Economic Journal is the first to find that "self-identified gay men are estimated to earn significantly *more* than similarly situated heterosexual men - a difference on the order of 10% of annual earnings."(428, emphasis original).The authors themselves seem to be surprised by this novel findings. They provide evidence to suggest that data quality, reduced discrimination, and changing patterns of household specialization *cannot* account for the gay men premium. But what does account for the substantially higher earnings of gay men? - that remains unanswered.

In this paper, I revisit Carpenter and Eppinks (2017). Using the same dataset, I replicate and extend their original analysis with the aims to provide clarity into their findings. First, by modifying their specification and regression sample, I show that much of the gay men premium are driven by earnings of workers who are older and those who are single, and by the inclusion of industry and occupation controls. Second, using Oaxaca - Blinder decomposition technique, I show that the different demographics and labor market characteristics contribute to the earning gaps in contradictory and inconsistent ways. These observations raise two critical questions. One, we need to carefully reflect on the impact of including industry and occupation controls in the regression model. Restricting the comparison of wages and earnings to be within occupation or industry may overlook the effect of occupational segregation - itself a potential forms of structural discrimination (Elwert and Winship (2014), Tilcsik et al. (2015), Martell (2018)). Moreover, for empirical works with marginalized social groups such as LGBTQ workers, including too many set of controls may exacerbate the problems with small sample size. Two, the National Health Interview Survey (NHIS) data that Carpenter and Eppinks uses seem to have a number of peculiar characteristics, and this raises question about external validity for their findings.

The rest of this paper will be structured as follows. Section 2 provides a brief literature review to contextualize Carpenter and Eppink's findings. In section 3 I discuss in more details their results, and some robustness check. I present my Oaxaca-Blinder decomposition in section 4, and discuss the issue of controlling for occupation and small sample bias in section 5.

2 Literature review

Discrimination based on sexual orientation and gender identities (SOGI) in the labor market are often operationalized somewhat narrowly as residual wage differentials in an earning regression. The intuition is that for a given sample of cisgender male workers, if after accounting for all characteristics relevant to productivity one still observes an unexplained wage gap between gay male workers and straight male workers, then it is suggestive that discrimination exists. Causality is more difficult to establish, given potential endogeneity concerns such as unobserved ability, endogenous selection, etc. Nonetheless, empirical works

seem to establish sufficient robust evidence to make a prima facia case that gay and lesbian workers are treated differently compared to their straight counterparts. In a meta-analysis of 34 studies of gay men and 29 studies of lesbian women,Klawitter (2015) found that there is a gay men penalty in wages and earnings of average 11% (in a range of -30 % to 0%, standard error 2 %). On the other hand, lesbian women receives a wage premium of average 9% (range from -25% to 45%, with standard error 2 %). These gay penalty and lesbian premium sum up the conventional wisdom of early studies. The gender difference itself presents a paradox, upsetting any easy explanation based on homophobia or animus toward homosexuality.

Klawitter (2015) also observes that both the gay penalty and the lesbian premium seem to trend toward zero, perhaps reflecting changing social attitudes. More recent works provide contradictory supports for this observation. As summarized in table 1 below, some studies find no statistically significant evidence of a gay penalty (Cerf (2016), Dilmaghabi (2018)) while others continue to find substantial and precisely estimated evidence for one (Waite and Denier (2015), Jepsen and Jepsen (2017), Martell (2018)). At the same time, almost all studies find persistent evidence of a lesbian premium, with magnitude of the same order of magnitude as the average reported in Klawitter (2015). Importantly, however, no study have found a statistically significant evidence of earning advantage for gay men, except Carpenter and Eppink (2017). The magnitude of the earnings premium that they found - nearly 10% of annual earnings - is also very substantial, and larger than the lesbian premium found in their sample. Within the context of all other studies, Carpenter and Eppink (2017) stands out as both novel and unusual.

It is also unusual because their own data analysis rules out any easy explanation for the gay men earnings premium. Improvement in social attitudes towards LGBTQ people in the US over time cannot explain the persistent lesbian earnings premium (which has been documented as early as in Badgett (1995)). Even in the implausible scenario that the improving attitude only applies selectively to gay men and not lesbian women, in the absence of discrimination one should expect earnings parity between gay and straight men, especially after controlling for demographic and labor market characteristics. Neither is the gay premium driven by pattern of household specialization, because Carpenter and Eppinks also found that among the partnered subsample, the estimated lesbian and gay premia are not statistically significant (more details on this in section 3). Lastly, they acknowledge that in their data there is no control for hours worked. Even though they have limited their sample to full time workers, differences in actual hours worked may lead to different annual earnings. On the other hands, it is often found that gay men supply less hours to paid works than straight men on average (Black et al. (2007), Waite and Denier (2015)). If this also holds for Carpenter and Eppink sample, one may expect an even bigger gay premiums on the basis of hourly wages. Ultimately, the authors leaves open the question as to what is the source(s) of the gay earnings premium.

Table 1: Recent findings of annual gay and lesbian earnings gap

Study	Data sources	Log annual earnings/income	
		Men	Women
Waite and Denier (2015)[ii]	Canadian Census 2006	-0.074***	0.091***
		(0.011)	(0.011)
Cerf (2016)[i]	Canadian CHS 2003-9	-0.090	0.027
		(0.057)	(0.044)
Cerf (2016)[ii]	Canadian CHS 2003-9	-0.130**	0.079*
		(0.066)	(0.061)
Carpenter and Eppink (2017)	NHIS 2013-2015	0.097**	0.086**
		(0.038)	(0.043)
Jepsen and Jepsen (2017)[ii]	ACS 2007-2011	-0.111***	0.107***
		(0.004)	(0.004)
Martel and Hansen (2017)	GSS 2008-2012		-0.339
			(-0.305)
Martel (2018)[ii]	ACS 2013-2015	-0.115***	
		(0.009)	
Dilmaghani (2018)	CADUMS 2008-2012	0.045	0.116*
		(0.066)	(0.058)

Note: *p<0.1; **p<0.05; ***p<0.01. Standard errors in bracket.
(i) Single/never married sample, (ii) married/partnered sample

The literature suggests a few additional factors that Carpenter and Eppink have not considered. For instance, due to confidentiality, the only geographical data available is at regional level. As such, Carpenter and Eppink cannot control for urban or rural living conditions. To the extent that gay men and lesbian women (and LGBTQ people *writ large*) tend to migrate to and live in urban metropolitans, the omission of any urban/rural control will lead to an overestimate of gay and lesbian earnings while in fact such higher earnings is merely compensation for higher costs of living in the big cities (Black et al. (2007), Jepsen and Jepsen (2015), Jepsen and Jepsen (2017), Waite and Denier (2015)).

Another possibility is the so-called collider effect, i.e. endogenous selection bias result from conditioning on occupations and industries in the earnings regression (Elwert and Winship (2014)). Previous research has indicated that there is a high degree of occupational segregation between gay and straight men. Gay men are more likely to work in female-majority occupations, lesbians are more likely to work in male-majority occupations, and both are over-represented in occupations that provide more independence and capacity to actively manage the disclosure of their sexual orientation (Tilcsik et al. (2015), Martell (2018)). Tentative evidence of collider effect in Carpenter and Eppink's analysis comes from the fact that the unconditional earnings gap between gay and straight men is smaller in magnitude and less precisely estimated then the gap found when job characteristics (including industry and occupation controls) are accounted for (see table 1 and 2 in Carpenter

and Eppink (2017)).

For the purpose of this paper, I will bracket off the issue with urban/rural control, and focus on the role of occupation and industry. Methodologically, we need to critically assess whether occupational and industry dummies ought to be included in a wage regression, as Elwert and Winship (2014) has cautioned. Theoretically, insofar as occupational segregation is motivated by fear of homophobia, discrimination and harassment, then comparing the wage gap within an occupation would understate the true extent of discrimination. And on a practical level, even large dataset often contains only a small number of gay and lesbian workers (regardless of whether they self-identified or being identified by researchers through the sex/gender of their partners). Depending on this number, when we include too many occupation dummies and industry dummies, we effectively cut the data so finely that within occupation-industry comparison may be biased by the small number of gays and lesbians in each occupation-industry cell. I will return to these consideration in section 5 of this paper.

3 Carpenter and Eppink's findings

Carpenter and Eppink (2017) uses pooled data from the 2013-2015 waves of the National Health Interview Survey (NHIS). The NHIS is an annual survey of about 35,000 households in the US. While this is designed primarily for public health research purposes, it also contains questions about respondents' demographics and labor market characteristics such as age, race, ethnicity, education, relationship status, labor force status, full time / part time employment or unemployment status, annual earnings from all sources, and the occupation, industry, workplace, and employment conditions of currently employed or retired respondents. Most crucially, the survey asks adult respondents to self-identify their sexual orientation by choosing one of the following categories: heterosexual/straight, homosexual/gay/lesbian, bisexual, others/something else, don't know, refuse to answer, or no response. The data contains months and years of interview and regional indicator for the respondent's location (Northeast, West, South, Midwest) but no further disaggregated geographical data (such as state, county, rural/metropolitan) are available in the public use files.

In their paper, there are two main regression models: a linear probability model of employment and a wage regression model. For the earning regression model, the functional form is as follows.

$$\text{Log Earnings}_i = \alpha + \beta_1 SOGI_i + \beta_2 X_i + \epsilon_i \tag{1}$$

where $LogEarnings_i$ is the natural log of annual earnings and $SOGI_i$ is a vector of sexual orientation dummies (homosexual, bisexual, other, don't know, refused, or missing, with the reference category is heterosexual). X_i is a vector of controls that includes (i) month and year of survey dummies and (ii) demographic characteristics: age and age square, race dummies (indicators for black only, American Indian or Alaskan Native only, Asian

only, race group not releasable, and multiple race, with white as the excluded category); Hispanic ethnicity; education dummies (less than high school degree, some college, associate degree, Bachelor degree or more, do not know educational attainment, and refused to provide educational attainment, with high school degree as the excluded category); relationship/marital status (widowed, divorced, separated, partnered [married or living with a partner], and missing marital status, with never married as the excluded category); a dummy variable for any children 0–5 in the household; a dummy variable for any children 6–17 in the household; and region dummy variables (Northeast, Midwest, and South, with West as the excluded category). In addition, the X_i vector also includes job characteristics (number of years of job tenure and its square; dummy variables for firm size (10–24, 25–49, 50–99, 100–249, 250–499, 500–999, 1000 or more, do not know firm size, refused to provide firm size, and missing firm size, with less than 10 workers as the excluded category); sector of employment (indicators for public sector, do not know sector, refused to provide sector, and missing sector, with private sector as the excluded category); 24 industry dummies; and 26 occupation dummies) and a dummy variable for whether workers' personal earnings or job tenure responses are topcoded. Sample is restricted to adults aged 25-64, with full-time jobs, and non-missing, non-imputed income.

Using the same data as Carpenter and Eppinks, I was able to replicate their data sample and all of their main findings.[1] Replicated results are shown in table 2 below. The only notable difference between this and their original paper is in the baseline specification controlling for sexual orientation and month and year interviews, I estimated the earning premium for gay men to be 0.078 log point and not statistically significant (p-value is 0.158).In contrast, Carpenter and Eppinks find a 0.077 log point premium and statistically significant at 10% (see table 2, column 5 in original paper). For the full model that includes demographics and job characteristics, we both find a nearly 10% earnings premium for gay men, and it is statistically significant at 5% level.

Carpenter and Eppink also estimate their earnings regression with a number of stratified subsamples to explore heterogeneity in the earnings gap (see table 3 in original paper). However, they use the same specification as detailed above. As a robustness check, I modify their specification slightly as followed:

- Excluding from regression sample all observations that answers "Don't know", "Refused", or missing responses to question about education, relationship status, tenure, firm size, sector, occupation, and industry.
- Excluding from regression sample industry and occupation related to the armed force (occupation numbered 23 and industry numbered 21).
- Excluding from regression sample the Agriculture, Forestry, Fishing, and Hunting In-

[1]The authors graciously shared with me their data and the code they use to clean and assemble their dataset. I downloaded the NHIS data myself, and write my own data processing code. I have checked that our procedures and final data sample are similar.

Table 2: Earnings regression - replicating Carpenter and Eppink

	Dependent variable: Log annual earnings			
	Female, baseline	Male, baseline	Female, full model	Male, full model
	(1)	(2)	(3)	(4)
Homosexual	0.159***	0.078	0.087**	0.095**
	(0.052)	(0.055)	(0.043)	(0.037)
Bisexual	−0.083	−0.099	−0.028	−0.009
	(0.112)	(0.105)	(0.091)	(0.068)
Controls included:				
Month & year interview dummies	Yes	Yes	Yes	Yes
Demographics			Yes	Yes
Job characteristics			Yes	Yes
Observations	17,255	19,387	17,224	19,336
R^2	0.003	0.016	0.354	0.349

Note: *p<0.1; **p<0.05; ***p<0.01. Robust standard errors in bracket.

dustries (industry numbered 1), and the Farming, Fishing, and Forestry Occupations (occupation numbered 18) as there is no gay men in the sample working in these industries and occupations.

- Considering separately the effect of including demographic characteristics, and including labor market characteristics.
- Considering separately the effect of including occupation controls and industry controls.
- In addition, estimate model with all various subsamples shown in Carpenter and Eppink table 3.

The results for male workers are presented in table 3 below.

Panel (A) of table 3 shows the estimated gay earnings premium using the entire male workers sample[2]. In a baseline specification which includes only sexual orientation dummies and controls for years and months of the interviews, the gay earnings gap is 0.70 log point and not statistically significant. Notice that this is slightly smaller than the baseline coefficient estimates by Carpenter and Eppinks (0.077, statistically significant at 10%). When demographics control (age, race and ethnicity, education, relationship status, having children, region) are included, the earnings gap actually shrink to 0.063 log point and not statistically different from zero. However when labor market characteristics are included, the earnings gap expands to 0.088 log point and statistically significant at 5%. Although

[2]Note that the sample is modified by removed certain observations, as explained above

Table 3: Gay earnings gap - robustness check

Specifications		Coef. on Gay dummy
	(A)	
Baseline (interviews dummy only)	$N_{gay} = 430$	0.070
		(0.049)
Baseline + demographic	$N_{gay} = 430$	0.063
		(0.046)
Full model	$N_{gay} = 430$	0.088**
		(0.038)
	(B)	
Young - Aged 25-44	$N_{gay} = 248$	0.032
		(0.054)
Old - Aged 45-64	$N_{gay} = 182$	0.170***
		(0.053)
Partnered workers	$N_{gay} = 145$	0.040
		(0.050)
Single workers	$N_{gay} = 256$	0.159**
		(0.065)
Workers with less than BA degrees	$N_{gay} = 193$	0.087
		(0.062)
Workers with BA degrees or more	$N_{gay} = 237$	0.105**
		(0.053)
Public sector	$N_{gay} = 65$	0.090
		(0.058)
Private sector	$N_{gay} = 365$	0.096**
		(0.043)
At least 500 workers at firms	$N_{gay} = 106$	0.150***
		(0.052)
Less than 500 workers at firms	$N_{gay} = 324$	0.072
		(0.048)
	(C)	
No occupation dummies	$N_{gay} = 430$	0.093**
		(0.039)
No industry dummies	$N_{gay} = 430$	0.081*
		(0.044)
No occupation and industry dummies	$N_{gay} = 430$	0.056
		(0.041)

Note: *p<0.1; **p<0.05; ***p<0.01. Robust standard errors in bracket.
Unless specified, model used is the full saturated earning regression

this is smaller than Carpenter and Eppink's estimate, it is of the same order of magnitude. More importantly, the fact that demographic controls help narrowing the earnings gaps while labor market controls widen it is lost in Carpenter and Eppink analysis.

Panel (B) of table 3 shows regression result, using the full models with both demographic and labor market controls, for various subsample. Here my results are fairly close to Carpenter and Eppink's result in their table 3. Note that the coefficients on gay men earnings is consistently large (in some case implied a nearly 20% earnings premium) across subsamples, but only statistically significant for older workers, single workers, those who have completed at least post-secondary education, those who work in the private sector, and those who work for bigger firms. In panel (C), I explore the impact of including occupation and industry control. As shown here, we observe a statistically significant gay earnings premium when either occupation or industry dummies are included in the regression. But interestingly, when industry and occupation dummies are not included, the earning gap shrinks to only 0.056 log points and not statistically significant. Juxtapose this result with the results in panel (A), we see that when demographics and certain labor market characteristics are controlled for, the unconditional wage gap is reduced, and the coefficients are never precisely estimated. However, by adding in a set of occupation and industry dummies, the magnitude of the earnings gap nearly double and corresponding standard error becomes smaller. We can conclude then, that much of what driving Carpenter and Eppink's finding is in this set of occupation and industry controls. I will return to this point in section 5.

4 Oaxaca decomposition

A different way to ascertain the sources of the observed earnings premium for gay men is to evaluate relative impact of various demographics and labor market characteristics using Oaxaca-Blinder decomposition (Blinder (1973), Oaxaca (1973)). The raw gap in weighted average annual earnings between gay and straight men is 0.067 log points, or equivalent to a gay earnings premium of nearly 7 percentage point. I decompose this total gap as follows.

$$\begin{aligned}\overline{\text{Log Earnings}_{gay}} - \overline{\text{Log Earnings}_{straight}} &= \beta_{gay}\overline{X_{gay}} - \beta_{straight}\overline{X_{straight}} \\ &= \beta_{gay}(\overline{X_{gay}} - \overline{X_{straight}}) + (\beta_{gay} - \beta_{straight})\overline{X_{straight}} \\ &= \text{Difference in characteristics} - \text{Difference in returns}\end{aligned} \tag{2}$$

The total difference in average annual log earnings is decomposed into two components. The first is difference in mean characteristics of gay and straight men, evaluated at the coefficients retrieved from earning regressions of the gay men subsample. The second is difference in coefficients of the earnings regression equation for gay and straight men (estimated separately), weighted by the average characteristics of straight men. To the extent

that the total earnings gap can be explained by differences in demographic and labor market characteristics, this will show up in the first component (often referred to as the "explained" part of the earnings gap). To the extent that gay and straight men, conditional of having the same characteristic, receive differential (potentially discriminatory) returns to their characteristics, this will show up in the second component (commonly referred to as the "unexplained" part of the earnings gap).

To obtain the neccessary coefficients, I estimate the earnings regression separately for gay men and straight men, using the specification detailed in (1) and making the appropriate sample adjustment as mentioned in section 3. The decomposition result is presented in table 4 below.[3]

Table 4: Oaxaca-Blinder decomposition of gay men's wage premium

Decomposition	Attributable to difference in:		% Relative to total gap	
	Characteristics	Returns	Characteristics	Returns
	(1)	(2)	(3)	(4)
Age	-0.011	1.078	16	1,618
Race and ethnicity	-0.019	0.074	29	112
Education	0.058	-0.034	87	52
Relationship	-0.018	-0.098	27	147
Children	-0.163	0.181	245	272
Region	0.001	-0.033	2	49
Tenure	-0.013	-0.110	20	166
Sector	0.000	0.010	0	15
Firm size	0.009	-0.198	13	297
Industry	-0.089	0.240	134	360
Occupation	0.058	-0.494	88	742
Survey wave	0.005	-0.113	7	170
Intercept	0.000	-0.252	0	379
Total	-0.183	0.249	-275	375

As we can see from the last row of table 4, the 0.067 log point in total earnings gap is due to an 0.183 log point penalty attributable to difference in characteristics (column (1)), and an 0.249 log point premium attributable to returns (column (2)). In other words, should straight men have the returns to characteristics as gay men do, we would expect their earnings to be 0.183 log point (20 percentage point) higher. On the other hand, gay men seem to have much higher returns to characteristics than straight men. If both groups share the same average characteristics as straight men currently do, then the differential returns would results in an even bigger gay earnings premium (0.249 log points or 28 percentage point higher). Columns (3) and (4) of the table show the magnitude of each source of earnings difference relative to the total size of the earnings gap.

[3]For brevity, regression results and detailed decomposition are not presented here but available upon request.

Looking at the sources of earnings difference, several unusual points emerges. First, in terms of difference in characteristics, the single most influential factor here is having children. If gay and straight father receives the same returns for having children, the difference in average number of children suggests that gay men would earn 0.163 log points lower (a magnitude nearly 2.5 times the size of the total earnings gap). This, however, is mainly due to the fact that the share of gay men having children is much smaller than that of straight men (8% vs 55%), while the coefficient for children in gay men earning regression is very large and imprecisely estimated. Other notable differences in characteristics are education, industry, and occupations, but note that their contribution is in different direction. Differences in educations and industry reduces the gay earnings gap, while differences in occupations widen it. Overall, the result from column (1) suggests that straight men possess characteristics that are more conducive to higher earnings.They would have earns up to 20 percentage point higher than gay men if both groups receive the same returns to their characteristics.

Second, in terms of difference in returns, the single most influential factor here is age. Gay men seems to command substantially higher returns to age - almost twice as much as older working straight men. This is driven by the fact that the coefficient for age in gay and straight men' earnings equations is 0.096 and 0.058 respectively (both statistically significant at 1% significant level). Yet, if this reflects the rewards for experience, then it is puzzling to see that the return to tenure (numbers of years spent at current jobs) for gay men is 0.11 log point lower than that for straight men. Other important sources of earnings difference here include the return to occupation, industry, firm size, having children, and having (or have had) relationship. But again we see that different factors contribute in different direction. Relationships narrows the earnings gap between gay and straight, but having children widens it. Conditional on being in the same industry, gay men seems to earn more than straight men (0.240 log point or 27 percentage point more), yet for a given occupation the former are paid far less (0.494 log point or nearly 64 percentage point less).

Given the contradiction and inconsistency of the detailed decomposition results, the aggregated decomposition result that there is gay men enjoy a substantially higher returns to characteristics becomes tenuous. There is even more doubt when we view the Oaxaca decomposition in conjunction with the earnings regression results. In particular, for gay men earnings equation, most of the coefficients are not statistically significant, including some of the most influential sources of the gay earnings premium such as relationship, children, and occupational dummies. On the one hand, judging from the adjusted R-square, more than 70% of the variations in earnings are not accounted by the set of current explanatory variables. On the other hand, I also contend that the fully saturated model includes too many controls given the sample size. There are 430 gay men in the sample, but the model has 18 industry dummies and 20 occupational dummies, let alone other controls. The number of observations in each cell is therefore too small and estimation too imprecise to

be informative.

5 Occupation and industry: distribution and earnings

In previous sections, results from earnings regression and Oaxaca decomposition both point to occupation and industry as important drivers of the gay earnings premium, for two distinct reasons. First, there might be endogenous selection bias, where occupation and industry segregation operates such that gay men are are over-represented in higher paying jobs and receives higher earnings than straight men, such that the unconditional earnings gap is positive and the conditional gap is even bigger. Second, there might be small sample bias, where in each occupation-industry cell there might be too fews observation of gay men. This will affect the external validity of the findings. Moreover, to the extent that gay men in the dataset might not be representative of all gay men in the population (e.g. if only highly paid gay men in high paying occupation self-identify their sexual orientation in the NHIS survey), this will also cast doubt about the internal validity of the gay earnings premium.

To explore this, tables 5 and 6 below show the number of gay and straight men in each industry and each occupation, together with their weighted average earnings. To be consistent with previous section, I limit the sample to adults aged 25-64 with full-time jobs, non-imputed and non-missing income. Occupations and industries related to the arms forces or where there is no gay men working have been excluded, and other sample restriction described in section 3 also apply.

In table 5, we can see immediately that in most industries, the number of self-identified gay men workers are very low. Out of 19 industries present, only 6 or one-third of them have more than 30 gay workers. Second, looking at the share of gay men in each industry and compare it against the share of straight men, we see evidence of occupational segregation. Gay men are over-represented in Professional and technical services, accommodation and food services, Retail trade, Educational services, and Healthcare and social assistance. Straight men, on the other hand, are over-represented in Construction and Manufacturing. To confirm this observation, a two-sample Kolmogorov-Smirnov test rejects the null hypothesis that the industry distribution of gay and straight men are drawn from the same distribution. Figure 1 provides a visualization of these information, with the red dotted line marking the last industry for which there is a positive difference between gay men's earnings and straight men's earnings. Overall, 46.5% of gay men and 45.4% of straight men are working in industries in which gay men earn a premium over straight men.

Figure 1 - Distribution of workers by industry

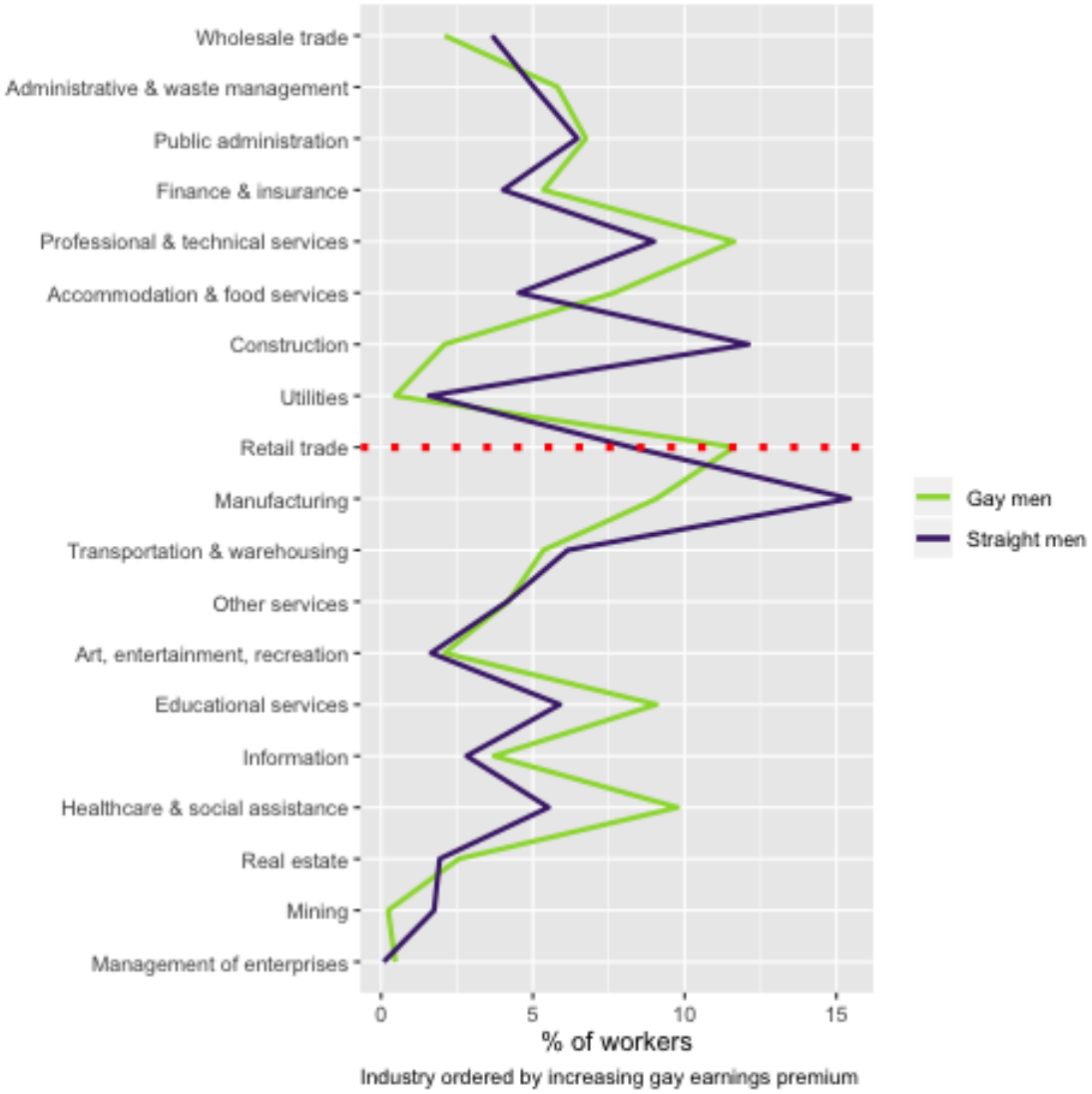

Table 5: Distribution of workers and earnings by industry

Industry	Gay men		Straight men		Average earnings		
	Count	% of total	Count	% of total	Gay	Straight	Difference
Wholesale trade	9	2.1	635	3.7	105,258	65,048	40,211
Administrative & waste management	25	5.8	873	5.0	63,372	43,513	19,859
Public administration	29	6.7	1122	6.5	89,930	73,883	16,046
Finance & insurance	23	5.3	698	4.0	107,692	93,296	14,396
Professional & technical services	50	11.6	1,561	9.0	100,469	88,056	12,413
Accommodation & food services	33	7.7	788	4.5	44,089	34,375	9,714
Construction	9	2.1	2,100	12.1	57,838	48,445	9,393
Utilities	2	0.5	274	1.6	77,684	76,201	1,483
Retail trade	50	11.6	1,431	8.2	52,826	51,455	1,371
Manufacturing	39	9.1	2,680	15.4	60,887	61,376	-489
Transportation & warehousing	23	5.3	1,067	6.1	51,898	55,404	-3,507
Other services	18	4.2	721	4.2	43,263	47,041	-3,778
Art, entertainment, recreation	9	2.1	286	1.6	41,760	45,995	-4,234
Educational services	39	9.1	1,022	5.9	52,101	58,267	-6,166
Information	16	3.7	489	2.8	67,217	76,226	-9,008
Healthcare & social assistance	42	9.8	957	5.5	55,921	69,002	-13,081
Real estate	11	2.6	335	1.9	46,174	62,382	-16,208
Mining	1	0.2	303	1.7	65,000	87,107	-22,107
Management of enterprises	2	0.5	15	0.1	50,311	92,962	-42,651

*Note:*Industries are in descending order of the difference in weighted average earnings between gay and straight workers.

Turning next to occupation, from table 6 we also notice that most occupations have very gay men working in: out of 21 occupations, only four have more than 30 gay men. There is some evidence of occupation segregation: gay men are over-represented in Business and finance, Office and administrative support, Sales, Management, and Art and entertainment. They are under-represented in Installation, maintenance and repair, Architecture and engineering, Construction and extraction, Production, and Transportation. A two-sample Kolmogorov-Smirnov test also rejects the null hypothesis that the occupational distribution of gay men and straight men are identical. Figure 2 provides a visualization of the result. Overall, 59.1% of gay men and 55.1% of straight men are working in occupations in which gay men earn a premium over straight men.

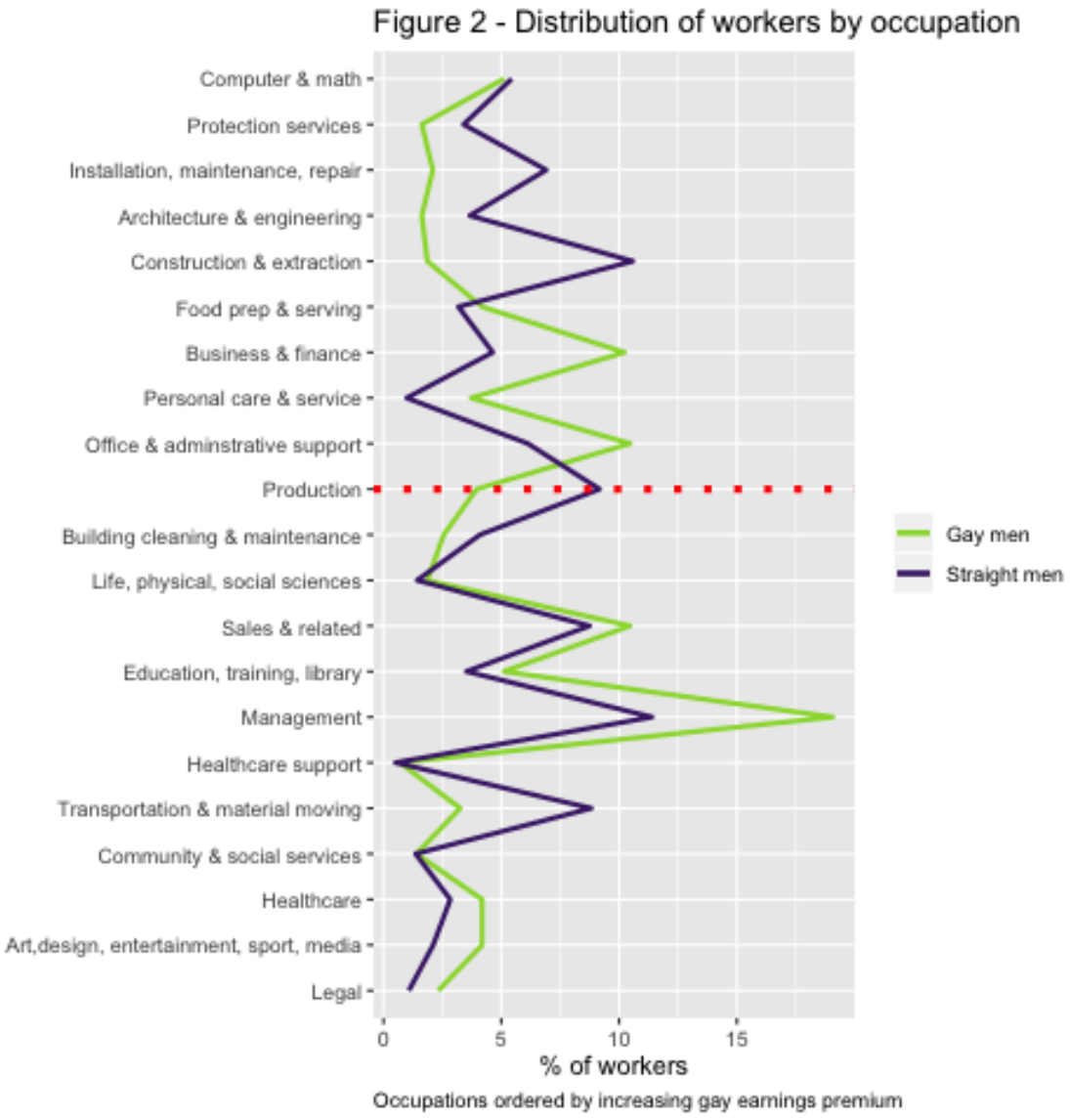

Table 6: Distribution of workers and earnings by occupation

Occupation	Gay men		Straight men		Average earnings		
	Count	% of total	Count	% of total	Gay	Straight	Difference
Computer & Math	22	5.1	945	5.4	111,545	86,752	24,793
Protection services	7	1.6	590	3.4	84,335	63,288	21,047
Installation, maintenance, repair	9	2.1	1,199	6.9	70,812	51,319	19,494
Architecture & engineering	7	1.6	636	3.7	102,111	84,387	17,724
Construction & extraction	8	1.9	1,837	10.6	54,441	44,385	10,056
Food prep & serving	18	4.2	552	3.2	34,105	28,915	5,190
Business & finance	44	10.2	805	4.6	83,826	82,111	1,715
Personal care & service	16	3.7	171	1.0	40,165	38,493	1,672
Office & adminstrative support	45	10.5	1,063	6.1	48,447	48,154	292
Production	17	4.0	1592	9.2	42,840	44,031	-1,191
Building cleaning & maintenance	11	2.6	711	4.1	30,358	32,954	-2,596
Life, physical, social sciences	8	1.9	248	1.4	81,287	84,169	-2,883
Sales & related	45	10.5	1518	8.7	64,551	67,845	-3,294
Education, training, library	22	5.1	611	3.5	54,910	59,584	-4,673
Management	82	19.1	1978	11.4	79,369	89,020	-9,651
Healthcare support	3	0.7	88	0.5	25,486	35,226	-9,740
Transportation & material moving	14	3.3	1,531	8.8	31,936	43,780	-11,844
Community & social services	6	1.4	237	1.4	38,364	50,823	-12,459
Healthcare	18	4.2	492	2.8	79,399	92,197	-12,798
Art,design, entertainment, sport, media	18	4.2	366	2.1	44,095	65,414	-21,319
Legal	10	2.3	187	1.1	86,401	110,686	-24,285

*Note:*Occupations are in descending order of the difference in weighted average earnings between gay and straight workers.

However, the pattern of industry and occupation segregation in the sample does not easily lend itself to an explanation for the overall (unconditional) gay men earnings premium. We notice that gay men are concentrated in both occupations and industries where they seem to enjoy an earnings advantage and those where they seem to be paid much less than straight men. From the descriptive analysis in this section, I find no immediately discernible systematic sorting of gay men that we can pinpoint to as the driver of the earnings premium. At the same time, the observation that in most industries and occupations, the number of self-identifying gay men are very small suggest that a regression model with full set of controls for industries and occupations may produce inconsistent estimates. When I restrict the sample to workers in industries or occupations with at least 30 gay men working, the earnings regression result is qualitatively similar to that shown in table 3 but the magnitude of the estimated coefficient is smaller and the statistical precision is also reduced. This is summarised in table 7 below.

Table 7: Earnings regression - industries and occupations with $N_{gay} \geq 30$

Specifications	Coef. on Gay dummy
Baseline (interviews dummy only)	0.027
	(0.057)
Baseline + demographic	0.068
	(0.052)
Full model without occupation and industry control	0.065
	(0.047)
Full model	0.085*
	(0.044)

Note: *p<0.1; **p<0.05; ***p<0.01. Robust standard errors in bracket.
N = 11,894; $N_{gay} = 341$

6 Conclusion

In this paper, I revisit Carpenter and Eppink (2017). I show that their finding of statistically significant 10% annual earnings premium for gay men over straight men is driven primarily by including industry and occupation controls in the earnings regression model. I have also presented evidence indicative of endogenous sorting into occupation and industry, as well as the possible small sample bias resulting from having too few gay workers in most occupation-industry cells. Both of these reasons suggest that we ought to be more cautious and thoughtful in thinking about the earnings equation.

One point that has not been discussed thus far is the unconditional earnings gap between gay and straight men. The fact that it is positive is unusual on its own, and this points us to different explanation for Carpenter and Eppinks. Is it possible that the self-identifying

gay men in the NHIS data are not representative of the population? One way to approach this question is to compare the distribution of gay men's earnings, demographics and labor market characteristics in NHIS data against similar distribution from a different dataset. Unfortunately there is currently no "gold standard" for data on LGBTQ workers. Nationally representative and large scale survey like the Census, the CPS or the ACS do not ask self-identifying question about sexuality. Researchers using those data source have had to rely on the sex/gender of spouses or cohabitating partners in order to infer sexuality, which leaves us in the dark about non-partnered gay, lesbian, and bisexual workers. We also don't know much about pattern of selection into partnership of queer and straight people. On the other hand, health surveys such as the NHIS, or the Behavioral Risk Factor surveillance system (BRFSS) have incorporated self-indentifying questions but as this paper have shown, it is not certain that their socioeconomic variables are accurate or representative. Nonetheless, as a next step it might be worthwhile for researchers to systematically compare the gay-straight wage gap found in the NHIS data with those found in the BRFSS, the CPS, ACS, and other data sources while making allowance for appropriate sample comparison.

References

Babcock, L., Recalde, M. P., Verterlund, L., and Weingart, L. (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107(3):714–747.

Badgett, M. (1995). The wage effect of sexual orientation discrimination. *ILR Review*, 48(4):726–739.

Becker, G. (1957). *The economics of discrimination*. University of Chicago Press.

Black, D. A., Sanders, S. G., and Taylor, L. J. (2007). The economics of lesbian and gay families. *Journal of Economic Perspective*, 21(2):53–70.

Blau, F. and Kahn, L. (2017). The gender wage gap: Extent, trends, and explanation. *Journal of Economic Literature*, 55(3):789–865.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(436-455).

Carpenter, C. S. and Eppink, S. T. (2017). Does it get better? recent estimates of sexual orientation and earnings in the United States. *Southern Economic Journal*, 84(2):426–441.

Cerf, B. (2016). Sexual orientation, income, and stress at work. *Industrial Relation*, 55(4):546–575.

Dilmaghabi, M. (2018). Sexual orientation, labour supply and occupational sorting in canada. *Industrial Relations Journal,*, 49:298–318.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53.

Jepsen, C. A. and Jepsen, L. K. (2015). Labor-market specialization within same-sex and difference-sex couples. *Industrial Relation*, 54(1):109–130.

Jepsen, C. A. and Jepsen, L. K. (2017). Self-employment, earnings, and sexual orientation. *Review of Economics of the Household*, 15:287–305.

Klawitter, M. (2015). Meta-analysis of the effects of sexual orientation on earnings. *Industrial Relation*, 54(1):4–32.

Martell, M. (2018). Identity management: Worker independence and discrimination against gay men. *Contemporary Economic Policy*, 36(1):136–148.

Mueller, R. E. (2014). Wage differentials of males and females in same-sex and different-sex couples in Canada, 2006-2010. *Canadian Studies in Population*, pages 105–116.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14:693–709.

Tilcsik, A., Anteby, M., and Knight, C. R. (2015). Concealable stigma and occupational segregation: Toward a theory of gay and lesbian occupations. *Administrative Science Quarterly*, 60(3):446–481.

Waite, S. and Denier, N. (2015). Gay pay for straight work: Mechanism generating disadvantage. *Gender & Society*, 29(4):561–588.

Duc Hien Nguyen

Department of Economics

University of Massachusetts

412 North Pleasant Street

Amherst, MA 01002

Email: duchiennguye@umass.edu