Do online job postings capture job vacancies?*

Michael R. Dalton
Bureau of Labor Statistics

Lisa B. Kahn
University of Rochester, NBER and IZA

Andreas I. Mueller
University of Zurich, CEPR and IZA

October 31, 2025

WORK IN PROGRESS -
PLEASE DO NOT CITE/CIRCULATE WITHOUT PERMISSION

Abstract

This paper analyzes the relationship between online job postings and job vacancies in the U.S. labor market, using near-universal online job postings data from Burning Glass (BG) matched to representative data from the Job Openings and Labor Turnover Survey (JOLTS). We document differential representation of BG postings and JOLTS openings over time and across establishment characteristics. We find BG data converged to the JOLTS series over our study period of 2007 and 2010-2022, with the largest changes occurring early in the sample period. We also document substantial heterogeneity in representativeness across industries, locations, and establishment wage and size. We propose an adjustment weight that aligns BG postings with the representative JOLTS openings series along establishment characteristics and will make a version of these weights publicly available. We then examine how conclusions previous researchers have drawn from BG data are impacted by our adjustment. While the level and cyclicality of demand for skill hold up well to our adjustment, the level of market concentration as measured by the Herfindahl-Hirschman Index is dramatically overstated in BG because small establishments are less likely to post their positions online.

*We thank Christina Patterson and seminar participants at the AEA's and Richmond Fed for helpful comments. We are especially indebted to Matt Sigelman and Bledi Taska for providing the Burning Glass data. All views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Bureau of Labor Statistics.

1 Introduction

Job vacancies take a central role in the theory and the measurement of the state of the labor market. The job vacancy rate is an important gauge of firms' labor demand and its relationship to the unemployment rate is the key determinant of job-finding rates in the canonical search-and-matching model. With the advent of the internet and online job boards, firms have posted their vacancies increasingly online, leading to the development of aggregate indicators based on online job postings, such as the Help-Wanted Online Index (HWOL) in the U.S., as well as a surge in research activity based on the underlying micro data. An important advantage of online job board data is the rich information contained in the job ads, such as job requirements and posted wages, which allows one to analyze firms' labor demand in unprecedented detail. At the same time, little is known about the representativeness of online job ads data.

In this paper, we link online job postings to representative data on job vacancies at the establishment level to better understand alignment of the two data series. We make use of the increasingly popular, near-universal database of online job ads in the U.S. known as Burning Glass (BG) and now collected by a company called Lightcast, as well as the Job Openings and Labor Turnover Survey (JOLTS) administered by the Bureau of Labor Statistics. We document differential representation of BG postings and JOLTS openings over time and across establishment characteristics. We find BG data converged to the JOLTS series over our study period of 2007 and 2010-2022, with the largest changes occurring early in the sample period. We also document substantial heterogeneity in representativeness across industries, locations, and establishment wage and size. We propose an adjustment weight that aligns BG postings with the representative JOLTS openings series along establishment characteristics. We then examine how conclusions previous researchers have drawn from BG data are impacted by our adjustment. Because BG represents large establishments well and because most openings are attributed to large establishments, questions about the level and cyclicality of demand for skill hold up well to our adjustment. And while BG data accurately reflect spatial variation in labor market concentration, the overall level of market concentration is dramatically overstated in BG because small establishments are less likely to post their positions online. Our research thus validates the use of BG data (or similar large-scale online jobs data) for many purposes and we advise using our adjustment weights, which we will make available, for questions where establishment-level representativeness is important.

We begin by matching BG postings to business registry data for the U.S. covering 95 percent of jobs at the firm name-by-location level (the Quarterly Census of Employment and Wages, QCEW). We use a sophisticated matching algorithm that targets exact matches on a regularized firm name or a fuzzy match validated using location and industry information. We are able to match 90% of BG postings to an establishment in the QCEW.[1] We employ a battery of checks to ensure a high level of accuracy of our merge and rule out false positives and through this process chose to restrict attention to private sector establishments with at least 2 employees. The QCEW serves as the sampling frame for the JOLTS and thus we can merge the matched BG-QCEW records to JOLTS seamlessly. JOLTS asks employers to report their stock of open vacancies at the end of the month as well as the number of hires over the month. Thanks to this panel nature and using the law of motion, we infer the flow of new openings in JOLTS in a given month – which better aligns with the nature of postings in the BG data.

We develop a conceptual framework that relates BG postings to JOLTS openings through a multiplicative adjustment factor comprising: (1) an extensive margin – the probability of posting a job vacancy online, (2)

[1] A separate issue is that 30 percent of postings in BG lack information on firm name or location and thus we cannot match these to establishments in JOLTS.

an intensive margin – the number of vacancies captured by one online job posting, and (3) a nuisance factor – capturing the share of postings that originate from firms that actually have job vacancies. The adjustment factor and its components provide a useful lens for describing variation in the postings-openings relationship over time and across establishment-level characteristics – variation that could not be examined with BG data alone. We also propose a prediction method that reweights BG postings to generate moments that are representative of JOLTS openings. This prediction allows us to move beyond the BG-JOLTS merge to the full business registry (QCEW) and address research questions for which the JOLTS is underpowered. We show that a reweighting leveraging only information contained in BG (and other large scale online postings data) performs quite well and will make these weights available for researchers.

Our analysis of the matched micro-level data reveals a number of facts: First, the representation of BG data rose steadily from 2007 to 2022, with the largest changes happening in the earliest years of the sample. As a summary measure, our adjustment factor declined eight-fold between 2007 and 2012 and while it still declines from there, the movement is much more gradual, halving over the next decade. Convergence was primarily driven by the extensive margin – the likelihood that an establishment with openings posts them online – but also because the intensive margin (how many openings a given posting corresponds to) fell modestly. Second, the adjustment factor decreases strongly in the average wage and urbanicity of the establishment; higher paying establishments and urban areas are better represented online. To a lesser extent the adjustment factor is decreasing in employment size, but this relationship is key for understanding the postings-openings relationship. Large firms are well represented online and make up the majority of job openings, while small firms are poorly represented online but make up the majority of firms. BG will thus represent well the distribution of openings in the economy but not the distribution of hiring firms. There is no pattern with respect to establishment age and a U-shaped but muted relationship with respect to employment growth rate. Finally, we find strong differences across industries with professional services sectors being the most well represented in BG, while lower skilled industries such as construction and leisure and hospitality sectors requiring four-fold larger adjustments.

We then apply our adjustment factor to revisit empirical patterns studied with BG data in the literature in two main applications: First, many researchers have been interested in the skill requirements of BG postings, but it is thought that BG tends to overrepresent skilled professions. We find that the skill requirements of posting are indeed lower when reweighting the BG data with our adjustment factor, but the adjustment is modest, reflecting the fact that most openings come from large establishments, which are well represented. We also replicate the analysis in Hershbein and Kahn (2018) who show that skill requirements in online job postings related to automation increased in the aftermath of the Great Recession. We probe whether a compositional shift in the types of hiring firms being captured by BG could impact their inference, especially since they had available only one year of pre-recession data to benchmark against (2007). As we show, representation of BG was changing rapidly over that period. However, we find that their results remain similar when reweighting the data with our adjustment factor.

Second, researchers have leveraged the richness of the BG data to study labor market concentration at highly disaggregated geographic and occupational levels (Azar et al., 2022; Hershbein et al., 2020; Azar et al., 2020; Schubert et al., 2023). For instance, using a Herfindahl-Hirschman Index, Azar et al. (2020) find that 60% of local labor markets are highly concentrated (an index above 2500). These advances in measurement have been instrumental in shifting academic and policy debates toward a broader focus on employer market power. We re-examine these concentration measures using our adjustment factor and find that the original estimates substantially overstate labor market concentration – by a factor of four. This discrepancy is driven

by the fact that small establishments are underrepresented in online job postings data. On the other hand, spatial variation in market concentration performs well in BG as rankings across markets are preserved fairly well after applying our adjustment. Concentration measures based solely on online postings severely exaggerate true market concentration by omitting the cumulative hiring activity of the multitude of small establishments. In fact, our adjustment shows that only 10% of markets are highly concentrated. As such, focusing on the level of concentration is ill advised without applying our adjustment.

Our paper relates to several important literatures. First, vacancy rates play a central role in the canonical search and matching models (Pissarides (1985); Mortensen and Pissarides (1994)) and are a key ingredient in the matching function, which has been subject to a long empirical literature (e.g., Petrongolo and Pissarides (2001); Lange and Papageorgiou (2020)). Vacancy rates are used to understand worker flows over the business cycle, as well as shifts in the Beveridge curve and mismatch in the economy (e.g., Abraham (1987); Blanchard and Diamond (1989); Shimer (2005, 2007); Şahin et al. (2014); Elsby et al. (2015)). However, vacancy data in the U.S. can typically only be measured at a high level of aggregation and with a substantial time lag. In their influential paper, Davis et al. (2013) leveraged JOLTS micro data to learn about recruiting intensity and vacancy fill rates at the firm level.[2] Still, being a survey, JOLTS data are simply not powered to analyze labor market dynamics at granularity.[3] We contribute to this literature by providing a method for using large scale data on online postings to inform moments that are representative of the leading source of vacancy information in the U.S.

Online job listings data have risen in popularity and are by now well-known, in particular the BG data. With several million postings each month, the data are unusually thick at disaggregated levels so make many research questions possible that would not be with publicly available data. In addition to the works mentioned above, researchers have examined demand for skill at high granularity (Deming and Kahn, 2018; Braxton and Taska, 2023; Hemelt et al., 2023; Acemoglu et al., 2022), job attributes such as the prevalence of work from home (Hansen et al., 2023), and impacts of public policy on narrowly defined groups (Dillon et al., 2025; Clemens et al., 2021; Azar et al., 2020; Callaci et al., 2024). The data are also available in real time, which helps, for instance, in our understanding of rapid cyclical movements such as those during the COVID pandemic (Forsythe et al., 2020, 2022).[4] While the data are to some extent well-aligned with external aggregate data sources (Carnevale et al., 2014), researchers have pointed out that online job postings are not representative of all vacancies. For instance the data overrepresent skilled positions. To date, researchers can compare the quantity of postings to employment or job flows at the occupation or industry level, for example, or compare worker characteristics to education, experience, and skill requirements in postings at a similarly aggregated level. However comparisons to datasets of vacancies are necessarily made at an aggregated level, such as the one-digit NAICS industry. Furthermore, little is known about the composition of firms in BG. A major contribution of this paper is to provide such comparisons.

Our BG-JOLTS linkage offers a great deal of insight by characterizing the relationship between online postings and representative vacancy data. We conclude that BG is apt for many research questions related to the distribution of openings in the economy. However, research questions that rely on understanding the full distribution of recruiting firms in the economy require caution. In practice, researchers will not have access to the firm characteristics we find most useful here. As such, we offer a reweighting of online postings data that uses only information on geography, industry, number of postings, and basic ad characteristics

[2] See also, more recently, Mongey and Violante (2019).

[3] Barnichon (2010) observes that the Help Wanted Index, which is available for earlier periods, closely tracks aggregate movements in JOLTS.

[4] Studies using data from online job board data from other sources and other countries include Barron et al. (1999), Banfi and Villena-Roldan (2019), Marinescu (2017), Marinescu and Wolthoff (2020), Mueller et al. (2024) and Mueller et al. (2025).

(occupation and education requirements). We show this reweighting does as well as an alternative method that leverages the full information contained in our BG-JOLTS linkage. We hope this can serve as a useful tool for researchers relying solely on postings data, who would like to make the data representative of the universe of job vacancies.

This paper proceeds as follows: Section 2 discusses the data sources and matching procedure in detail. Section 3 discusses conceptual issues and develops a framework for relating BG postings to new openings in JOLTS. Section 4 provides descriptive statistics on the adjustment factor at the aggregate and establishment level. Section 5 applies our adjustment factor to measures of skill requirements and measures of local labor market concentration. Section 6 concludes.

2 Data Sources and Matching Procedure

We use three primary data sources in this paper. Online job postings come from the company most well-known to academics as Burning Glass (hereafter, BG); the company has since merged and changed its name to Lightcast. We match these data at the employer-location level to two Bureau of Labor Statistics datasets: the Quarterly Census of Employment and Wages (QCEW) – business registry data that serves as the sampling frame for all BLS establishment surveys – and the Job Openings and Labor Turnover Survey (JOLTS) – a BLS survey that asks employers about their hiring needs. The purpose of these matches are to (1) identify establishment characteristics not available in the BG data and (2) compare BG job posting behavior to self-reported information on vacancies and hires. In this section, we describe each dataset, summarize our matching procedure, and provide comparison statistics.

2.1 Online Job Postings

Burning Glass attempts to collect the universe of online job postings on a rolling basis. They do so by scraping over 40,000 websites including online job boards, recruiter firms, and private company postings. BG uses proprietary machine-learning algorithms to code a wide range of potential fields for each ad (e.g., location, industry, skill requirements, etc.), and they make significant efforts to remove duplicates from their database since ads are often posted on multiple sites. We use data for 2007 and 2010 to 2022.[5]

For matching BG postings to BLS establishments, we focus on employer name-geography cells. Employer names in BG are coded when available, 70% of the time.[6] We also leverage the fact that BG impute an industry code based largely on the employer name, when available, as a way to validate our match.

From BG, we can measure the flow of new job postings. However, the number of jobs desired from a given posting is not well understand. An employer could be hoping to hire many workers from the same posting, or could be very unlikely to hire at all and keep a posting up in case an ideal candidate comes along. Another complication is that some establishments may have semi-permanent online postings (e.g., high churn fast-food positions), and keep postings online even after a person has been hired. To account for this potentially important feature of online postings, BG treats any ad that has been up for 60 days as a *new* posting.

[5] There is a gap for the years 2008 and 2009.

[6] BG indicates that the primary reason an ad would not have an employer name is because it is posted by a recruiter who wants to keep the company name confidential. There could also be errors in the algorithm that miss the name.

2.2 QCEW

The Quarterly Census of Employment and Wages is derived from wage records and is a near universe of establishments paying into the state Unemployment Insurance program, plus Federal government entities. It covers more than 95% of all jobs and serves as the sampling frame for all Bureau of Labor Statistics establishment surveys.

For each establishment, it gives the physical address, a NAICS industry code, monthly employment and the quarterly wage bill. Because establishments can be tracked over time, we also know the age of the company. These characteristics are not known in the BG data and so are useful to bring in. Merging BG to QCEW also facilitates linking variables at the establishment level for any other BLS survey.

2.3 JOLTS

The primary BLS survey of interest is the Job Openings and Labor Turnover Survey. The JOLTS is a monthly survey of approximately 16,000 establishments that asks about vacancies at the end of the month, hires and separations over the course of the month, and employment. There is also a panel component that allows for looking at changes within establishments across time. Most establishments are surveyed for 24 consecutive months in JOLTS.

We can easily connect QCEW observations to the JOLTS since they are both internal BLS datasets. Since the JOLTS is a survey, it will naturally not include all companies in QCEW or posting in BG.

We use establishment responses to the question on vacancies as a comparison to the job postings in BG. A key issue that we face is how to align the definitions of job openings in each dataset.[7] JOLTS defines openings as *"all positions that are open (not filled) on the last business day of the month"* that meet the following three requirements:

1. A specific position exists and there is work available for that position.
2. The job can start within the next 30 days.
3. There is active recruiting for the opening, which includes posting job ads online.

JOLTS also collects information on the hires during a month defined as *"all additions to payroll during a month,"* which includes newly hired workers as well as workers who have been recalled. Hires do not have to satisfy the requirements for openings in JOLTS, so it is possible that a hire in JOLTS occurred without active recruiting or for a position with a long lag between posting date and start date (e.g., those posted on Job Openings for Economists). Davis et al. (2013) in their canonical work use establishment-level data from JOLTS to measure the relationship between vacancies and hiring.

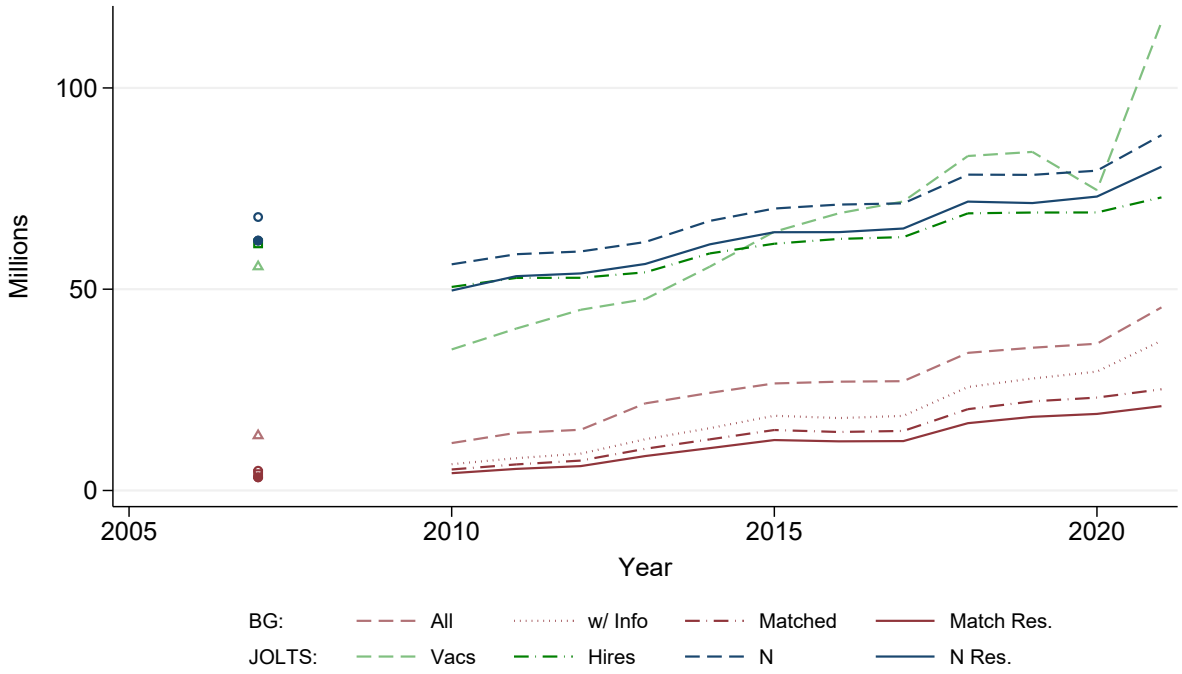
To better understand the representation of BG at a granular level, we want to compare the flow of online postings in BG to the stock of unfilled positions at the end of the month in JOLTS. We use the law of motion of the stock of unfilled positions, to solve for the flow of new openings from establishment, i, in month, t:

$$N_{i,t} = J_{i,t} - J_{i,t-1} + H_{i,t}, \tag{1}$$

where J are the unfilled vacancies at the end of the month, defined above, and H are hires over the month. Note that to the extent that some hires occur without a vacancy, equation 1 will be an overestimate of the

[7] See JOLTS documentation here https://www.bls.gov/opub/hom/jlt/concepts.htm.

Figure 1: BG and JOLTS Time Series



Notes: We plot the aggregate number of hires and vacancies (Vacs) in JOLTS as well as our imputed flow of new vacancies (N, eqn 1) and several BG series. "All" reflects all BG openings in our extract, the "w/ Info" series limits to openings with non-missing firm names and geography. "Matched" are those that match to an establishment in the QCEW and "N Res." and "Match Res." restrict the matched sample to private sector establishments with at least 2 employees.

true flow of openings. On the other hand, J alone is potentially an under-estimate of the total number of openings since it only counts the ones that remain unfilled. Hereafter, we use the term JOLTS vacancy to refer to the J (i.e., the unfilled openings at the end of the month) and JOLTS opening to refer to N as defined in equation 1.[8]

2.4 Merging BG to QCEW

For the purposes of matching, we begin with BG job ads with an identifiable employer name and geographic location. We then follow a matching algorithm detailed in appendix A. This involves cleaning employer names and then looking for exact matches on name for establishments in the QCEW within as narrow a geography as possible. If exact matches are not available, we use a fuzzy matching algorithm to find similar names, relying on common geography and industry codes to identify higher quality matches.

Figure 1 plots several relevant aggregate series. First, the dashed blue and green lines compare JOLTS aggregate vacancy (Vacs, light green) and hires (dark green) series to our imputed flow measure (N, navy) from equation 1. The flow of new openings N, tracks hires fairly closely, though the series diverge a bit in more recent years as the number of unfilled vacancies has steadily increased.

The maroon lines plot different BG openings series. The dashed line "All" shows the total number of

[8]When using JOLTS data, we will restrict attention to respondents who also had a response in the previous month, to calculate N.

vacancies we have access to. This series is about 40% the level of the JOLTS N series. We further limit our ability to capture JOLTS openings due to missing information in BG. The series "w/ Info" is only about a quarter of the overall JOLTS N level.[9] Still, within this sample, we match the vast majority of ads to establishments in QCEW. The "Matched" series is 80% of the number of BG ads with info.[10]

Because of the coverage of QCEW, most BG cells should have a match. However, employers may advertise in localities where they do not have a physical location. Trucking positions are an example and indeed we do find a slightly lower match rate for transportation occupations. We may also miss recently founded establishments that begin posting ads before they officially register as a business. Government establishments pose difficulty for us in matching because the lack of specificity of posted names. Finally, noise in the match due to typos or commonality in the employer name can generate false negatives or false positives. We found that data appeared especially misaligned for the smallest establishments and especially those with 1 employee – these represent a sizable chunk of establishments in QCEW, almost none in BG, and those that are matched look unusual in terms of the number of postings versus employment levels.

For all of these reasons, we limit our analytic sample to a restricted one that focuses on the private sector and establishments with at least 2 employees. The final "Match Res." series captures about 20% of the final JOLTS "N Res" series (solid navy), both of which make these restrictions.

Regardless of the comparison used, there is convergence between BG and JOLTS over time. In the aggregate, JOLTS outnumbered BG by 15 or 20 to 1 in the early period but by only 3 or 4 in the later years.

2.5 Descriptive Statistics

Appendix Table A.1 provides summary statistics for key variables in BG for the full sample, the sample of ads that can be matched to QCEW, the restricted matched sample – excluding the public sector and establishments with fewer than 2 employees, and finally the JOLTS matched sample. We have 150 million total postings over these years. Of these, 56% specify an education requirement – roughly evenly split between high school and college – and a similar proportion require some experience in the field. As is well known, the vast majority of ads do not specify a salary (85%). BG imputes an occupation for each position based on the job title and other information. As a broad summary, we report the distribution of ads across four major occupation groups: professional jobs, sales and administrative positions, blue collar positions, and low skilled service positions.[11] The occupation distribution skews towards professional jobs, which make up 46% of all postings.

As mentioned, most firms posting in BG should have a match to QCEW. Although we do not match all ads, Table A.1 shows reassuringly that our matched samples look very similar on ad characteristics, compared to the full BG universe, both for ads matching the QCEW and for ads posted by firms that made it into the JOLTS that month.

In contrast, not all QCEW firms will match to BG. A firm may not be recruiting and even conditional on recruiting a firm may not post online to a forum that BG captures. Table 1 provides some intuition for the different kinds of firms that show up in each group. We use establishment characteristics that are defined in the QCEW, meaning they can be measured whether or not the establishment was surveyed by JOLTS. However, in this table, we limit our attention to establishments that were sampled by JOLTS in the

[9] Between our primary sample years 2011-2019, 71% of ads have a non-missing firm name, of these only 1% are missing information on the geography of the position.

[10] There is a slightly worse match rate in post-COVID years, which we attribute to an increase in postings for remote work.

[11] Professional includes SOC codes 11-19, 23, 27, and 29. Sales and admin include 21, 25, 31, 41 (excluding 412), and 43. Low-skill services include 35-39 and 412. Blue collar are the remainder (33, 45-53).

Table 1: Characteristics of Establishments: All JOLTS versus those Recruiting or Posting

	All	Recruiting	BG Posting	BG Posting and Recruiting
Employment	19.92	50.43	46.73	123.1
Wage Bill per Emp	42727.4	39036.8	49918.8	46106.0
Annual Growth Rate	0.0188	0.0355	0.0151	0.0174
Age	15.77	16.10	16.65	17.71
Metro Area	0.853	0.859	0.909	0.920
# Establishments	79888	63632	28579	20228

Notes: Monthly summary statistics for JOLTS establishments (column 1), those JOLTS establishments that are recruiting (meaning they had hires or new vacancies in the month – $N > 0$ from equation 1, column 2), those that post in BG in the month (column 3), and those that post in BG and were identified as recruiting in JOLTS ($N > 0$). Uses JOLTS weights and restricts to 2011-19. All columns restrict to JOLTS respondents in the month that are private sector employers of at least 2 employees. Employment is the number of workers on payroll in the month, measured in QCEW. Wage is the total wage bill in a given quarter divided by average monthly employment in the quarter. Annual growth rate is the year-over-year rate of employment change for the focal month. Age is establishment's age in the month, measured in QCEW, and metro area is an indicator equaling 1 if the establishment is located in a metropolitan statistics area.

month so that we can provide a comparison between all JOLTS firms (meeting our restriction) and all those posting in BG in the year (columns 1 versus 3), and also limiting our JOLTS comparison to firms that are actually recruiting (column 2) versus BG posting firms that are actually recruiting (column 4). We define an establishment as recruiting if $N > 0$ in equation 1, meaning it had hires that month or had more unfilled vacancies at the end of the month compared to unfilled vacancies in the previous month. As such, the entire table limits to establishments that responded to the JOLTS survey in both the focal month and the previous month. We further limit to our primary sample: the 2011-2019 period and private sector establishments with at least 2 employees.

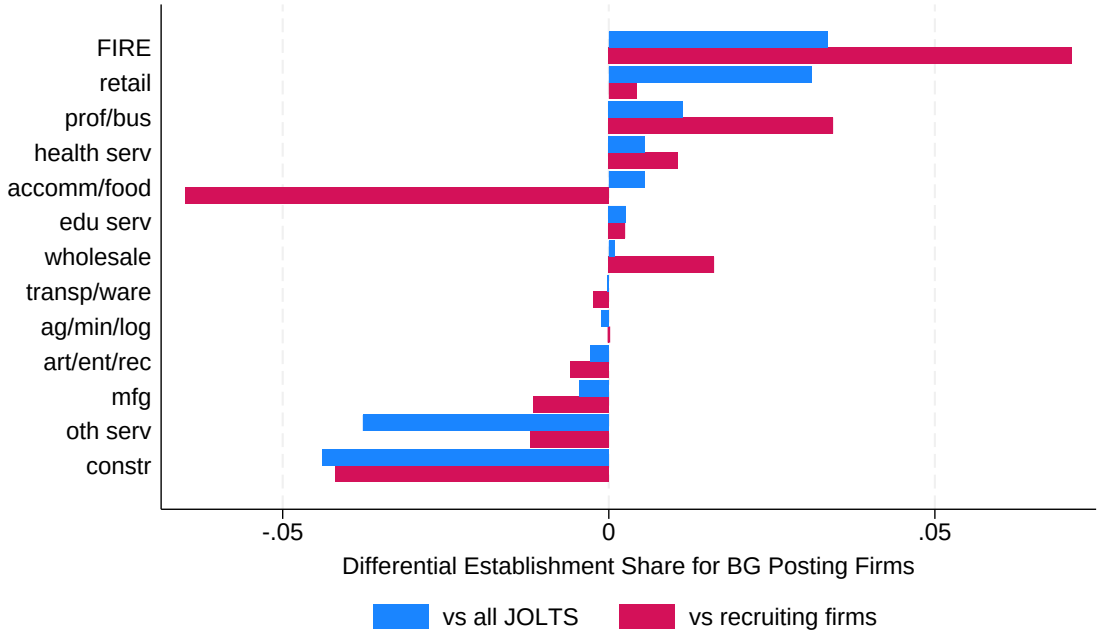
Comparing columns 1 and 2, recruiting firms are larger and growing more rapidly, relative to all firms in JOLTS. They are also slightly lower paying.[12] Firms posting in BG are much larger than the average firm in JOLTS but similarly sized to recruiting firms. Of note, BG posting firms are also higher paying, on average, and more likely to be in an urban area. They are, on average, also slower growing than either JOLTS comparison.

Note the column 4 firms – those that posted in BG and were identified as recruiting in JOLTS that month – are much bigger than any of the other comparison. Their size likely explains why they have slower growth than when restricting only to firms identified as recruiting in JOLTS (column 2). Furthermore, the difference in size foreshadows a result we will visit later, that small firms are much more likely to have BG postings without any JOLTS openings. Hence moving from columns 3 to 4 excludes many small firms. As such, for conceptualizing the relationship between JOLTS openings and BG postings, we will have to take into account not only whether a firm with openings posts online but also whether a firm that posts online indeed reports having openings.

Figure 2 shows differences by industry. The blue bars compare the share of establishments in each broad

[12]The wage concept in QCEW is incomplete. We can measure the total wage bill in a quarter and we normalize this by average monthly employment over the quarter. This calculation generates downward bias in wages of high churn firms, which could explain the slightly lower average pay in recruiting firms compared to all firms.

Figure 2: Industry Differences: All JOLTS versus those Recruiting or Posting



Notes: We compare industry distributions of JOLTS establishments that posted in BG in the month to all JOLTS establishments (blue) or those that were recruiting in the month (maroon) defined as having at least 1 hire or new vacancy ($N > 0$ in equation 1). We restrict to JOLTS respondents in the month that are private sector employers of at least 2 employees and years 2011-19. Industries are two-digit NAICS sector codes and are sorted based on the blue bars (the BG-all JOLTS comparison).

industry category in BG to the full JOLTS sample. This comparison can also be made by pairing BG with publicly available data sources (see, e.g., Hershbein and Kahn (2018); Clemens et al. (2021)). According to this comparison, BG most overrepresents retail, food and accommodation, and all of the professional services categories. The latter are high skilled sectors, while the former are high churn sectors. The red bars limit the BG comparison to JOLTS firms that are recruiting in the month, i.e., those with $N > 0$. This comparison, which is not possible without firm-level micro data on hiring, brings a great deal of clarity to the previous comparison. BG still overrepresents the high skilled professional service sectors and the differences are even exacerbated. In contrast, the high churn sectors are no longer overrepresented and in fact food and accommodation is sharply underrepresented. So, compared to the average firm, BG overrepresents recruiting firms but conditional on recruiting, BG overrepresents higher skilled firms.

In sum, BG looks somewhat different than representative establishments in JOLTS, even when compared to those that are hiring or trying to hire. Again, these differences could arise for many reasons. A hiring firm may not be recruiting formally or conditional on recruiting they may not post online in a way that BG can capture. Or, it could be that a given post in BG corresponds to multiple openings. For instance, in the case of high churn industries, it may be that one BG ad corresponds to multiple hires and that is why it underrepresents hiring activity. We next develop a framework conceptualizing these differences.

3 Methodology

One of the main goals of this paper is to shed light on the representativeness of online job postings data. Besides the measurement issues raised above there are many reasons why the number of BG postings might differ from the number of JOLTS openings. In what follows, we develop a conceptual framework that connects BG postings and JOLTS openings through a multiplicative "adjustment factor". We show how to estimate this relationship across time and firm-level covariates. This approach offers new insights, as variables like firm size and average wage – absent from BG – are available through our firm-level linkage with JOLTS.

However, a key advantage of the BG data is that it can be used to answer a range of research questions for which JOLTS is underpowered. Thus a second main goal is to show how job postings data alone can be used to capture the distribution of vacancies. We propose a prediction method that reweights BG postings to generate moments that are representative of JOLTS openings. In one exercise, we leverage our estimated adjustment factor based on the full set of firm-level covariates available inside the BLS. In a second exercise, we use only information available in BG (that should also be available in similar large scale online postings datasets). We conclude this section by validating both predictions.

3.1 Conceptual Framework

We assume that the data generating process takes the following form. Firm i of type j at time t decides whether to have an opening with a probability and, if yes, how many according to:

$$N_{ijt} = \begin{cases} N_{jt}^{+} & \text{with probability } p_{jt}^{N} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

For BG postings, we distinguish among establishments with zero and positive openings. If $N_{ijt} > 0$, then the firm decides whether to post the openings online with probability p_{jt}^{B} and, if yes, posts B_{jt}^{+} postings. If $N_{ijt} = 0$, then the firm may still have positive openings, which is either (1) due to the fact that the postings may not necessarily satisfy the requirements of a hire or open vacancy in JOLTS or (2) time aggregation issues. We assume that firms of type j at time t with zero openings post B_{jt}^{0} postings with probability $p_{jt}^{B,0}$. Therefore, we can summarize the BG posting decision as follows:

$$B_{ijt} = \begin{cases} B_{jt}^{+} & \text{with probability } p_{jt}^{N} p_{jt}^{B} \\ B_{jt}^{0} & \text{with probability } (1 - p_{jt}^{N}) p_{jt}^{B,0} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

Importantly, this model assumes that the number of openings (N) and postings (B) depends only on a firm's type, j. We treat the parameters of the model as fixed, i.e. there is no strategic behavior on the part of firms, and assume that there is no variation in these variables within type j.

First-order moments. Given these assumptions, the expected number of openings and postings for firm i of type j at time t are:

$$\begin{aligned} E_{jt}(N_{ijt}) &= p_{jt}^{N} N_{jt}^{+} \\ E_{jt}(B_{ijt}) &= p_{jt}^{N} p_{jt}^{B} B_{jt}^{+} + (1 - p_{jt}^{N}) p_{jt}^{B,0} B_{jt}^{0}. \end{aligned}$$

Combining the two equations and defining $n_{jt} = \frac{N^+_{jt}}{B^+_{jt}}$, we can then express the expected number of Ns as a function of expected Bs:

$$E_{jt}(N_{ijt}) = \frac{n_{jt}}{p^B_{jt}}(E_{jt}(B_{ijt}) - (1-p^N_{jt})p^{B,0}_{jt}B^0_{jt}). \tag{4}$$

Let us further simplify by defining $s^B_{jt} = \frac{p^B_{jt}p^N_{jt}B^+_{jt}}{p^N_{jt}p^B_{jt}B^+_{jt}+(1-p^N_{jt})p^{B,0}_{jt}B^0_{jt}}$, then[13]

$$E_{jt}(N_{ijt}) = \frac{n_{jt}}{p^B_{jt}}s^B_{jt}E_{jt}(B_{ijt}). \tag{5}$$

This thus defines an adjustment factor of the form

$$a_{jt} = \frac{n_{jt}}{p^B_{jt}}s^B_{jt}, \tag{6}$$

which shows that postings differ from openings due to the following three margins:

1. The *extensive margin* ($\frac{1}{p^B_{jt}}$), i.e. the inverse likelihood of having any BG postings in the same time period as a JOLTS opening.
2. The *intensive margin* (n_{jt}), i.e. openings per posting, conditional on having both; this reflects that a single posting corresponds to multiple openings or that a firm posts only a fraction of their openings online.
3. *Nuisance* (s^B_{jt}), i.e. the fraction of postings that correspond to an opening; this is a nuisance parameter that arises due to measurement issues (e.g., time aggregation, conceptual differences between JOLTS vacancies and BG postings).

In section 4, we will describe variation in the adjustment factor across time and firm-level covariates, and decompose it into the different components.

Higher-order moments. We can also derive expectations for transformations of the form $f(N) = N^r$, where r is a scalar, again leveraging our assumption that openings and postings vary only with jt:

$$\begin{aligned} E_{jt}((N_{ijt})^r) &= p^N_{jt}(N^+_{jt})^r \\ E_{jt}((B_{ijt})^r) &= p^N_{jt}p^B_{jt}(B^+_{jt})^r + (1-p^N_{jt})p^{B,0}_{jt}(B^0_{jt})^r \end{aligned}$$

Following the same steps as above, we get:

$$E_{jt}((N_{ijt})^r) = \frac{(n_{jt})^r}{p^B_{jt}}s^{B,r}_{jt}E_{jt}((B_{ijt})^r) \tag{7}$$

where $s^{B,r}_{jt} = \frac{p^N_{jt}p^B_{jt}(B^+_{jt})^r}{p^N_{jt}p^B_{jt}(B^+_{jt})^r+(1-p^N_{jt})p^{B,0}_{jt}(B^0_{jt})^r}$ and thus defines adjustment factor of the form:

[13] One can show that $s^B_{jt} = \frac{E_{jt}(B_{ijt})-(1-p^N_{jt})p^{B,0}_{jt}B^0_{jt}}{E_{jt}(B_{ijt})} = \frac{p^B_{jt}p^N_{jt}B^+_{jt}}{p^N_{jt}p^B_{jt}B^+_{jt}+(1-p^N_{jt})p^{B,0}_{jt}B^0_{jt}}$.

$$a_{jt}^r = \frac{(n_{jt})^r}{p_{jt}^B} s_{jt}^{B,r} \tag{8}$$

We will use these expectations of higher order moments further below when we predict concentration measures of openings such as the Herfindahl-Hirschman Index based on the distribution of postings in Burning Glass.

3.2 Estimation

To estimate each component of our adjustment, we need to estimate the following moments for each firm type j and time t in the JOLTS sample:

$$\begin{aligned}
p_{jt}^N &= Pr(N_{ijt} > 0 | j, t) \\
p_{jt}^B &= Pr(B_{ijt} > 0 | N_{ijt} > 0, j, t) \\
p_{jt}^{B,0} &= Pr(B_{ijt} > 0 | N_{ijt} = 0, j, t) \\
N_{jt}^+ &= E[N_{ijt} | N_{ijt} > 0, j, t] \\
B_{jt}^+ &= E[B_{ijt} | B_{ijt} > 0 \text{ and } N_{ijt} > 0, j, t] \\
B_{jt}^0 &= E[B_{ijt} | B_{ijt} > 0 \text{ and } N_{ijt} = 0, j, t]
\end{aligned} \tag{9}$$

The JOLTS data are at monthly frequency. In practice, time aggregation issues may prevent us from perfectly capturing JOLTS openings in BG. In order to allow for some degree of flexibility, we take the average number of BG postings over a 3-month rolling period centered on the month of the JOLTS measurement.[14]

We define a type j as a set of firm characteristics X and estimate a series of regressions as in equation 10, where outcomes Y_{ijt} are the six moments listed in (9) above, and i denotes a firm whose type is j.:

$$Y_{ijt} = \alpha_t + X_{jt}\beta^{y(t)} + \varepsilon_{ijt}, \tag{10}$$

Given our establishment-level match, we regress an outcome on time (month) indicators (α_t) and a vector of firm characteristics (X_{jt}) that depend on its type j, including size, age, average wage, industry dummies, and location characteristics.

We estimate these regressions at the establishment-month level using JOLTS sampling weights. We allow the coefficients β to vary with year ($y(t)$). To avoid reclassification bias, we define the time-varying characteristics at a lag with respect to the month of JOLTS measurement.[15] The advantage of this approach is that we can include just one dimension of firm heterogeneity at a time to summarize the relationships, or we can include many to better understand whether controlling for one characteristic accounts for variation in another.

[14] For example, $p_{jt}^B = Pr(\sum_{k=-1}^{+1} B_{ij(t+k)} > 0 | N_{ijt} > 0, j, t)$ and $B_{jt}^+ = E[\frac{1}{3}\sum_{k=-1}^{+1} B_{ij(t+k)} | \sum_{k=-1}^{+1} B_{ij(t+k)} > 0, N_{ijt} > 0, j, t]$. The other terms are defined analogously.

[15] Specifically, we estimate regressions separately by year and include the following covariates: calendar month fixed effects, 16 indicators for 2-digit NAICS sector, indicators for average wage quintile (total wage bill divided by average monthly employment) in the previous calendar year, indicators for urbanicity category (large central metro area, large fringe metro, medium metro, small metro, micro area, noncore), age bins: ≤ 2, 3-5, 6-10, 11-15, 15+, and indicators for lagged employment growth rate (defined as the percent change in employment over the previous month to the same month in the previous calendar year) bins: $< -30\%$, -30 to -20, -20 to -10, -10 to -5, -5 to -1, -1 to 1, 1 to 5, 5 to 10, 10 to 20, 20 to 30, $\geq 30\%$. We summarize employment size in the previous year with indicators for size < 5, 5-10, and ≥ 5000, and a linear spline with kink points at 10, 250, 1000, and 5000.

For the binary variables (p^N, p^B, $p^{B,0}$), we estimate the equation with a Probit model, whereas for the other variables (N^+, B^+, B^0), we estimate the equation with the exponential model, which respects the non-negativity constraint of these variables.[16]

3.3 Predicting Openings outside of JOLTS

In order to move outside the JOLTS sample, we propose predictors of openings that leverage the adjustment factor derived above.

Prediction #1: Full Predictor. Our first predictor is expressed in equation 11 below. We obtain fitted values from regressions of the 6 moments and use these to calculate a firm-specific adjustment factor ($\hat{a}_{ijt}$) that is a function of only characteristics X_{jt} and time:

$$\hat{N}_{ijt} \quad = \quad \hat{a}_{jt} B_{ijt}. \tag{11}$$

Note that firms vary only by their type, which can be expressed as $j(i)$, and time t, but for expositional ease, we simply refer to ijt observations. We can calculate this adjustment for any contemporary sample with postings, B, and the same sets of covariates (X). For higher order moments, we can apply an analogous adjustment based on a^r_{ijt} (equation 8).

Recall $a = \frac{n}{p^B} s^B$ so these adjustments are a function of 1) n, which adjusts for the propensity to have multiple openings per posting and 2) the inverse of p^B, which adjusts for the fact that not all establishments post openings online. Our assumption that firms differ only by their type jt means that those that do post in BG are representative of those that do not. In that sense, one can think of $\frac{1}{p^B}$ as a sampling weight that makes the BG sample representative of the sample of openings N. By modeling the posting decision itself, our prediction more effectively uses the BG data than a more general prediction of the form $N = g(B, X)$ where X are firm characteristics, since many establishments do not post online. Finally, the nuisance term s^B adjusts for the fact that a fraction of postings occur at establishments with zero openings.

Our approach assumes that p^B, n and s^B are constant within a group of establishments of the same type j in a given time period. Given this, our predictors have the property that they will equal N in expectation, when taking expectations over j and/or t within the estimation sample. The same argument applies to the higher order moments.

We can calculate $\hat{a}$ for the JOLTS sample and for the full QCEW sample, since the same covariates are available in each. Since JOLTS is only a survey, we lose power for many questions of interest and therefore find it useful to apply our adjustment to the full business registry for the United States, captured in the QCEW. Estimates are based on the 6 moments in equation 9 and in our preferred analysis, we estimate these equations on the full JOLTS sample. However, as we show next in the validation section, results are similar when we split the JOLTS sample to avoid overfitting.

Prediction #2: BG-Only Predictor. In practice, researchers may not have access to the demanding set of covariates required for the full predictor. As an alternative, we propose a predictor that uses only information that would be contained in online postings data.

[16] Our estimate is closely related to the Poisson count model, but since we sometimes allocate BG postings across multiple establishments, the monthly number of postings can take non-integer values. Since we use these variables for predictive analysis, it is crucial that the predicted values respect the non-negativity constraints, or else the predictions may become meaningless or implausible, especially when combined into the overall adjustment factor.

We estimate a second-stage regression of the form

$$\hat{a}_{jt} = \alpha_t + O_{it}\gamma^{y(t)} + \varepsilon_{ijt}, \tag{12}$$

where $\hat{a}$ are the fitted adjustment factors based on regressions of the 6 moments in the JOLTS sample. We leverage a vector of BG-covariates, O, which describe the distribution of ads posted by firm i at time t. We select covariates that generate the most meaningful variation in a, based on results summarized below. These include 2-digit NAICS industry and urbanicity fixed effects, the number of postings summarized across 5 bins, as well as the share of ads across 4 broad occupation groups, and the share of ads requiring high school or college. These ad shares are collected at the firm-month level. Because adjustment factors are strictly non-negative, we also use the exponential model for this regression.

From this regression, we obtain fitted values $\widehat{a^B}_{it}$, which give the average adjustment for firms whose postings follow a given set of characteristics O. We then use this adjustment for a BG-only prediction as in equation 13 below.

$$\widehat{N^B}_{it} \quad = \quad \widehat{a^B}_{it} B_{it} \tag{13}$$

We estimate equation 12 on the BG-QCEW matched sample, conditional on $B > 0$.[17] We will provide regression output and guidance on how researchers can use the coefficients to adjust data on job postings. This predictor will match expected N, when taking expectations over the covariates in O. We will show that it also does a reasonable job when taking expectations across covariates X.

3.4 Out of Sample Validation

Outside the prediction sample, differences in the underlying postings-openings relationship or sampling variation may generate discrepancies in the predictions relative to true openings. We employ a k-fold cross-validation to validate our model on an "out" sample within JOLTS. This approach has the advantage that we can still measure true JOLTS openings even in the out sample and helps avoid problems with over-fitting. To do so, we split the JOLTS sample into thirds. For each iteration, we estimate our model on a two-thirds "in" sample. The model consists of the probit and exponential regressions used to estimate the 6 moments outlined above. We can also predict fitted values for both in and out samples. We repeat this process 3 times, thus generating a complete set of out-of-sample predictions covering the entire JOLTS sample.

We conduct several validation exercises on the predicted $\hat{N}_{it}$s. First, along dimensions X that our model uses in its prediction, we match perfectly expected values of $\hat{N}_{it}$ and N in-sample and also very close out-of-sample. That is, the average $\hat{N}_{it}$ equals the average N conditional on a given wage bin, say. Below, we will show that the BG-only predictor does fairly well at capturing dynamics on some dimensions that cannot be measured in the BG sample, such as firm size.

To test the fit of our model with more granularity, we examine bin scatters of expected values of N on the x-axis and Bs of our predictors on the y-axis in various samples. This exercise tests in a non-parametric way whether our predictors can match expected variation in N. To calculate expected N we use the same covariates X as in our model. Using expected Ns has the advantage that we can conduct this exercise in the QCEW, completely outside the sample we used to estimate our model (JOLTS). Appendix figures A.1 and A.2 show that while the relationship between actual posts and expected N is quite muted (a slope of

[17] Relative to the JOLTS estimate, the much larger QCEW generates a handful of extreme outliers in our adjustment factor. To deal with these, we winsorize adjustments at the 95 percentile before estimating equation 12.

roughly 0.14), our predictors do a much better job. Across JOLTS in and out samples and the QCEW and for the full predictor and the BG-only predictor, we achieve slopes that are quite close to 1. That is, our adjustments do a good job capturing expected variation in N, even in this non-parametric illustration.

4 Do online job postings capture job openings?

In this section, we assess the representativeness of the BG data across time and firms, using the adjustment factors estimated in the previous section.

4.1 The aggregate adjustment factor

Figure 3 provides time series for our adjustments. For clarity, the figures restrict to a narrower time window, 2011-19, while the full time series is given in appendix figure A.3. The black solid line in Panel (a) plots annual averages for the full adjustment from equation 6. We limit to the JOLTS sample (the analytic sample that calculated a and its components) and summarize the adjustment factor for establishment-months with postings.[18] We find substantial convergence over time. The full adjustment factor averages 15.6 in 2011 and falls to 5.7 by 2019. Convergence is especially rapid until 2013, with more gradual declines from there. The appendix figure shows especially large adjustment factors in 2007 and 2010, meaning the BG postings and JOLTS openings series are quite misaligned. For this reason, we restrict attention to the more stable 2011-19 period for most of the analyses. We will however revisit the earlier years when evaluating previous researchers' use of 2007 as a benchmark pre Great Recession period (Hershbein and Kahn, 2018).

The dashed black line provides an alternative adjustment factor, one that holds firm composition constant. To do so, we compute fitted values from the regressions for a representative firm and only allow the time component to vary.[19] To place the two series on the same scale, we normalize both to the unadjusted 2011 value by multiplying all values by the 2011 ratio of the two series. Interestingly, the convergence over time holds largely within these firm characteristics. This means that the extent to which BG mis-represents the typical hiring firm on observables has not really changed.

The green series confirms that the BG-only adjustment (estimated using regression equation 12 on the full QCEW sample, but again summarized here for the JOLTS sample of establishments with postings) achieves a similar level and pattern. Though the series are somewhat mis-aligned in the earlier time period, and especially so for 2007 and 2010 (appendix figure A.3).

Panel (b) Figure 3 provides time series for the components of the full adjustment factor, which is the product of three terms: $\frac{1}{p^B}$ is the inverse of the extensive margin, n is the intensive margin, and s^B is the share of postings with non-zero N, reflecting measurement error. These are estimated from the 6 moments in equation 9.

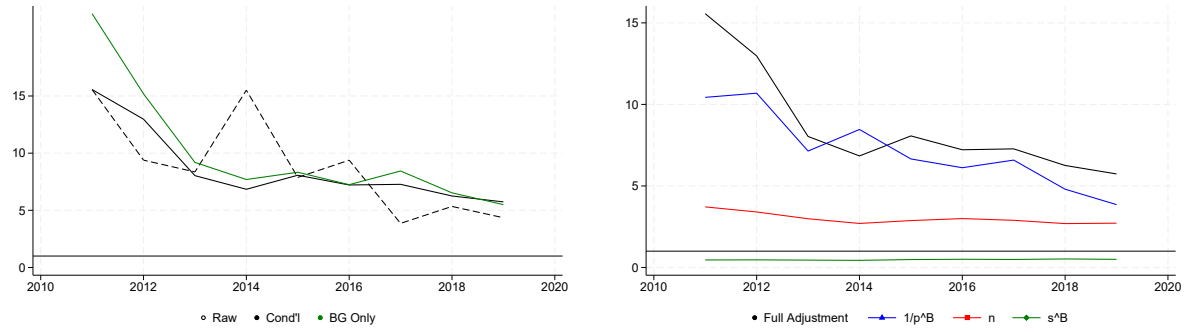
Beginning with the extensive margin, we plot $\frac{1}{p^B}$, the inverse likelihood of having BG postings conditional on openings so that declining values for all adjustment margins indicate convergence across the series. This component averaged 10.5 in 2011 and showed similar convergence to the overall adjustment factor, declining to under 4 by the end of the sample window.

The intensive margin n, which equals the number of openings per posting showed modest convergence as well, falling from about 4 at the beginning of the time period to 2.7 by the end. These estimates being well

[18] As the adjustment will only be applied to postings data, the sample conditioned on non-zero postings is the relevant one and we make this restriction for all the descriptives in this section.

[19] Specifically, the representative firm has 20 employees, is 8 years old, has an average wage bill per worker in the 3rd quintile, is located in a large fringe metro area, has 0 growth, and is in healthcare (NAICS 62).

Figure 3: Adjustment Factors over Time: 2011-2019



(a) (b)

Notes: Panel (a) shows the average full adjustment factor a in equation 6 (black, solid) and compares this to a version that holds establishment characteristics fixed at the values of a representative firm and is scaled to match 2011 (black, dashed). We compare to the BG-only adjustment factor, estimated with equation 12 (green, solid). Panel (b) shows the full adjustment factor and its components: $\frac{1}{p^B}$ is the inverse of the extensive margin, n is the intensive margin, and s^B is the share of postings with non-zero N, reflecting measurement error. The overall adjustment a is the product of the three. Estimates are from the BG-JOLTS matched sample, restricting to private sector establishments with at least 2 employees. Values are averaged annually for establishments with non-zero postings using JOLTS sampling weights. See appendix figure A.3 for the full time series: 2007 and 2010-2022.

above 1 indicates that a single posting could correspond to multiple openings or it could be that companies post only a fraction of their openings online.

Our nuisance parameter s^B is steady at about 0.5 for the whole time period. In a given month the average establishment with no JOLTS openings still has a 50% likelihood of having at least one posting. This estimate implies substantial measurement error or that a sizable fraction of postings do not fall into the JOLTS definitions of hires and vacancies. However, we will show next that the vast majority of these occurrences are in small firms. Still, it is worth keeping in mind that if one is interested in understanding how to capture openings contained in the official government JOLTS survey then one should not be concerned with the nuisance factor s^B.

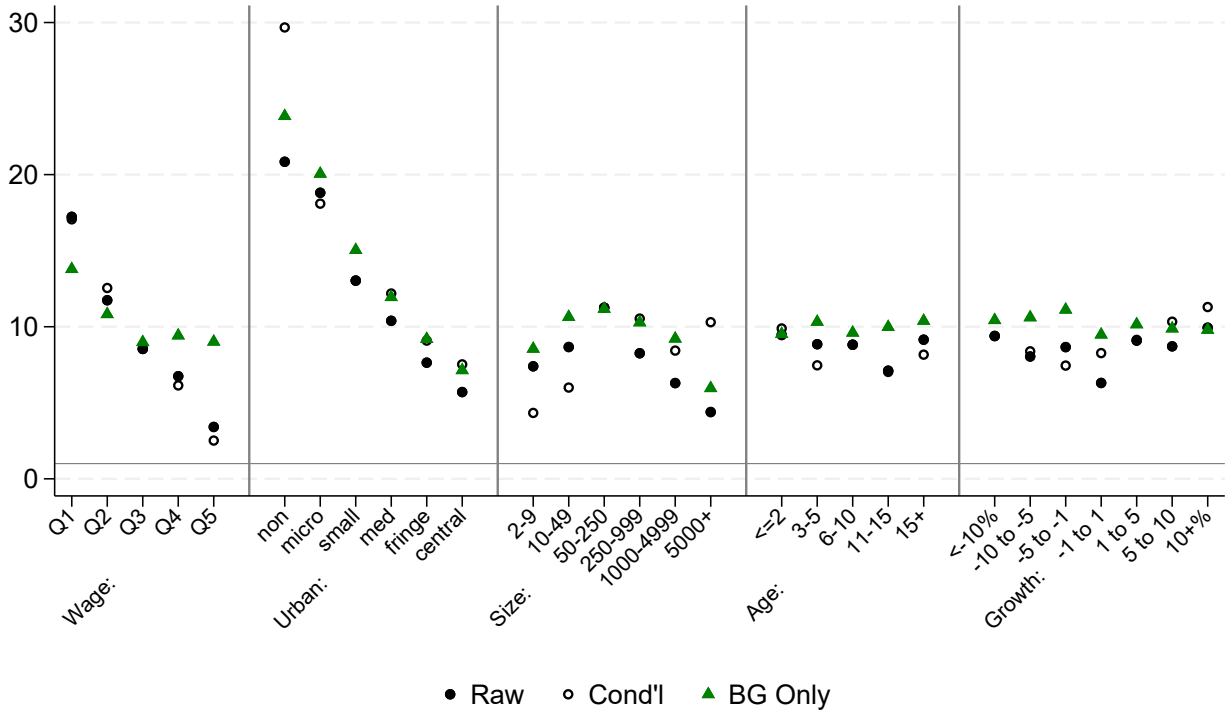
4.2 Variation across establishment characteristics

We can now, for the first time, provide an understanding of how BG representativeness varies across establishment characteristics. To do so, we estimate regression equation 10 for each of the 6 moments in the appropriately conditioned sample. We then calculated fitted values based on the firm characteristics used in the regression for all firms in JOLTS. These fitted values are combined to produce the components of interest, which we summarize here again within the JOLTS sample of firms with non-zero postings.

Figure 4 compares adjustment factors for a wide range of firm characteristics. We plot both the unconditional comparison within a category (filled in circles) and also a comparison that holds all other characteristics constant at the representative firm and only allows the given category to vary (hollow). We rescale the latter series so that it equals the raw comparison for the middle category. We also provide the BG-only version (green triangles). We average across the 2011-19 sample window.

We begin with total wage bill per worker, which we categorize into quintiles. The adjustment factor

Figure 4: Adjustment Factor by Firm Characteristics



Notes: The figure plots the adjustment factor a (equation 6) by firm characteristics. Filled in circles are sample means, hollow circles use regression analyses to hold constant all other characteristics at the representative firm level and only allow the focal characteristic to vary. For instance, the wage series plots average adjustment factors within each wage quintile (filled in circles) and averages for a firm of fixed urbanicity, size, age, growth, and sector (hollow circles). We calculate establishment-level averages within each year and then take an unweighted average across years from 2011-2019. We use the BG-JOLTS sample, conditioning on establishments with postings and to private sector establishments with at least two employees. The green triangles plot the adjustments calculated with BG-only characteristics (equation 12).

exhibits a strong gradient across categories: the lowest paying firms' adjustment is 5 times larger than the highest paying firms'. A given BG posting from a firm in the lowest wage quintile would need to be scaled up by 17 times to match the distribution of observable characteristics in JOLTS, while a firm in the highest paying quintile would only need to be scaled up by about 3.5 times.

Urbanicity categories show a similar gradient, falling 3-fold across categories: large central metro areas are the most well represented online. Those postings should be scaled by 5.7 to match the JOLTS establishment distribution, while micropolitan areas need a scaling by a factor 19. Furthermore, the hollow circles show that holding constant other firm characteristics does not account for any of the strong gradients across establishment wage bins or urbanicity categories. Finally, the BG-only adjustment does a good job capturing the urbanicity gradient since geography can be measured in BG, while it shows a more muted relationship across wage bins. We cannot measure firm-level wages in BG but the ad characteristics such as occupation and education requirements provide a proxy that still picks up some of this variation.

Appendix Figure A.4 shows the different adjustment margins for wage and urbanicity. Across urbanicity categories, p^B exhibits a strong gradient with more populated areas being much more likely to post their vacancies online, n shows a gradient in the same direction though it is quite modest, and s^B does not show

the same variability. Across wage bins, both p^B and n exhibit strong gradients in the same direction. High paying firms are both more likely to post vacancies when they have openings and also hire a smaller number of workers per posting when they have both. Low paying firms, in contrast, are less likely to post openings but when they do, they exhibit many hires per opening. So, relative to high paying firms, low paying firms need to be upweighted both because their vacancies are unlikely to appear online but also because when they do, each posting corresponds to more JOLTS openings.

Next, in Figure 4, the adjustment factor shows an interesting relationship with firm size. Without controlling for firm composition (filled in circles), the relationship is fairly flat for small-to-mid size firms and is modestly declining from mid to the largest size firms. Firms with fewer than 250 employees need an adjustment of roughly 10, while the largest firm size classes need half that. So the gradient is more modest than those for wage or urbanicity but is still declining substantially. That means that in order to make BG representative of JOLTS openings, one needs to upweight small firms and downweight large firms. Interestingly, the BG-only specification does a good job capturing this variation, since we control flexibly for the number of postings in the month and that does a good job reflecting variation in firm size.

Interestingly, once we control for firm composition (hollow circles), the relationship actually flips. For a representative firm type, the adjustment factor is actually weakly increasing with size, roughly doubling from smallest to largest size classes. This means that the negative gradient is actually being driven by other firm characteristics that happen to be correlated with size. When we probed further, we found that the wage control is doing most of the work here. Large firms tend to be high paying firms and high paying firms have small adjustment factors. In the horse race, the effect loads on wage rather than size.

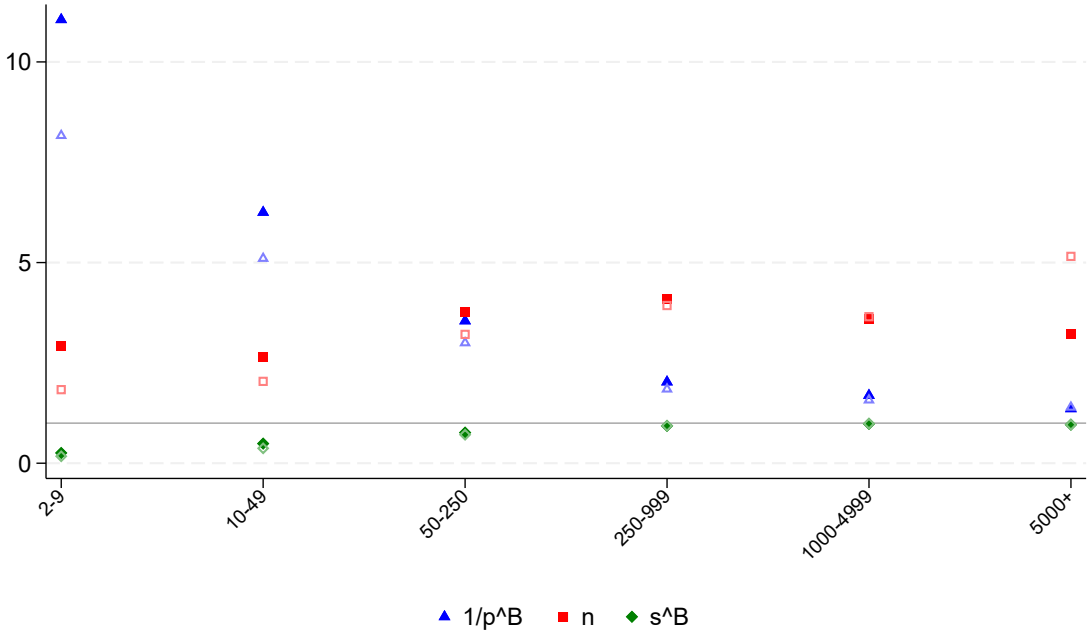
It might be surprising that the firm size gradient is less strong, especially in comparison to wage and urbanicity, and also that the modest negative gradient is actually driven by a correlation with other characteristics. The separate margins of the adjustment, plotted in figure 5 are instructive. Here we see there is actually quite a steep size gradient with respect to p^B where larger establishments are increasingly more likely to post when they have an opening. For the largest size classes, p^B is close to 1. However, s^B moves strongly in the other direction. Recall, s^B is a nuisance factor and is primarily meant to deal with measurement error perhaps due to time aggregation issues. It equals the fraction of BG postings at establishments with no JOLTS openings. The smallest firm size classes have very low s^B while the larger ones have values close to 1. Even while postings of small firms need to be scaled up because they are less likely to post online, they need to be scaled down because the posts they do make are less likely to match to an opening in JOLTS. So the two forces offset each other. Both gradients are quite strong and the magnitudes remain similar when holding constant other characteristics.[20]

Further, figure 5 shows the intensive margin (n) also exhibits modest variation along the firm size spectrum. Large firms have modestly more hires per posting than small firms, pushing up their adjustment factor because each posting actually corresponds to relatively more hires and vacancies. The effect is even stronger once controlling for firm composition. Again, the correlation between wage and size drives this increased gradient: Large firms are also high paying firms but, as we saw above, high paying firms tend to have small n's. Controlling for that fact exacerbates the gradient with respect to size which is positive. The n component increases strongly from below 2 to over 5 across the size distribution once controls are included.

Revisiting the overall adjustment in Figure 4, we can see that these forces balance out yielding a modest

[20]Errors in our BG-QCEW match could in principle cause postings and openings to misalign, lowering both p^B (i.e., raising $\frac{1}{p^B}$) and s^B. To test this, we looked at a "high confidence" sample, where postings are linked to QCEW/JOLTS firms using stricter criteria. For this sample, p^B was naturally much lower, meanwhile s^B was unchanged. As such, matching errors alone likely cannot explain the opposite movements of p^B and s^B for small firms.

Figure 5: Margins of Adjustment by Size



Notes: The figure shows the decomposition of the adjustment factor a in its three main components: blue triangles for (the inverse of) the intensive margin (p^B), red squares for the extensive margin (n) and green diamonds for the share of postings with non-zero N, reflecting measurement error (s^B). Filled in shapes represent sample means; hollow shapes report averages for a firm of fixed urbanicity, size, age, growth, and sector (hollow circles). We calculate establishment-level averages within each year and then take an unweighted average across years from 2011-2019. We use the BG-JOLTS sample, restricting to private sector establishments with at least two employees.

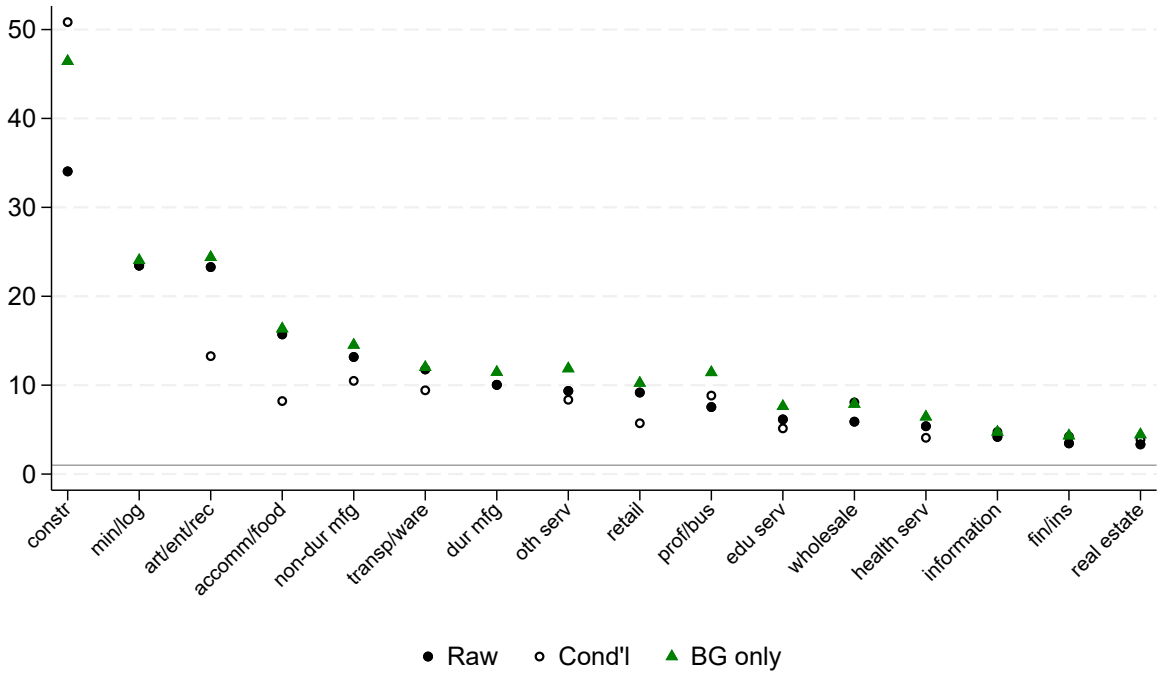
positive gradient across the smaller categories and a modest decline driven by p^B for the larger size classes that would be larger but for n. Once controlling for firm composition, the n gradient becomes steeper and outweighs the p^B effect, thus flipping the size gradient.

Even if due to a correlation with average wage, however, BG does underrepresent small firms. The largest size classes receive a smaller adjustment. This adjustment could have a big impact on moments generated with the BG data. Most firms are small and so questions related to the distribution of firms that are hiring will vary substantially in BG versus a more representative dataset. However, most workers are employed at large firms and therefore also most vacancies are posted by large firms. Since BG represents these firms quite well, it may represent the distribution of postings well.

The final panels of figure 4 show no strong pattern in overall adjustments with respect to age, whereas there is a U-shaped pattern with respect to employment growth. A differential representation of growing versus shrinking establishments could impact many inferences regarding the BG data. For instance, if expanding firms posted more of their vacancies online, this could lead researchers using online job ads data to exaggerate the increase in labor demand in economic expansions. However, given the symmetry exhibited in the adjustment factor across growth rates, these issues are likely not a concern.

Lastly, there are strong gradients across industries as illustrated in Figure 6. Construction and leisure and hospitality sectors require large adjustments to make BG representative of JOLTS. Meanwhile, professional service sectors require much smaller adjustments. Gradients are extremely steep across industries, declining

Figure 6: Adjustment Factor by Firm Sector



Notes: The figures shows the overall adjustment factor a (equation 6) for different two-digit NAICS industries. Hollow circles show the gradient by sector when holding firm composition on all other dimensions fixed. Green triangles show BG-only adjustments. Values are averaged using JOLTS sampling weights within year and equal weights across years for 2011-19. Estimates are from the BG-JOLTS matched sample, restricting to private sector establishments with at least 2 employees.

more than 5-fold across the spectrum. This gradient is exacerbated when holding firm composition across industries constant. Finally, since BG also imputes an industry, the BG-only adjustment does a good job of capturing this variation.

5 Adjusting online postings data to capture vacancies

As shown in the previous section, observable characteristics for firms that post in BG differ from the typical firm with openings in JOLTS. How much do these differences matter for drawing inferences from BG? In this section, we use our adjustment factors to reweight the full BG data series. We can do this regardless of whether the firm answered the JOLTS, since our prediction model only uses establishment characteristics that are available in the QCEW data. The resulting reweighting should match the distribution of establishment characteristics of true openings. We use these adjustments to ask how answers to various questions researchers have posed change when applying our adjustment.

5.1 Application I: The Demand for Skills

Many researchers have been interested in the skill requirements of BG postings since BG takes considerable effort to measure these and track demand for skill in the economy. However, BG represents higher skilled

Table 2: Average Skill Requirements in BG with and without our Adjustments

	Actual B	Adjusted B	
		Full	BG Only
Occupation Distribution:			
Professional Occ	0.62	0.56	0.57
Admin/Clerical Occ	0.184	0.187	0.182
Service Occ	0.082	0.097	0.106
Manual Occ	0.079	0.117	0.110
Skill Requirements:			
Education	0.67	0.63	0.62
Experience	0.58	0.54	0.54
Cognitive	0.39	0.35	0.35
Social	0.48	0.44	0.45
Computer	0.41	0.39	0.39
Salary Listing:			
No Salary Listed	0.93	0.91	0.92
Average Salary, condl	34910.20	35894.67	36578.40

Notes: Postings weighted summary statistics 2011-2019 in the BG-QCEW sample, restricting to private sector establishments with at least 2 employees. Column 1 weights each ad equally. Column 2 applies our adjustment in equation 11 and uses these adjusted postings as weights. Column 3 applies our BG-only adjustment in equation 13. Occupation ad shares aggregate listed SOC occupation codes into 4 broad categories. Skill requirements include social and cognitive indices from Deming and Kahn (2018) as well as BG created measures for education, experience, and computer requirements.

firms so we might think that their ads overrepresent firms that demand high-skilled positions. Correcting for this misalignment could impact conclusions about demand for skill in the economy and how it varies across time and space.

5.1.1 Average Skill Requirements in Postings

Table 2 summarizes ad characteristics in our full BG-QCEW matched sample from 2011-2019. Column 1 places equal weight on each ad while column 2 applies the full adjustment factor as in equation 11. Column 3 reweights postings based on our BG-only adjustment in equation 13. As noted above, BG does over-represent skilled positions. For instance, 62% of ads are in a professional occupation, 67% specify an education requirement. Applying our adjustments help match the distribution of BG firms to that of hiring firms in JOLTS.[21] Here we see downward movement in the share professional ads and in these skill requirements. Interestingly, comparing columns 1 and 2, the movements are modest – the professional ad share falls by 5.5 percentage points (9%), and education requirements decline by 4 percentage points (5.5%). Other keyword skill requirements fall by a few points. Again, BG overrepresents large firms, mostly due to a correlation between pay and size. Most firms are small but most job ads are posted by large firms. So when trying to understand the distribution of postings, our adjustment makes only a small difference.

Moreover, the BG-only adjustment in column 3 lines up very closely with our full adjustment. For instance, the share of ads posted in professional occupations is 0.62, while the adjusted share based on the full adjustment is 0.56 which nearly equals the adjusted share based on the BG-only adjustment. As

[21] In principle, we could show the same statistics for the BG-JOLTS sample and weight by the number of openings, but note that we do not observe skill demand for openings that have no corresponding online posting in the matched sample, so this calculation would not accurately represent the skill requirements.

another example, the share of ads with any education requirement is 0.67, while the two adjustments are 0.63 and 0.62, respectively. This result is reassuring that researchers who do not have access to JOLTS/QCEW covariates can still successfully reweight their data to match the JOLTS openings distribution.

5.1.2 Skill Requirements over the Business Cycle

Hershbein and Kahn (2018) show that skill requirements related to automation increased after the Great Recession. By using spatial variation in the severity of the recession, they show evidence of an episodic restructuring towards labor saving technologies. They use the 2007 datapoint as a pre-period to compare to the post recessionary years. However, as we showed above, the BG data substantially underrepresent JOLTS openings in the earlier years and especially in 2007. A compositional shift in the types of hiring firms being captured by BG could impact their inference.

HK exploit variation in the Great Recession across Metropolitan Statistical Areas (MSAs), a level at which JOLTS is not sufficiently powered. As such, we estimate their model using the BG-QCEW match. In equation 14, we regress the change from 2007 in the share of ads in MSA m and year t requiring a given skill on shock_m, which measures the size of the Great Recession shock experienced by the MSA, as well as a vector of MSA-level controls (C_m):[22]

$$Skill_{mt} - Skill_{m07} = [\text{shock}_m * I^t]\alpha_1 + I^t + C_m\beta + \varepsilon_{mt}, \tag{14}$$

where MSA-year level observations are weighted by the size of the MSA labor force in 2006 in order to upweight larger areas for greater precision.

Figure 7 compares the Hershbein-Kahn (HK) results to those that apply our adjustments. We plot the coefficients on the shock-year interactions. Outcome variables are the share of ads in the MSA-year requiring an indicated skill.[23] HK focus on education and experience requirements as well as the cognitive skill index from Deming and Kahn (2018) and computer requirements, as they argue these skills would be increasingly demanded when firms shift their workforce to work alongside automation technology.

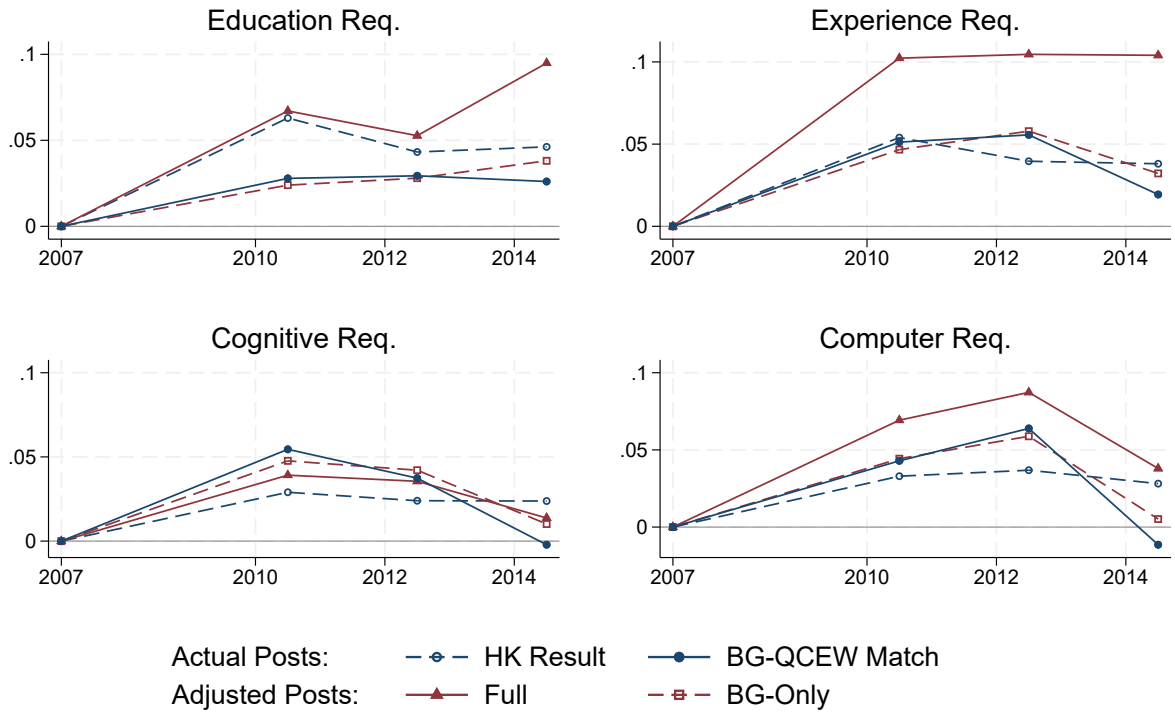
The blue dashed lines replicate HK's results showing upskilling in the immediate post-recession period. Crucial for the HK argument, this upskilling persists well after the local labor markets recovered from the Great Recession. So any purely cyclical explanations for upskilling, such as opportunistic hiring in a slack labor market, cannot explain the finding. When paired with evidence of capital deepening in harder hit MSA's, HK argue that the Great Recession accelerated routine-biased technological change. Blue solid lines show their results hold for the most part when restricting to the BG-QCEW matched sample.[24]

We find that this story of persistent upskilling in harder hit locations holds when applying either of our adjustments (maroon lines). Again, since the story in HK is about the composition of postings changing

[22] In practice, HK use a shift-share that predicts the change in employment growth from 2006-09 by applying national employment changes by industry to baseline shares at the MSA level. Controls C_m are measured in the American Community Survey (averaging 2005-06; see also Ruggles et al. (2021)) and include the share of the population that is female, black, Hispanic, Asian, married, migrated in the last year, is a high school dropout, has exactly a high school diploma, has some college, has exactly a bachelor's degree, is enrolled in school, is less than age 18, is age 19–29, is age 30–39, is age 40–49, and is age 50–64. We also control for the employment-to-population ratio and the average weekly wage of full-time workers. MSAs that cannot be matched to the ACS (8 percent) are still included with an indicator variable and imputes of zero for all controls.

[23] Specifically, we sum the number of posts with an indicated requirement across employer observations in the year and divide by the sum of all posts across employers. Our adjustments instead sum the share of an employer's posts with a skill requirement that year times their total number of adjusted posts across employers.

[24] The persistence of upskilling does not last through the end of the sample period for cognitive and computer requirements in our BG-QCEW sample. This is because public sector firms – which are excluded from this sample – drive a good chunk of the upskilling in the later years.



Notes: We summarize coefficients α_1 from equation 14, regressing a change in skill requirements in BG from 2007 to the indicated year on the Great Recession shock experienced at the MSA-level. The blue dashed line replicates Hershbein and Kahn (2018); the blue solid line replicates the results in the BG-QCEW match sample; the maroon solid line adjusts posts based on our full predictor (equation 11); the dashed maroon line adjusts posts based on our BG-only predictor (equation 13). Both adjusted series restrict to the BG-QCEW match.

over the cycle, and BG captures large firms and the bulk of postings fairly well, the adjustments make only a small difference and, if anything, make the results stronger.

5.2 Application II: Market Concentration

Researchers have leveraged the richness of the BG data to study labor market concentration at highly disaggregated market levels (Azar et al., 2020; Hershbein et al., 2020; Azar et al., 2019; Schubert et al., 2023). These advancements in measurement have been crucial for moving the academic and policy conversations towards a larger discussion on firms' wage setting power in labor markets. However, measures of market concentration can be quite sensitive to whether the large tail of small firms is or is not captured in the data. Here we explore the consequences for researchers' conclusions on labor market competition.

The most commonly used statistic to measure market concentration is the Herfindahl-Hirschman Index (HHI), which is the sum of squared market shares in a given market. Applied to online postings, the index describes the distribution of postings (B) across BG firms S within a market-month as follows (where the market and time subscripts have been suppressed for notational ease):

$$HH^B = \sum_{i=1}^{S} \left(\frac{B_{ij}}{\bar{B}} \right)^2, \tag{15}$$

where $\bar{B} = \sum_{i=1}^{S} B_{ij}$ is the market size expressed in terms of postings. We would instead like to know the

HHI for actual openings:

$$HH^N = \sum_{i=1}^{S} \left(\frac{N_{ij}}{\bar{N}} \right)^2 \tag{16}$$

where $\bar{N} = \sum_{i=1}^{S} N_{ij}$ is market size expressed in terms of openings. We can use our higher-order adjustment from the conceptual framework in Section 3.1 to calculate $E(N^2)$ and divide this by the size of the market based on our prediction $E(N)$. Specifically, our adjusted HHI is:

$$\hat{HH}^N = \sum_{i=1}^{S} \frac{s_j^{B,2}(n_j)^2}{p_j^B} \left(\frac{B_{ij}}{\hat{\bar{N}}} \right)^2, \tag{17}$$

where $\hat{\bar{N}} = \sum_{i=1}^{S} \frac{n_j}{p_j^B} s_j^B B_{ij}$ is the market size in terms of adjusted postings and where $a_j^{B,2} = \frac{s_j^{B,2}(n_j)^2}{p_j^B}$ is the adjustment factor for second-order moments for establishment of type j.[25] Given our assumptions above and taking the expectation of our prediction conditional on predicted market size, we can shows that $E(\hat{HH}^N|\hat{\bar{N}}) = E\big(\sum_i^S \big(\frac{N_{ij}}{\hat{\bar{N}}}\big)^2 \big|\hat{\bar{N}}\big) = E\big(HH^N|\hat{\bar{N}}\big)$, i.e. the expectation of our predicted HHI is equal to the actual HHI of openings, conditional on expected market size.[26] In our analysis below, we probe the robustness of our results when we condition on predicted market size, and find that the results are very similar when we condition on predicted market size. We also validate our adjusted HHI by comparing it to the HHI of openings in the JOLTS data.

In our empirical application, we compare the HHIs of postings (15) to the adjusted Bs (17) across local labor markets in the matched BG-QCEW sample. In the empirical literature, local labor markets are typically defined by a skill grouping (such as occupation or industry) and geography (such as Commuting Zones). Because the QCEW does not have information on occupation, only on industry, we calculate HHIs at the 4-digit NAICS industry by commuting zone (CZ) level.

Turning to our main results, figure 8 provides a bin scatter relating HHIs of actual Bs to the adjusted HHIs. We find that the adjusted HHIs are much less concentrated than the HHIs of BG postings. Using our full adjustment, the best fit line shows that a one-unit more concentrated market as measured by actual posts translates into only a 1/4 unit increase in concentration when adjusting for representativeness. Naturally, the reason is that small firms are underrepresented in BG and the HHI is particularly sensitive to the long tail of small firms that do not show up in BG.

The dashed line and hollow circles shows the BG-only adjustment. For this exercise, we calculate an a^2 BG-only adjustment in the same manner, by regressing the second moment adjustment on BG-only characteristics and obtaining fitted values. Here we reach a similar conclusion: the HHI of actual postings relates to the BG-only adjusted HHI with a slope of 0.3. This means researchers can use the weights we

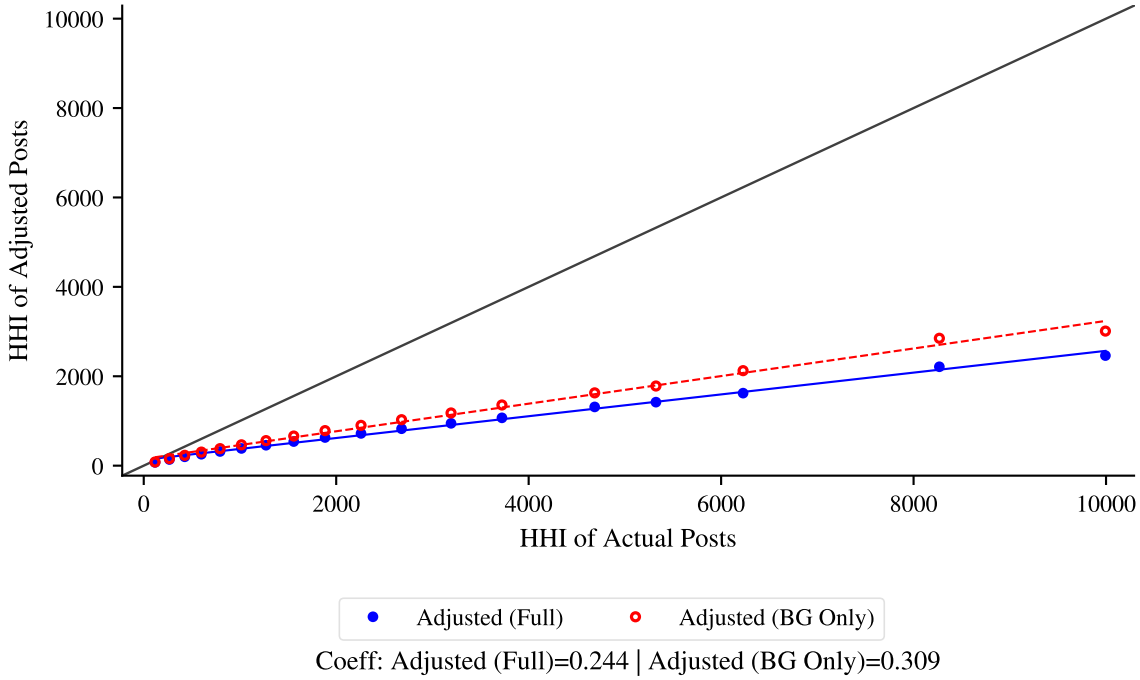
[25] In appendix figure A.5, we show that the adjustment factor of second-order moments (based on equation 8) shows similar patterns by firm characteristics as our main adjustment factor a^B.

[26] To get to this equivalence, we use equation (7) above from the conceptual framework and do the following steps:

$$\begin{aligned} E(\hat{HH}^N|\hat{\bar{N}}) &= E\Big(\sum_i^S \frac{s_j^{B,2}}{p_j^B} \frac{(n_j B_{ij})^2}{(\hat{\bar{N}})^2} \Big| \hat{\bar{N}}\Big) = \sum_i^S \frac{s_j^{B,2}}{p_j^B} \frac{(n_j)^2 E_t((B_{ij})^2|\hat{\bar{N}})}{(\hat{\bar{N}})^2} = \sum_i^S \frac{E((N_{ij})^2|\hat{\bar{N}})}{(\hat{\bar{N}})^2} = E\Big(\sum_i^S \Big(\frac{N_{ij}}{\hat{\bar{N}}}\Big)^2 \Big| \hat{\bar{N}}\Big) \\ &= \frac{1}{\hat{\bar{N}}^2} E\big(\bar{N}^2 HH^N|\hat{\bar{N}}\big) = E\big(HH^N|\hat{\bar{N}}\big). \end{aligned}$$

Note that to get this equivalence, in the last step, we also imposed that $E\big(\bar{N}^2 HH^N|\hat{\bar{N}}\big) = E\big(\bar{N}|\hat{\bar{N}}\big)^2 E\big(HH^N|\hat{\bar{N}}\big)$, which seems innocuous given that we condition on expected market size.

Figure 8: HHI's of BG Postings: Adjusted vs. Unadjusted



Notes: Bin scatter of adjusted and actual HHIs in BG postings. We divide markets (defined at the CZ-by-4 digit NAICS level) into 20 evenly sized bins based on the HHI of BG postings (equation 15) then plot the average actual HHI and adjusted HHI's (equation 17) within each bin. We average monthly HHI's over 2011-19.

Table 3: HHI Summary Statistics Summary Statistics

Panel A.	Market-Month level summary statistics		
	Actual Postings	Adjusted Postings: Full	Adjusted Postings: BG-only
Average HHI	3689.97	1028.09	1269.36
Median HHI	2475.35	582.04	770.48
Fraction above 2500 HHI	0.49	0.10	0.14
Fraction 1500-2500 HHI	0.14	0.11	0.14
Panel B.	Worker-weighted summary statistics		
	Actual Postings	Adjusted Postings: Full	Adjusted Postings: BG-only
Average HHI	733.68	244.62	304.74
Median HHI	244.89	100.21	128.92
Fraction above 2500 HHI	0.07	0.01	0.01
Fraction 1500-2500 HHI	0.05	0.02	0.02

Notes: We summarize labor market concentration measures when measured with actual BG posts (column 1, equation 15) or either of our adjustments (columns 2 and 3, equation 17). Labor markets are defined at the 4-digit NAICS-by-Commuting Zone level. Panel A places equal weight across markets while panel B weights markets by employment. We provide statistics for the BG-QCEW match restricting to private sector establishments with at least 2 employees. We calculate monthly HHI's which we then aggregate for the years 2011-19.

provide to reach more realistic values for HHIs in online postings data.

Table 3 provides summary statistics of market concentration of postings and adjusted postings; panel A

weights equally across market months while panel B weights by employment. By our market definition, the average market has an HHI of about 3700 and the average worker is in a market with an HHI of about 700. These figures are comparable to those reported in Azar et al. (2020) who also use BG data but estimate concentration among smaller markets (CZ by 6-digit occupations) and find an average concentration of 4,378. Using the full predictor, these figures are substantially smaller, each roughly a quarter of the size of the unadjusted: 1028 and 245. The BG-only predictor offers a similar conclusion.

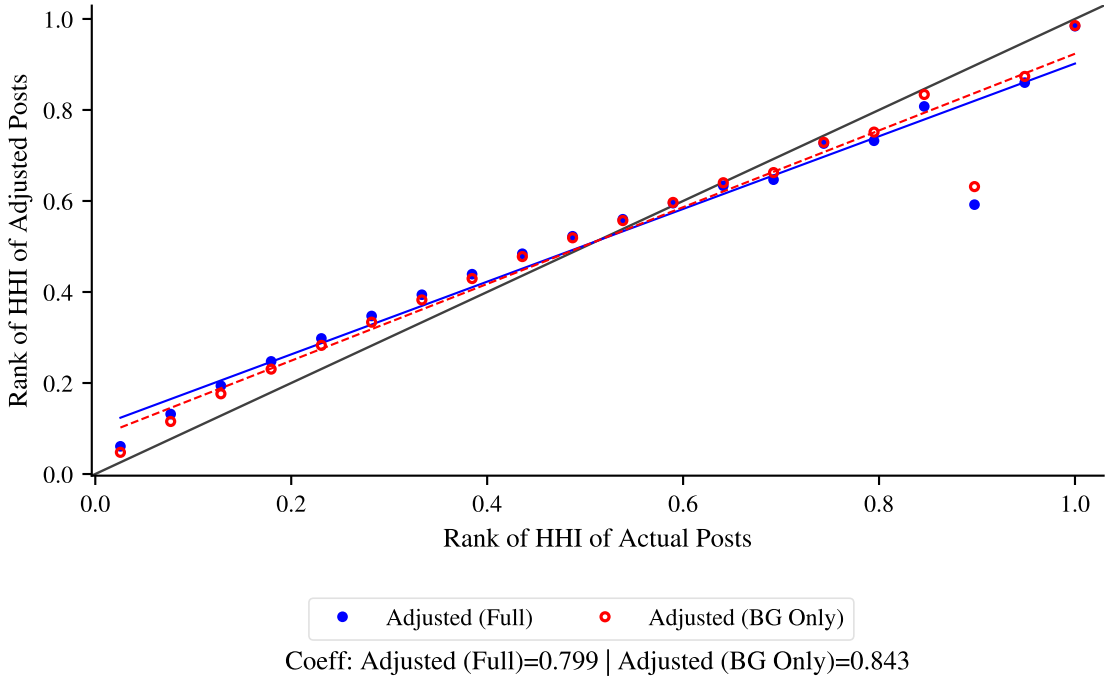
Azar et al. (2020) pay particular attention to markets with HHI above 2500, which they deem highly concentrated and markets with HHIs between 1500 and 2500, which they deem moderately concentrated.[27] When using actual postings and equal weight across markets, 50% of markets are highly concentrated and another 14% are moderately concentrated. Again, because Azar et al. (2020) use even smaller markets, they find 60% are highly concentrated. We reach very different conclusions about the prevalence of concentrated markets when using adjusted postings. When applying our full adjustment, we find that 10% of markets are highly concentrated and 11% are moderately concentrated. The BG-only adjustment produces slightly higher fractions, but overall provides a similar picture. Note also, as has been pointed out, the worker-weighted statistics reflect substantially less concentrated markets (Berger et al., 2022). This means that most workers live in labor markets that are not highly concentrated.

To probe the robustness of our results in this section, we perform several robustness checks. Note that the data underlying Figure 8 includes all markets with at least one posting in BG. However, some markets might be quite small such that they should not meaningfully be described as markets or they may generate outliers that impact our inference. In fact, when researchers use BG to calculate market concentration, they typically drop markets that do not meet a certain number of postings threshold. As such appendix figures A.6 and A.7 provide two approaches to explore robustness. First, we cut out the smallest 25% of markets based on total postings. Second, we provide an employment weighted distribution of HHIs, weighting markets by employment instead of giving each market equal weight. Though the distribution of HHIs shifts inward with either of these adjustments, the relationship between HHIs of observed Bs and adjusted Bs remains unchanged, i.e. HHIs based on adjusted Bs tend to be only a quarter of the magnitude than HHIs based on unadjusted Bs. We also show in appendix figure A.8 that our inference is largely unchanged when we drop firms with fewer than 10 employees, thus showing that we need not rely on the smallest firm size classes for this result. Finally, we probe the robustness of our results by conditioning on predicted market size. More precisely, we split our sample into the bottom quartile, the two middle quartiles and the top quartile in terms of our adjusted market size ($\hat{\hat{N}}$) and repeat the analysis. Appendix Figure A.9 shows that the slopes are all very similar across the three samples and similar to the one in the main Figure 8, suggesting that our adjustment works equally well, whether we condition or do not condition on predicted market size. Overall, this shows that our results are very robust and that BG overstates levels of labor market concentration measured in terms of HHIs by a factor 4.

While the level of HHIs in BG is about 4 times too large, we find that BG represents *ordinal* comparisons across markets quite well. To see these, we rank markets based on HHIs of their actual postings and of their adjusted postings. We plot bin scatters of these rank-rank comparisons in figure 9, across 20 evenly sized bins based on a market's percentile of actual posts HHI's; each dot gives the average percentile within each bin. Here we find evidence of a much closer relationship. Using either adjustment, the slope across market

[27] These cutoffs come from antitrust law and guidance for mergers relating to product-market concentration. The cutoffs have very different interpretations in two-sided labor markets, since a concentrated number of firms is likely to be in a labor market with a small number of job seekers as well. Regardless, previous researchers have focused on these cutpoints as important indicators.

Figure 9: HHI's: Market-level ranks of adjusted vs unadjusted



Notes: See figure 8. Here we divide markets into 20 evently sized bins based on the HHI of their actual B's and plot the average percentile within each bin of HHI's versus adjusted HHI's.

percentiles is 0.8, indicating that one can fairly accurately make relative comparisons of HHIs across markets in BG. Appendix figure A.10 shows the employment weighted distribution has an even stronger relationship (slope of 0.9 for the full adjustment and 1 for the BG-only adjustment). We therefore find that BG data can still be useful for understanding the concentration of employers in a local labor market. However, we urge researchers to focus on relative comparisons. Drawing strong conclusions from the *levels* of HHIs across markets in BG is ill advised.

6 Conclusion

Big data sources of economic activity are increasingly popular due in part to advancements in web scraping in recent decades. These data often come with a great deal of richness beyond those contained in surveys or administrative data sources. However, little is known about their representativeness. In this paper, we link large scale online postings data to a representative vacancy survey to better understand representation. The increasingly popular Burning Glass-Lightcast dataset of online postings contains rich information about the individual positions being advertised for and is thick even at highly disaggregated levels. However, it lacks even basic information about hiring firms. Our linked data show that the job openings of small and low paying establishments, as well as certain industries and those located in rural areas are not well represented online.

We derive an approach that allows reweighting BG postings so that establishment characteristics match those in the representative JOLTS data. Reassuringly, our reweighting reproduces qualitatively similar results

for many research questions. Most workers are employed at large firms and therefore also most vacancies are posted by large firms. Since BG represents these firms quite well, it also represent the distribution of vacancies well. As we showed, questions about the level and cyclicality of skill demand hold up well to our adjustment. However, most firms are small and so questions related to the distribution of firms that are hiring will vary substantially in BG versus a more representative dataset. As such, we advise against drawing conclusions from the *level* of HHIs measured in online postings data. The HHI is sensitive to whether the large tail of small firms is captured. Instead, ordinal comparisons of concentration across labor markets as measured in BG are much more representative.

Researchers rely on vacancy data to understand a wide range of questions, such as those related to job finding rates and mismatch in the economy. These researchers might like the granularity and detail afforded by large scale data on online postings. We show that researchers can reweight these data using our proposed adjustment to match the distribution of establishments in representative datasets. We recommend applying this reweighting when representativeness with respect to observable establishment characteristics is important. Otherwise, a key takeaway from our analysis is that online postings data provide a reliable representation of the vacancy distribution for questions at the posting level. In contrast, establishment-level analyses require our reweighting to yield representative results.

References

Abraham, K. G. (1987). Help-Wanted Advertising, Job Vacancies, and Unemployment. *Brookings Papers on Economic Activity 18*(1), 207–248.

Acemoglu, D., D. Autor, J. Hazell, and P. Restrepo (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics 40*(S1), S293–S340.

Azar, J., E. Huet-Vaughn, I. Marinescu, B. Taska, and T. von Wachter (2019). Minimum wage employment effects and labor market concentration. Working Paper 26101, National Bureau of Economic Research.

Azar, J., I. Marinescu, and M. Steinbaum (2022). Labor market concentration. *Journal of Human Resources 57*(S), S167–S199.

Azar, J., I. Marinescu, M. Steinbaum, and B. Taska (2020). Concentration in us labor markets: Evidence from online vacancy data. *Labour Economics 66*, 101886.

Banfi, S. and B. Villena-Roldan (2019). Do High-Wage Jobs Attract More Applicants? Directed Search Evidence from the Online Labor Market. *Journal of Labor Economics 37*(3), 715–746.

Barnichon, R. (2010). Building a composite Help-Wanted Index. *Economics Letters 109*(3), 175–178.

Barron, J. M., M. C. Berger, and D. A. Black (1999). Employer Search, Training, and Vacancy Duration. *Economic Inquiry 35*(1), 167–192.

Berger, D., K. Herkenhoff, and S. Mongey (2022, April). Labor market power. *American Economic Review 112*(4), 1147–93.

Blanchard, O. J. and P. Diamond (1989). The Beveridge Curve. *Brookings Papers on Economic Activity 20*(1), 1–76.

Braxton, J. C. and B. Taska (2023, February). Technological change and the consequences of job loss. *American Economic Review 113*(2), 279–316.

Callaci, B., M. Gibson, S. Pinto, M. Steinbaum, and M. Walsh (2024, 11). The effect of franchise no-poaching restrictions on worker earnings. *The Review of Economics and Statistics*, 1–35.

Carnevale, A. P., T. Jayasundera, and D. Repnikov (2014). Understanding online job ads data: A technical report. Technical report, Washington, DC: Georgetown University Center on Education and the Workforce.

Clemens, J., L. B. Kahn, and J. Meer (2021). Dropouts need not apply? the minimum wage and skill upgrading. *Journal of Labor Economics 39*(S1), S107–S149.

Davis, S. J., R. J. Faberman, and J. C. Haltiwanger (2013). The Establishment-Level Behavior of Vacnacies and Hiring. *The Quarterly Journal of Economics 128*(2), 581–622.

Deming, D. and L. B. Kahn (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics 36*(S1), S337–S369.

Dillon, E. W., L. B. Kahn, J. Venator, and M. Dalton (2025, July). Do workforce development programs bridge the skills gap. Working Paper 34012, National Bureau of Economic Research.

Elsby, M. W. L., R. Michaels, and D. Ratner (2015, September). The beveridge curve: A survey. *Journal of Economic Literature 53*(3), 571–630.

Forsythe, E., L. B. Kahn, F. Lange, and D. Wiczer (2020). Labor demand in the time of covid-19: Evidence from vacancy postings and ui claims. *Journal of Public Economics 189*, 104238.

Forsythe, E., L. B. Kahn, F. Lange, and D. Wiczer (2022). Where have all the workers gone? recalls, retirements, and reallocation in the covid recovery. *Labour Economics 78*, 102251.

Hansen, S., P. J. Lambert, N. Bloom, S. J. Davis, R. Sadun, and B. Taska (2023, March). Remote work across jobs, companies, and space. Working Paper 31007, National Bureau of Economic Research.

Hemelt, S. W., B. Hershbein, S. Martin, and K. M. Stange (2023). College majors and skills: Evidence from the universe of online job ads. *Labour Economics 85*, 102429.

Hershbein, B. and L. B. Kahn (2018). Do recessions accelerate routine-biased technological change? evidence from vacancy postings. *American Economic Review 108*(7), 1737–72.

Hershbein, B., C. Macaluso, and C. Yeh (2020). Monopsony in the u.s. labor market. Working paper, FRB Richmond.

Lange, F. and T. Papageorgiou (2020). Beyond cobb-douglas: Flexibly estimating matching functions with unobserved matching efficiency. Technical report, National Bureau of Economic Research.

Marinescu, I. E. (2017). The General Equilibrium Impacts of Unemployment Insurance: Evidence from a Large Online Job Board. *Journal of Public Economics 150*, 14–29.

Marinescu, I. E. and R. Wolthoff (2020). Opening the Black Box of the Matching Function: The Power of Words. *Journal of Labor Economics 38*(2), 535–568.

Mongey, S. and G. L. Violante (2019, September). Macro recruiting intensity from micro data. Working Paper 26231, National Bureau of Economic Research.

Mortensen, D. T. and C. A. Pissarides (1994, 07). Job creation and job destruction in the theory of unemployment. *The Review of Economic Studies 61*(3), 397–415.

Mueller, A. I., D. Osterwalder, and J. Zweimüller (2025). The Ins and Outs of Vacancies. Mimeo., University of Zurich.

Mueller, A. I., D. Osterwalder, J. Zweimüller, and A. Kettemann (2024, 05). Vacancy Durations and Entry Wages: Evidence from Linked Vacancy–Employer–Employee Data. *The Review of Economic Studies 91*(3), 1807–1841.

Petrongolo, B. and C. A. Pissarides (2001, June). Looking into the black box: A survey of the matching function. *Journal of Economic Literature 39*(2), 390–431.

Pissarides, C. A. (1985). Short-run equilibrium dynamics of unemployment, vacancies, and real wages. *The American Economic Review 75*(4), 676–690.

Ruggles, S., S. Flood, R. Goeken, J. Grover, and M. Sobek (2021). IPUMS USA: Version 11.0 [dataset].

Şahin, A., J. Song, G. Topa, and G. L. Violante (2014, November). Mismatch unemployment. *American Economic Review 104*(11), 3529–64.

Schubert, G., A. Stansbury, and B. Taska (2023). Employer concentration and outside options. Mimeo, Harvard University.

Shimer, R. (2005, March). The cyclical behavior of equilibrium unemployment and vacancies. *American Economic Review 95*(1), 25–49.

Shimer, R. (2007, September). Mismatch. *American Economic Review 97*(4), 1074–1101.

Appendix

A BG-QCEW Match

We use information on firm name, location, and imputed industry in the Burning Glass (BG) dataset to match to establishments in the Quarterly Census of Employment and Wages (QCEW). Our BG extract was obtained from Emsi-Burning Glass in late 2022. The company has sinced merged and changed its name to Lightcast. Our postings data include ads from January through December of 2007, and from January 2010 through October 2022.

We restrict attention the 69% of postings that include a firm name and a geographic identifier – at least as detailed as the state where the job is located. Importantly, BG does regularize firm names and this alogirthm has gotten better over time. More recent extracts may therefore have different (cleaner) firm names than ours. However, all updates to their algorithms are applied to the entire dataset so they are internally consistent.

We begin by regularizing employer names by removing common components such as "LLC" or "the," singularizing, removing geographic information, and removing punctuation. A BG cell is constructed for each unique edited employer name and city-county-state. We then find the most similar employer name in QCEW within the same commuting zone (CZ).

We determine exact matches to be those where the ad shares exact cleaned firm name with a QCEW in the same substate geography. When an exact match is not available, we use fuzzy matching techniques to find similar names across datasets, while relying on common substate geography and industry codes to identify higher quality matches. Our fuzzy match algorithm is as follows:

1. Identify 30 best matches based on Term Frequency-Inverse Document Frequency (TF-IDF) text match score (score of 0 to 1)

2. Create Levenshtein text match score across all 30 matches (normalized to be scored from 0 to .15) We add points to this baseline match score as follows:

    - +0.11 if match has same sector
    - +0.2 if match is in same city
    - +0.15 if match is in same county
    - +0.1 if match is in same state and commuting zone
    - +0.16 if match exists in years consistent with BG employer name - city

3. Among matches that score at least 0.9, we keep the highest value

These match score components were chosen based on a maximization routine on a training dataset of employer names and geography matched to QCEW that minimizes false positives.

If a posting does not have an exact firm name match nor a high enough fuzzy match score, we make one final attempt to link them to a QCEW establishment. These "sibling" matches share exact firm name but only match to an extablishment in the same state (and not sub-state geography).

It is entirely feasible for a single BG cell (firm name-geography) to be correctly matched to multiple establishments. If an employer in a particular city has multiple establishments (e.g., a large coffee chain), then the BG cell will capture the online postings for all of those establishments - with no easy way to

determine which singular establishment the postings should be assigned to. For this reason, when a BG cell has multiple establishment matches, we divide the BG postings proportionally to the relative employment of each establishment matched to the BG cell. This can mean that occasionally some establishments are assigned a fractional posting. In our analyses, we treat a positive number of postings as that establishment being assigned at least 1 posting. We could instead aggregate the QCEW to the employer-location level as well, but that approach creates problems when we move to comparisons with the JOLTS. The JOLTS may only capture one of many establishments within a given name-location cell and we would not want to attribute all BG postings to that one observation.

Table Match.1 summarizes the breakdown of match types, by percent of BG postings. Matches are grouped into 3 different categories for the firm name linkage and across levels of geography based on the most detailed level of common geography the match attains (i.e., if a match has the same city then it also has the same county, CZ, and state but will be summarized in the city column). Most of the postings are exactly matched on (cleaned) employer name and most are matched to an establishment in the same city (the smallest geography available in BG). Overall, 91% of BG postings match to at least one establishment in the QCEW.

Table Match.1: Distribution of BG Postings across QCEW Match Type

	City	County	CZ	State	Total
Exact Match	29.6	8.7	8.4	-	46.7
Fuzzy Match	13.8	5.3	6.1	-	25.3
Sibling Match	-	-	-	19.7	19.7
Total	43.4	14.0	14.5	19.7	91.6

Notes: We summarize the percent of BG postings across the type of match to establishments in the Quarterly Census of Employment and Wages. Exact matches share the same cleaned firm name and sub-state geography. Fuzzy matches are based on the algorithm described in this section. Sibling matches share exact firm names but only a geography match at the state level. 46.7% of BG postings match to an establishment in QCEW based on the exact match type. Overall 92% of postings successfully match.

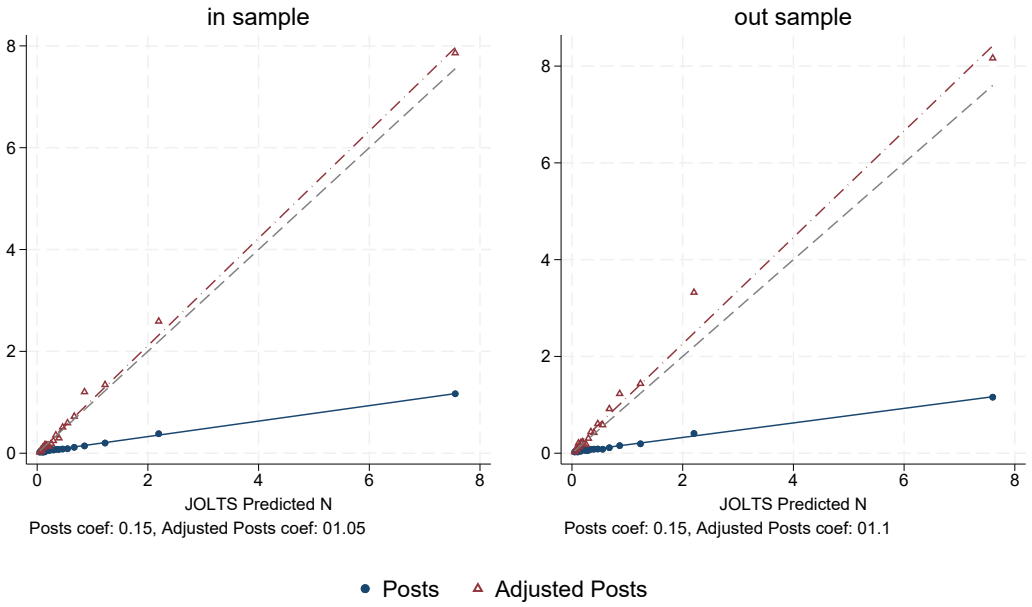
B Additional Tables and Figures

Table A.1: BG Ad Characteristics across Merge Samples

	All BG	QCEW-match		JOLTS-match
		Full	Restricted	Restricted
Education Requirement	0.563	0.566	0.564	0.571
Experience Requirement	0.493	0.496	0.495	0.476
Missing Salary	0.857	0.862	0.887	0.877
Professional Occupation	0.459	0.456	0.450	0.449
Clerical/Admin Occupation	0.245	0.246	0.240	0.251
Service Occupation	0.124	0.129	0.141	0.142
Manual Occupation	0.128	0.126	0.126	0.117
# Postings	153971100	132468167	109130392	101971186

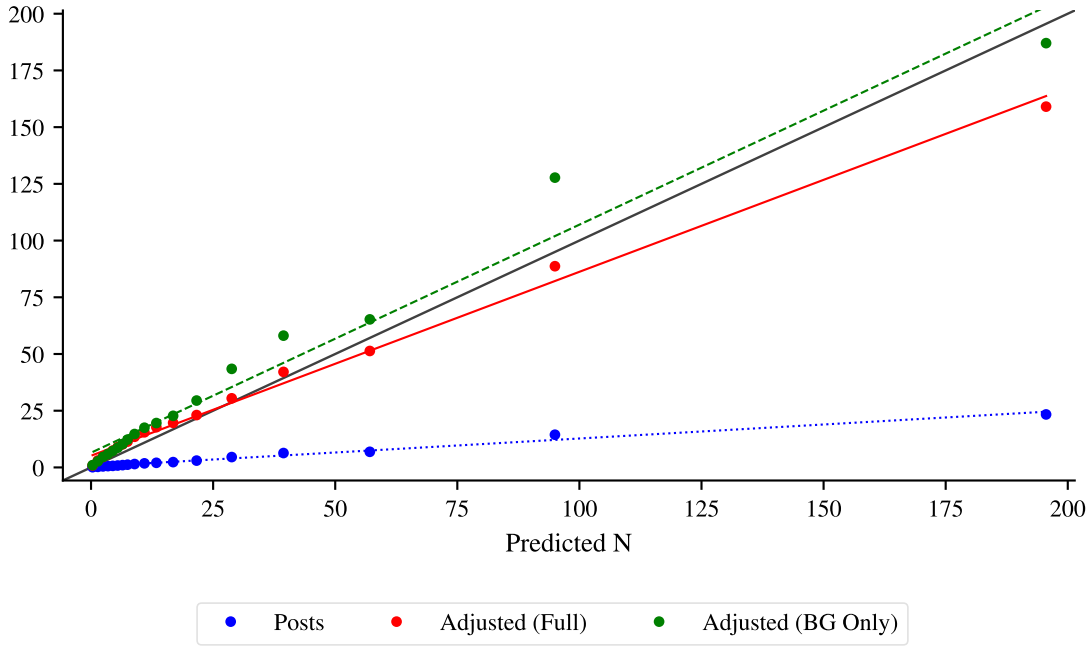
Notes: We summarize the share of ads with various characteristics across BG subsamples. All columns restrict to years 2011-2019 and BG ads with a firm name and location. Columns 3 and 4 restrict to private sector establishments with at least 2 employees. Columns 2 and 3 restrict to BG ads that can be matched to a QCEW establishment and column 4 restricts to months where the matched establishment appears in the JOLTS survey. Cells give the ad shares across select requirements or characteristics. Education and experience requirements indicate the share of ads that specify any requirement. Missing salary are the fraction of ads that contain no information on salary. Occupations are 4 aggregated categories across SOC occupation codes as follows: Professional includes SOC 11-19, 23, 27, 29; Administrative/Sales is 21, 25, 31, 41 (excluding 412), 43; Low-skill Service is 35-39, 412; and Blue Collar is the remainder (33, 45-53).

Figure A.1: Validating our Model in JOLTS



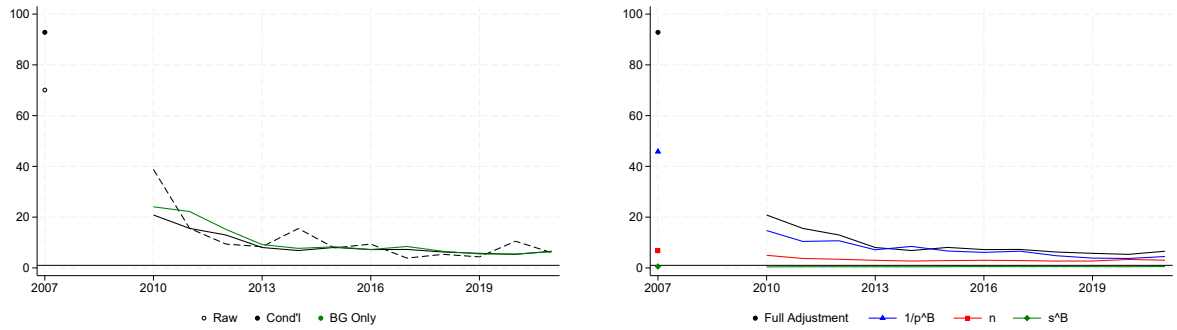
Notes: This figure presents bin scatter plots, dividing JOLTS establishments into 20 evenly sized bins based on predicted N. (We predict N based on the full set of covariates in our model.) We then take average values of the number of postings and the number of adjusted postings using our full adjustment. Both samples are drawn from JOLTS. We split the JOLTS sample into thirds and estimate our models on each 2/3 sample. We predict fitted values based on this estimates for the full JOLTS sample where the remaining 1/3 contributes to the "out sample". We repeat 3 times and aggregate to make up the full out sample.

Figure A.2: Validating our BG-only Model



Notes: See figure A.1. This figure uses the full BG-QCEW matched sample. We predict N based on the full set of covariates in our model and divide establishments into 20 bins based on this predicted N. Each bins do not reflect equal weight but instead are chosen to summarize a wide range of data. Within each bin we take the average N, the average number of postings (blue), the average adjusted postings based on our full model (maroon), and the average adjusted postings based on our BG-only model (hollow triangles). The latter uses only characteristics observed in BG to predict the adjustment factor for each establishment.

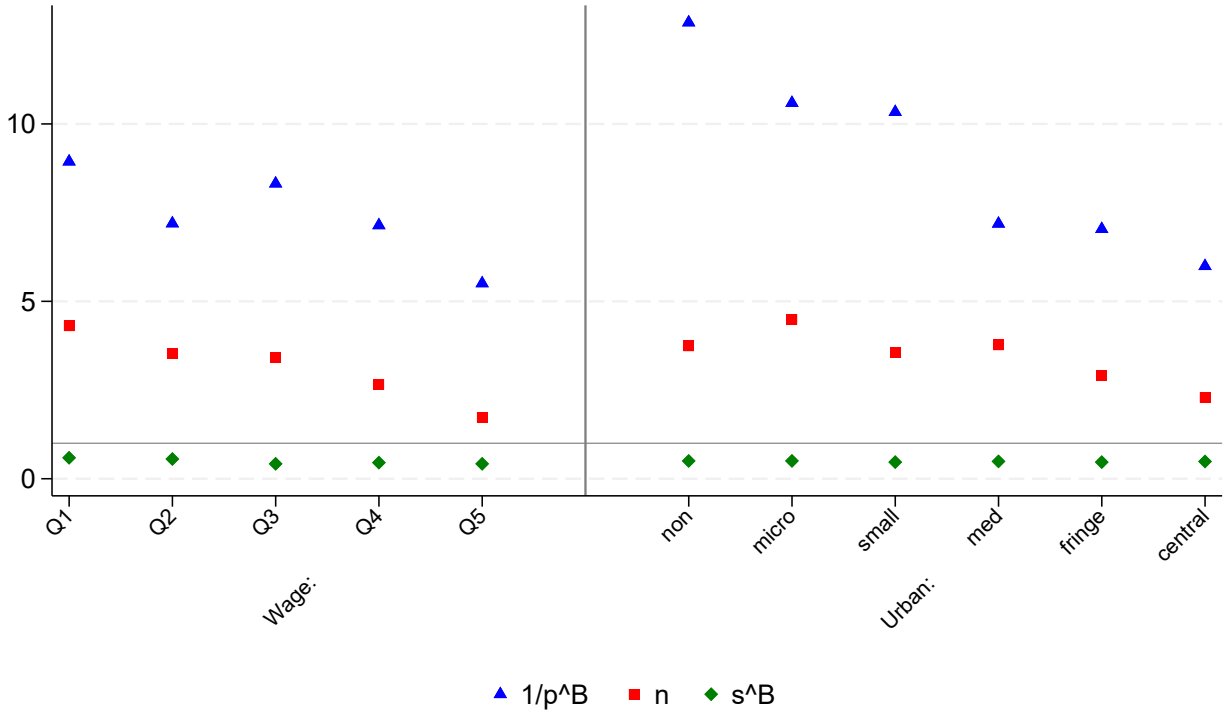
Figure A.3: Adjustment Factors over Time, all years



(a) (b)

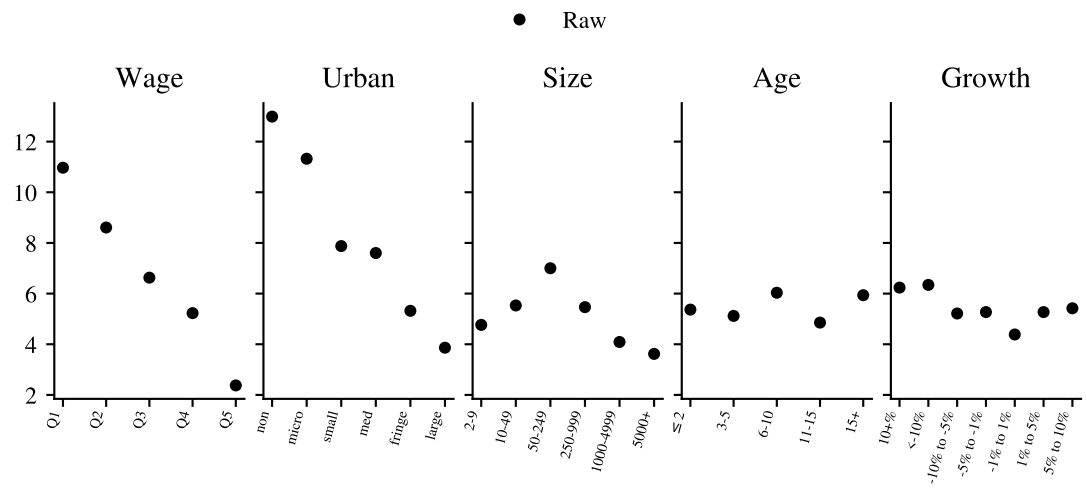
Notes: This figure provides the full time series companion to figure 3, which zooms in on our preferred sample window for clarity.

Figure A.4: Margins of Adjustment by Wage and Urbanicity



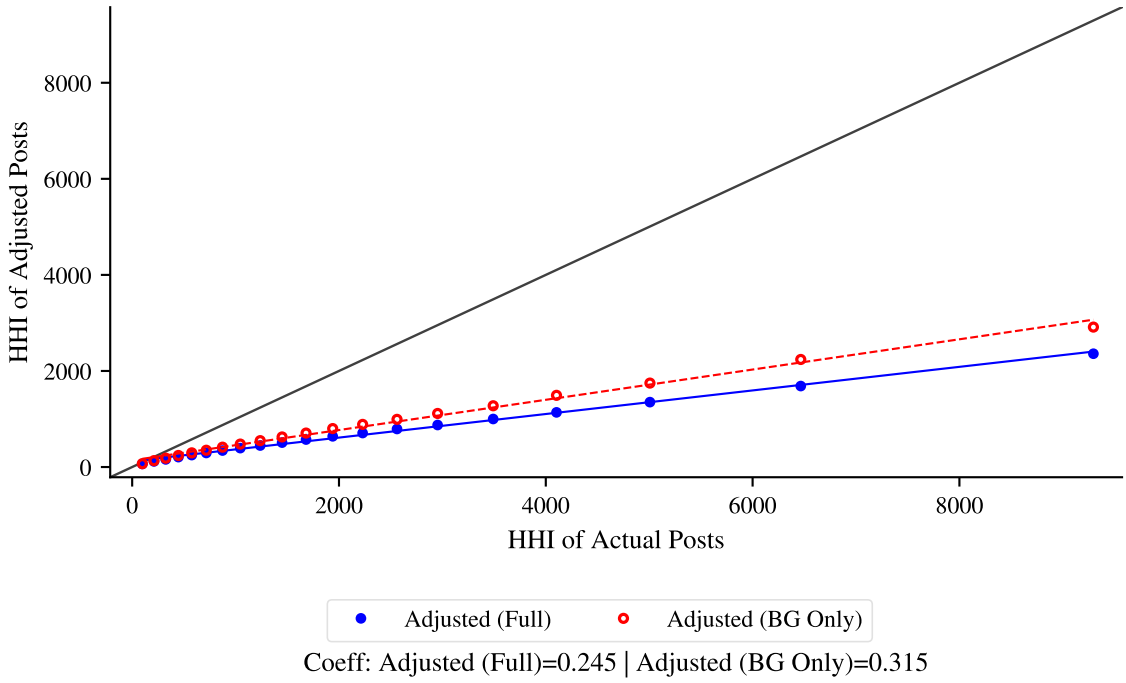
Notes: See figure 4. The figure shows the decomposition of the adjustment factor a in its three main components: blue triangles for (the inverse of) the extensive margin (p^B), red squares for the intensive margin (n) and green diamonds for the share of postings with non-zero N, reflecting measurement error (s^B). We calculate establishment-level averages within each year and then take an unweighted average across years from 2011-2019. We use the BG-JOLTS sample, restricting to private sector establishments with at least two employees.

Figure A.5: Adjustment Factor for $E[N^2]$



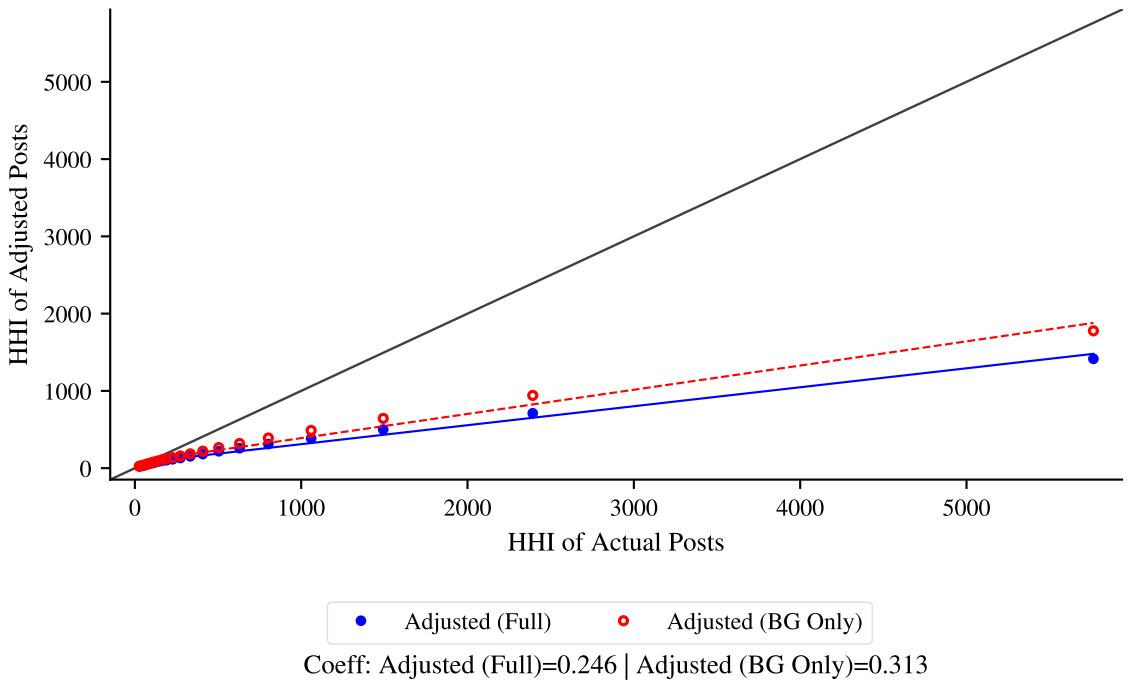
Notes: The figure plots the adjustment factor for a^2 by firm characteristics. See equation 8. Results are shown for the baseline sample with data from 2011 to 2019.

Figure A.6: HHI: Excluding Small Markets



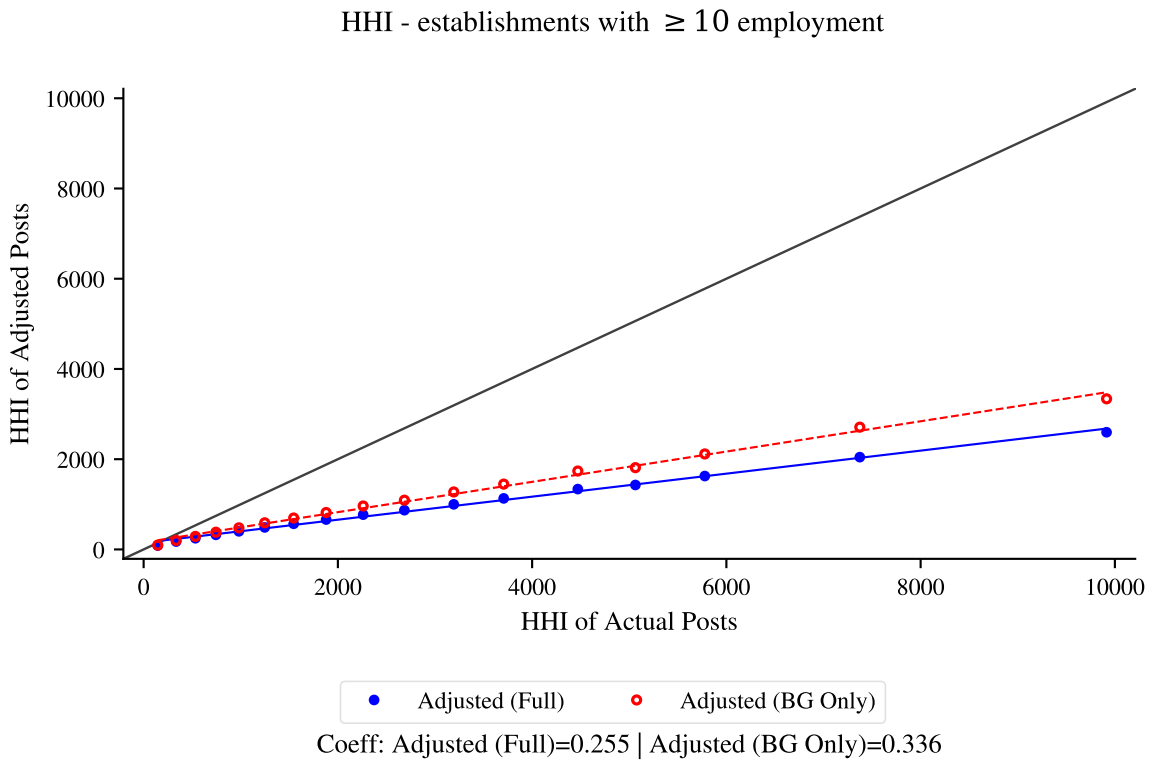
Notes: See figure 8. Here we exclude the bottom 25% of markets based on their number of posts over our entire sample period.

Figure A.7: HHI: Employment Weighted



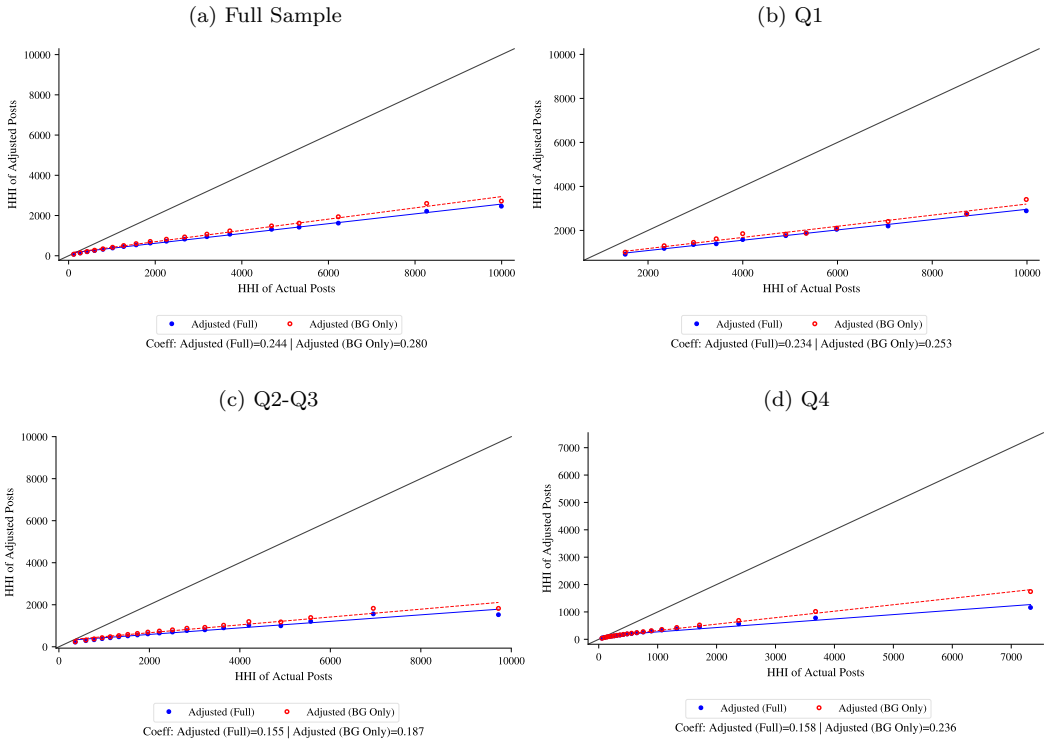
Notes: See figure 8. Here we aggregate markets into 20 evenly sized bins when weighted by market size and then plot size-weighted HHI's within each bin.

Figure A.8: HHI: Excluding Small Firms



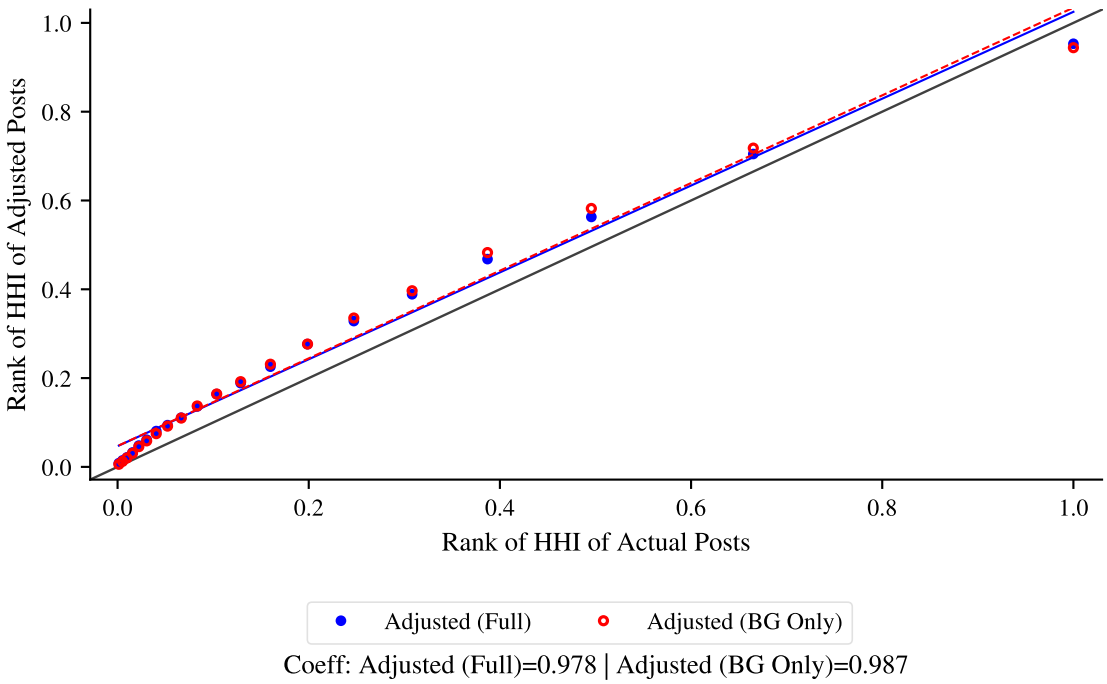
Notes: See figure 8. Here we restrict the sample to firms with at least 10 employees.

Figure A.9: HHI's of BG Postings: By Predicted Market Size



Notes: Bin scatter of adjusted and actual HHIs in BG postings. We divide markets (defined at the CZ-by-4 digit NAICS level) into 20 evenly sized bins based on the HHI of BG postings (equation 15) then plot the average actual HHI and adjusted HHI's (equation 17) within each bin. We average monthly HHI's over 2011-19. Panel a shows the results for the full sample, whereas Panels b, c and d show them by quartiles of adjusted predicted market size ($\hat{\tilde{N}}$).

Figure A.10: HHI: Rank on Rank, Employment Weighted



Notes: See figure 8. Here we weight markets based on employment, rather than giving each market equal weight, when dividing into 20 bins based on HHI of actual postings. We also calculate employment-weighted means within each bin.