# Gender Differences in Returns to Skills

*Evidence from Job Vacancy Data and Matched Employer-Employee Data*

Mathias Fjællegaard Jensen *
*Copenhagen Business School*

January 2020
*Preliminary version - please do not distribute*

**Abstract**

Recently available data from online job vacancies have enabled analyses that move beyond across-occupation variation to also include within-occupation variation in terms of skills: in which occupations, but also, in which firms do workers employ certain task-specific skills, and how does this skill composition develop with technological change? Such questions can be answered by the use of job vacancy data alone. More interestingly, I also test how the employment of skills and their returns depend on the gender of the worker by exploiting a novel combination of Danish job vacancy data and matched employer-employee register data.
I use the combination of vacancy and matched employer-employee data to show that women face lower returns to cognitive, customer service, financial, and computer skills when compared to men after controlling for both occupation and firm fixed effects. Thus, ignoring the gendered dimension of returns to skills would lead to biased results and conclusions.

*JEL classification*: I24, J16, J24, J31, J71, O33

*Keywords*: returns to skills, tasks, technological change, wage differentials, gender pay gap.

*Department of Economics, mfj.eco@cbs.dk, +4538155620. Thanks to Fane Groes and Moira Daly for data access, advice, and support. Thanks to Oliver-Alexander Press for his assistance with the job vacancy data.

# 1. Introduction

Since the seminal work of Autor, Levy, and Murnane (2003), the effects of technological change have primarily been studied in the context of disaggregate task-specific skills and the routine-biased technical change (RBTC) hypothesis. Autor et al. (2003) argue that technological change primarily affects the labour market by substituting workers undertaking routine cognitive and routine manual tasks, and by complementing workers undertaking non-routine tasks. The RBTC hypothesis can explain observed patterns of job polarisation in the US (Acemoglu and Autor, 2011) as well as in many European countries, including Denmark (Goos, Manning, and Salomons, 2009, 2014). The link between task-specific skills, technological change, and job polarisation have spurred research into the demand for certain skills and research on their returns. Certain task-specific skills have been highlighted as complementing new technologies, and thus, the demand and returns to these skills should increase with technological change. Recently, Deming (2017) and Weinberger (2014) have emphasised the growing employment and wages in jobs requiring both social and cognitive skills, rather than cognitive skills alone.

Black and Spitz-Oener (2010), and Cerina, Moro, and Rendall (2017) find that the polarization patterns noted by Autor et al. (2003) and predicted by the RBTC hypothesis are more pronounced for women than for men. Bacolod and Blum (2010) use US occupation-level skills data to show that the returns to people skills and cognitive skills increased from 1968-1990. Since women are particularly well-endowed with people skills and cognitive skills, they find that increasing returns to these two skills can explain up to 20 % of the decline in the gender pay gap. Beaudry and Lewis (2014) find that gender pay gap narrowed with the adoption of PCs from 1980 to 2000, because women are well-endowed with cognitive skills, which complement PC adoption. Studies on the *interaction between technological change and gender* typically use skills and task data at the occupation-level, i.e. they do not observe within-occupation variation in skills. Hence, gender differences in skills can only be inferred from the fact women and men tend to work in different occupations. Lindley (2012), however, utilises survey data on individual-level variation in skills and shows that women tend to be less endowed with skills that complement tasks related to technological change, and thus, concludes that overall women lost out on technological change between 1997 and 2006.

The empirical analysis that follows in the sections below draw on two additional, well-known conclusions from the literature on gender differences in labour market outcomes. Firstly, women still receive substantially lower hourly wages when compared to men, al-

though some convergence in women's and men's labour market outcomes has been observed internationally over the last few decades, both in terms of hours worked, earnings and educational attainment (Blau and Kahn, 2017; Goldin, 2014; Lindley and Machin, 2012; Olivetti and Petrongolo, 2016). Even *within* occupations, large and significant gender pay gaps remain (Blau and Kahn, 2017; Goldin, 2014; Lindley, 2016). Secondly, women and men remain, to a large extent, segregated in the labour market as they tend to work in different occupations and industries (Levanon and Grusky, 2016; Olivetti and Petrongolo, 2014). Counterintuitively, occupational segregation is particularly pronounced in Scandinavian countries, including Denmark, despite the fact that we also observe some of the smallest gender pay gaps in these countries (Jarman, Blackburn, and Racko, 2012).

The analysis in the paper at hand relies on the match of data from two sources: 1) register data on the full population on Danish employers and employess, and 2) Danish online job vacancies. With the internet's omnipresence in the Global North, online posting of job vacancies is now an integrated part of firms' recruitment of new employees. The text of each job post is highly informative when studying modern labour markets: Typically, job posts state expected skills, education, and experience of potential applicants, as well as certain characteristics of the job itself, e.g. its occupation, industry, and region. Crucially, the text of digital job posts can easily be scraped from various sources on the web. Before the availability of job vacancy data, researchers typically relied on skills and tasks data from relatively small surveys or from the DOT- and O*NET-databases, which were infrequently updated and provided job characteristics that only varied at the occupational level. Most of the studies that utilise job vacancy data also explore technological change in the labour market by exploiting the fact that job vacancies typically include information on skills requirements. However, these studies do not tend to point at gender differences in outcomes. Hershbein and Kahn (2018) show that during the Great Recession, skill requirements in job posts increase more in areas that were hit harder by the recession. Modestino, Shoag, and Ballance (2016a,b) find a similar relationship between skill requirements and the availability of workers, i.e. that skill requirements increased during the recession and decreased again through the recovery. Cortes, Jaimovich, and Siu (2018) utilise new measures of tasks from job ads in a range of US newspapers from 1960 to 2000 together with DOT data. They find that when social skills become more important within an occupation, the occupation's female share of employment also increases. After merging their skills measure to a sample of US census data, they also indicate that returns to social skills have increased over time, which is consistent with Deming's (2017) findings.

Of the papers utilising job vacancy data, Deming and Kahn's (2018) is the one closest

related to my analysis. They use Burning Glass Technologies' online job vacancy data from 2010 to 2015 to extract 10 general skill measures at the firm*occupation-level. Next, they match these skill measures to data on individual firms and to wage data from metropolitan statistical areas (MSA). Thus, they can estimate the relationships between skills and wages, as well as between skills and firm performance. Deming and Kahn (2018) find that their skills measures generally correlate positively with both wages and firm performance. High paying and high performing firms require higher levels of social and cognitive skills. When a job requires both social and cognitive skills, they find a particularly high level of wages. Although Deming and Kahn (2018) explore variation in returns to skills both across and within occupations, they cannot say whether or not their results hold at the individual level. This follows from the fact that they cannot match their skills and firm data with employees, but only with wage data at the MSA-level.

Although job vacancy data enable analyses of labour market outcomes, which would be impossible to undertake with traditional data sources, e.g. of within-occupation variation in skills, all job vacancy datasets are constrained by the fact that information on the hired worker is hidden, including information on the worker's gender and earnings. Vacancy data is often matched with firm-level data, for example by using firm names in Deming and Kahn (2018), but matching at the individual-level is impossible in settings where only datasets with samples of workers are available. However, I have access to a recently available dataset that contains all Danish online job posts covering the period 2010-2016. The Danish job vacancy data distinguish themselves by the fact that they can be matched with Danish register data at the firm*occupation*month-level. This exercise can only be undertaken because the Danish register data include monthly information on employment, including earnings, occupation codes, and firm identifiers for the universe of Danish employees. The resulting pseudo-individual match between vancancy data and register data makes it possible to evaluate gender differences in returns to skills both *across* and *within* occupations. In other words, the matched vacancy and register data makes it possible to answer the question: do women and men face different returns to the same task-specific skills, e.g. social skills, cognitive skills, and computer skills? If so, existing studies that ignore the gender dimension in returns to skills may include (gender-)biased estimates of returns to task-specific skills.

I find that task-specific skills do not yield particularly high returns to men beyond what can be explained by occupation and firm fixed effects with the exception of cognitive and financial skills. However, returns to 5 out of 9 task-specific skills are significantly lower for women when compared to men after controlling for occupation and firm fixed effects. Importantly, the gender differences in returns to task-specific skills are pronounced for cognitive

and computer skills; skills that have been emphasised as being technology-complementing in the existing literature. It follows that gender differences in returns to skills are particularly important to consider in the context of technological change. I do not, however, find any significant effects of the interaction between social and cognitive skills, neither for women nor for men.

The paper is outlined as follows. In Section 2, I describe the Danish vacancy data and register data. Furthermore, this section includes some details on the data pre-processing, although some is reserved for Appendix A. Section 3 includes some descriptive analyses of the data. In section 4, I present regression models and results, as well as some robustness checks. Section 5 contains a discussion of the results, and Section 6 concludes.

# 2. Data sources and pre-processing

The analysis that follows below rely on two sources of data. Firstly, Statistics Denmark provide register data on employment, education, demographics, firm characteristics etc. Crucially, these registers include the entire population of both workers and firms in Denmark. Furthermore, it is possible to match the different registers at the firm- and individual-levels. Monthly employment data are available, and they include a firm identifier and an occupational code for each employment relation. Secondly, Danish online job vacancy data from 2010-2016 are supplied by the Danish consultancy firm, Højbjerre Brauer Schultz (HBS). These data also include a firm identifier and an occupational code for each job post as well as a posting date. Thus, it is possible to match data from the two sources using firm identifiers and occupational codes, and by exploiting the data's time dimension. In the following subsections, I separately describe the register and job vacancy data in more detail before I move on to describe the match between the data sources.

## *2.1. Register data*

In this subsection, I briefly outline which data I extract from Statistics Denmark's registers. [1] The BFL register provide detailed monthly employment data for the entire Danish population. Monthly wages, job start and end dates, monthly hours, a firm identifier (CVR-number), and an occupational code are provided for each monthly observation. A person will appear in the register multiple times if they have more than one job in a given month, i.e. jobs are not aggregated at the individual-level, but are included as separate observations. In

[1] For data documentation, see:
http://www.dst.dk/extranet/forskningvariabellister/Oversigt%20over%20registre.html

what follows, I define a job spell as the period over which a worker remains within the same firm*occupation cell. I use *3-digit* occupation codes (DISCO-codes). Thus, a new job spell starts when a worker enters a new role in the same firm (new occupation code), or when a worker gets a job in another firm (new firm identifier). From this definition, I construct my main dataset as follows. I identify new jobs in BFL, i.e. jobs where workers are registered with either a new occupational code or a new firm identifier in a given month.[2] I construct a sample of those new jobs with the first 12 months of observations in BFL (or fewer, if the job spell ends before). Next, I aggregate to get the 12-months-averages of hourly wages, full-time equivalents [3], and other relevant variables. Thus, this dataset contains all new jobs and information on the first 12 months of employment. I have access to data until the end 2016, and since I need 12 months of observations, the latest job spells included start in January 2016. The constructed dataset yield information on workers only during their first year of employment in a certain job. I impose a number of restrictions on the sample - see Appendix A.1. To complement the employment data, I extract data on demographics, years of education, student status, employment experience etc. from other registers, which completes my register-based dataset. In the following subsection briefly outline the Danish job vacancy data before moving on the match between the vacancy and register data.

## *2.2. Job vacancy data*

HBS collects online job vacancy data from numerous Danish online jobs boards, and thus, they believe that their data contains the near universe of publicly accessible Danish online job posts.[4] They remove duplicates and clean the data before machine reading the job posts. HBS extracts the date on which a given job vacancy was posted online, the identification number (CVR-number) of the posting firm, and a 6-digit DISCO-code. If the firm identifier is not listed directly in the job post, HBS imputes it from publicly accessible registers using the firm name listed in the job post. Importantly, HBS also extract keywords from the raw text in the job post. In many ways, the resulting data is similar to the US job vacancy data supplied by Burning Glass Technologies. In order to be able to match with the register

[2]"New" in the sense that the worker was not observed in same firm*occupation cell in the month before. Furthermore, I detect gaps between spells of work in the same firm*occupation cell. If the gap between two spells is less than 6 months, I do not code reoccurring work in a firm*occupation cell as a new job, rather I code them as the same job. I also correct for changing firms identifiers.

[3]A full-time job is defined as 1923.96 hours per year by Statistics Denmark. Hence, full-time equivalents = total number of hours per year / 1923.96 (see https://www.dst.dk/da/Statistik/dokumentation/Times/moduldata-for-arbejdsmarked/fuldtid). This measure of full-time equivalents will be used as weights in the analyses that follows.

[4]For more details, see:
http://www.hbseconomics.dk/wp-content/uploads/2017/09/Eftersp%C3%B8rgslen-efter-sproglige-kompetencer.pdf

Table 1: Skills categories and examples of their corresponding keywords

| Skill | Examples of keywords |
|---|---|
| Cognitive | problem solving, research, analytical, critical thinking, math, statistics, systematic |
| Social | communication, teamwork, collaboration, negotiation, presentation, social, extrovert, network, relations |
| Character | organised, detail-orientated, multi-tasking, time management, meeting deadlines, energetic, busy, engaged, overview |
| Writing/ language | writing, language, English, German, Swedish, Norwegian |
| Customer Service | customer, sales, client, patient |
| Management | management, supervisory, leadership, mentoring, staff, control, planning, implementing |
| Financial | budgeting, accounting, finance, cost, tender/bids |
| Computer (general) | computer, spreadsheets, common software, (e.g. Microsoft Excel, PowerPoint) |
| Computer (specific) | programming, java, python, computer science |

*Note: Categories their keywords are based on Deming and Kahn (2018), Table 1.*

datasets, the vacancy data sample is restricted to include job posts with non-missing firm identifiers and occupational codes only.

In order to extract skill requirements from the job vacancy data, I initially follow the method of Deming and Kahn (2018). They map a selection of keywords into skills categories. For example, the keyword "teamwork" is indicative of a job requiring social skills. The nine skill categories, which I use, as well as the categories' mapping to a selection of keywords can be found in Table 1. Unlike Deming and Kahn (2018) who only map a selection of keywords into skill categories, I assign all keywords either a skill category or a noise tag. This is done as follows: 1) The most frequent keywords (approx. 2000) are assigned a skill category or noise tag manually. These words amount to the vast majority of keyword-observations. 2) Using online dictionary APIs each word's synonyms are obtained.[5] Each word's synonyms are assigned the same category as the word itself. 3) Using online dictionary APIs the each word's definitions are obtained. 4) Using the definition of the words, the remaining non-categorised words are assigned a category using machine-learning methods. See Appendix A.2 for more details. After these steps, all keywords are assigned either a skill category or a noise tag. The categorised keywords undergo further pre-processing, but only after the vacancy and register data are matched. The matching procedure is described in the next section.

[5]Many thanks to the Society for Danish Language and Literature for providing access to these ressources.

### *2.3. Data match*

As unique firm identifiers and occupational codes (DISCO-codes) are included in both the register data and job vacancy data, the data can be match along those two dimensions. Furthermore, I exploit the time dimension of the data. For the match on DISCO-codes to be reliable, the codes must be consistently coded across the register data and job vacancy data. Thus, in Appendix A.3, I briefly outline how DISCO-codes are coded in the two data sources. Although DISCO-codes are generally imputed in a similar manner in both the register data and the job vacancy data, some inconsistencies are to be expected at the very detailed *6-digit* level. For example, there are three subdivisions of school teachers at the *6-digit* level and only one at the *4-digit* and *3-digit* levels.[6] In order to avoid coding inconsistencies, I perform the following matching procedure at the *3-digit* DISCO-codes (170 unique values).

First, I assume that vacancies are posted in same month as the vacancy is filled or maximum four months prior. For example, if a job spell starts in May, the corresponding vacancy would be posted anytime between the beginning of January to the end of May in the same year. With this assumption, I use the job vacancy data to construct a rolling sum of job vacancies for each firm*occupation cell. If a new job spell appears in the BFL data, I match it with job vacancies summed over the relevant 5 months. For example, if a firm posts two job vacancies in the same firm*occupation cell, one in January and one in February, a job spell starting in January will only be matched with first vacancy, whereas a job spell starting in February will be matched with both vacancies. Because 4 months of job vacancy data before job start is needed, my matched data is limited to jobs commencing in the period May 2010 to January 2016. This matching strategy gives a pseudo-individual-level match between new employees and their corresponding job post. Table 2 shows match rates aggregated to the yearly level for the dataset using *3-digit* DISCO-codes.

My definition of a new job is not very restrictive, so it is not surprising that only 30.8 % of new BFL-jobs can be matched with a job post. Many of the new jobs are likely to be internal hires, informal hires (the job is not publicly posted), or DISCO-recodes (e.g. because of promotions). However, 51 % of job posts are matched to new BFL-jobs. This is a very high match rate when compared to, for example, Kettemann, Mueller, and Zweimüller (2018) who undertakes a similar exercise using Austrian data.

It is assumed that new employees' skills levels are reflected in the job posts in their firm*occupation cell just around the start of their job spell. Furthermore, focussing on the first 12 months of wages in a job spell should limit bias from additional skills learned in the firm*occupation cell. Since only few workers tend to start in the same firm*occupation cell

[6] For more details, see:
https://www.dst.dk/da/Statistik/dokumentation/nomenklaturer/disco-08

Table 2: Match rates

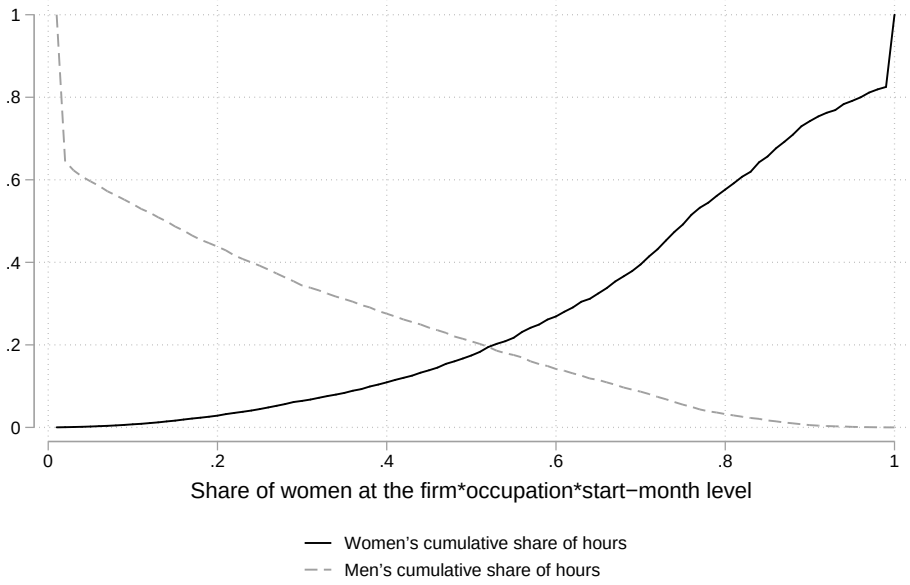
| Year | New jobs | Matched new jobs | % new jobs matched | Job posts | Matched job posts | % job posts matched |
|---|---|---|---|---|---|---|
| 2011 | 714380 | 214182 | 30.0 | 86687 | 38917 | 44.9 |
| 2012 | 657615 | 194150 | 29.5 | 116449 | 58156 | 49.9 |
| 2013 | 672442 | 221291 | 32.9 | 125265 | 66112 | 52.7 |
| 2014 | 767327 | 237177 | 30.8 | 133148 | 69542 | 52.2 |
| 2015 | 731268 | 227433 | 31.1 | 110888 | 59571 | 53.7 |
| Total | 3543032 | 1094233 | 30.8 | 572437 | 292298 | 51.0 |

*Note: The 5-months matching period is not available for all months of 2010 and 2016, and thus, match rates are lower and not comparable to those reported here.*

in a given month, the level of aggregation is low. However, aggregating the job vacancy/skills data at the firm*occupation*start-month levels is a potential drawback of my data: I cannot separate women and men in the job vacancy data, and thus, I assume that everyone has the same skills at the firm*occupation*start-month level. In other words, the same skills are assigned to women and men in the same cells; I do not observe any gender variation in skills at the firm*occupation*start-month levels. If women and men tend to work in the same cells, this would restrict my analysis. However, as pointed out above, women and men tend to work in different occupations in the Danish labour market, i.e. high levels of occupational segregation are observed (Jarman et al., 2012). Due to the smaller cell sizes, gender segregation is likely to be even more pronounced at the firm*occupation*start-month levels. To explore gender segregation at these levels, I first calculate the female share of hours in each firm*occupation*start-month cell. Next, I graph the cumulative distribution of hours worked for women and men respectively on the cell's female share of hours. Figure 1 shows that women and men rarely get employed at the same time in the same firm*occupation*start-month cell. So, despite the fact that I cannot observe any gender variation in skills within firm*occupation*start-month cells, I still observe plenty of gender variation in skills across these cells. Furthermore, I do observe gender differences in wages and in all other characteristics within a cell; these variables vary at the individual level.

An average match rate of 30.8 % of BFL-jobs can be problematic if the matched jobs spell are not representative of the population of new job spell. To check whether or not all occupations and industries are represented in the matched data, I compare the occupational and industrial distribution in the complete BFL data and in the matched subsample. Figures showing the distributions are included in Appendix A.3. The significant overrepresentation of public employees in the matched sample follows from the fact that all permanent public sector jobs by law must be publicly advertised. Thus, public sector job vacancies are also overrepresented in the vacancy data. Notice, however, how all occupations and industries

Figure 1: Cumulative distribution of hours worked




*Source: BFL 2010-2016, excluding observation with missing CVR- or DISCO-codes.*
*Note: Cumulative distribution of hours worked by women and men on the share of women in firm*occupation*start-month cells. Notice that hours worked by men is concentrated in cells with a low share of women and vice versa.*

are represented in the matched data. I include a variable in the matched data to indicate whether a job is in the public or private sector, which is utilised in as a control variable in the data analyses that follow below.

## *2.4. Skill measures*

After matching the job spells and job posts, the categorised keywords are revisited. Recall that a job spell can be matched with more than one job posts and vice versa. If a job spell is matched with more than one job post, keywords from all the relevant job posts are aggregated. Next, the number of (aggregated) keywords belonging to the nine skill categories as well as noise words are counted for each job spell. Using these counts, the fraction of keywords indicating a certain skill are calculated for each job spell. For example, a job spell may be matched with one job post, which contains 4 % "character" words. Or a job spell may be matched with two job posts, which in total contain 8 % "character" words. However, these skill fractions are hard to interpret, and particularly in regressions analyses.

A more easily interpretable alternative would be to classify each job spell as either "character" or "not character", i.e. to create an indicator variable for each skill category. Indicator variables are easy to interpret, particularly in regression analyses with interaction terms.

However, almost all job posts include one or more "character" keywords. Hence, there would be little variation in the skill measure if all job posts that include a single "character" keyword were classified as "character" rather than "not character". At the same time, other skill keywords are relatively rare, e.g. keywords indicating "computer (specific)" skills. Therefore, a simple data-driven approach is used to classify each job post as either "character" or "not character", and analogously for the remaining eight skills.

First, I consider the non-zero fractions of "charater" keywords for each job spell: At which point in distribution does the fraction of "character" keywords predict anything about wage levels? In order to determine this, I do the following: 1) Calculate each percentile in the distribution of non-zero "character" fractions. 2) Construct percentile-dummies indicating whether or not a job spell's "character" fraction is above or below each percentile. 3) Separately regress ln(hourly wages) on each of the percentile-dummies and a constant, but no control variables. 4) Choose the percentile-dummy which yields the most predictive power (the highest $r^2$). 5) Classify each job spell as "character" if the fraction of "character" keywords equals or exceeds that of the percentile determined by the percentile-dummy. This exercise is repeated for the remaining eight skill measures, which concludes the data preprocessing. In the next section, some descriptive statistics are reported.

# 3. Descriptive statistics

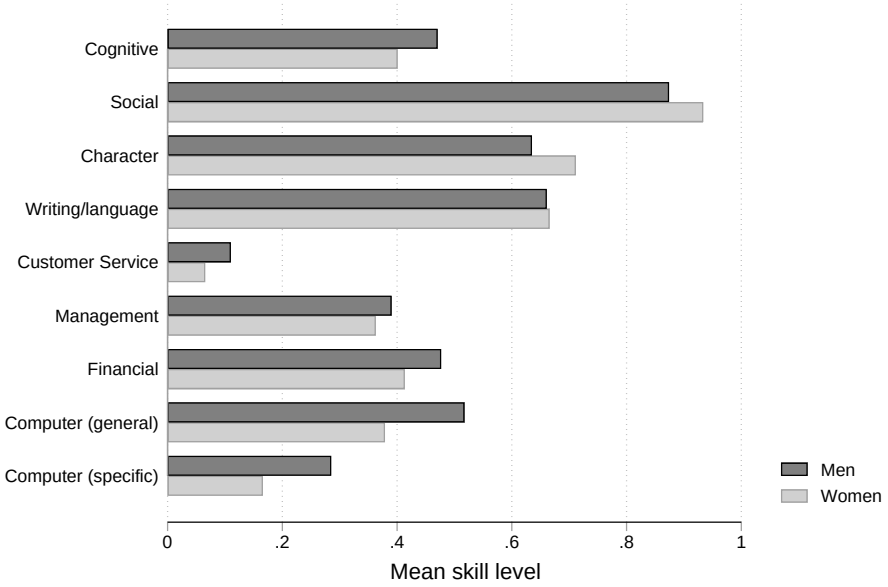
Before moving on to regression analyses of gender differences in returns to skills in the next chapter, some introductory descriptive statistics are provided here. First, I exploit the match between the job vacancy data and register data when reporting gender differences in skills. Second, it is highlighted how the skill measures correlate with each other and with other variables such as wage and years of education. Lastly, I show that not all variation in skill measures can be accounted for by individual-level variation in other variables. Thus, I demonstrate that the skill measures yield explanatory power beyond that of standard labour market data (cf. Deming and Kahn, 2018).

## *3.1. Gender differences in skills*

The vacancy-register data match enables analyses of skills together with the rich sets of variables provided by the Danish registers. In the context of this paper, an essential piece of information to exploit is - of course - the gender of workers. Figure 2 maps the average of jobs categorised as requiring each skill for women and men respectively.

Figure 2 shows that women are overrepresented in jobs that are categorised as requir-

Figure 2: Mean skill levels by gender




*Source: BEF, HBS-Jobindex 2010-2016.*
*Note: Observations weighted by full-time equivalents*

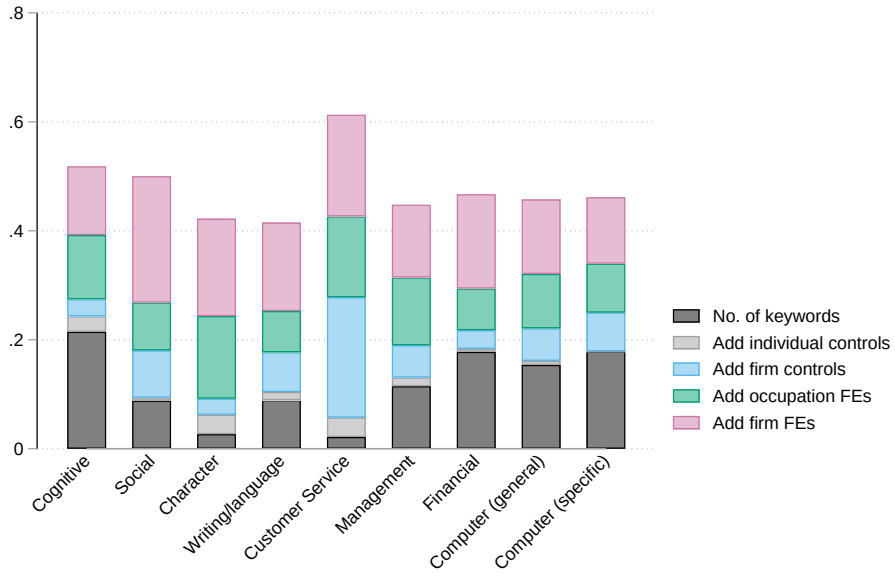
ing social, character, and writing/language skills when compared to men. The opposite is the case for the remaining six skills. Despite some small gender differences, jobs are largely similarly categorised for women and men. The largest relative gender difference observed is in "computer (specific)" skills, where men are more likely to be employed in a firm*occupation*start-month cell that is categorised as requiring computer skills.

## *3.2. Correlations*

Before moving on to regression analysis, simple correlation coefficients between the skill measures, wages, and gender are important to consider for at least two reasons. Firstly, the skill measures should not be too highly correlated, as that could result in multicollinearity issues in regressions. Second, the correlations themselves may give us some idea of whether or not the skill measures make sense to include in wage regressions. For example, one would expect that skilled workers tend to work in cells with higher wages, i.e. that skills measures and wages are positively correlated.

Table 3 includes correlations between log hourly wages, a female dummy variable (=1 for women), and finally, all nine skill measures. All skill measures are positively correlated with wages, with the exception of "character" and "customer service" skills. Most skills are positively correlated with each other, although there are a couple of exceptions: "character"

Figure 3: Adjusted $R^2$ from regressions of skills level on various controls




*Source: Various registers, HBS-Jobindex 2010-2016.*
*Note: Skills are regressed on various sets of controls.*
*The sets of controls are described in the section on regression models.*

and "customer service" are negatively correlated with a few skills. This is an early indication of "character" and "customer service" skills being common in low wage jobs and in jobs with few other skills. Importantly, no skill measures are correlated to a degree that should cause problems of multicollinearity in regression models.

## *3.3. Variance*

Although the correlation coefficients indicate that my skill measures are not correlated to a degree that would cause multicollinearity issues in regression models, the variance of the skill measures should also be explored. Before moving on to regression analyses it must be established that skill requirements cannot be entirely predicted by potential covariates. If so, the skill measures would not add any explanatory to a regression model. Thus, I regress the nine skill measures on various sets of control variables, and plot the adjusted $R^2$ from each regression:

Figure 3 shows that between 41 % and 62 % of the variance in the skill measures can be explained by the most extensive set of covariates. Notice that occupation and firm fixed-effects explain particularly large fractions of the variance in skill requirements. Still, a significant share of the variance in skill demands cannot be explained by even the most extensive set

Table 3: Correlation tables of skills, wages and gender

| | ln(wage) | Female | Cognitive | Social | Character | Writing/ language | Customer Service | Manage-ment | Financial | Computer (general) | Computer (specific) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ln(wage) | 1 | | | | | | | | | | |
| Female | -0.226 | 1 | | | | | | | | | |
| Cognitive | 0.320 | -0.0691 | 1 | | | | | | | | |
| Social | 0.0509 | 0.102 | 0.128 | 1 | | | | | | | |
| Character | -0.210 | 0.0805 | -0.208 | 0.0839 | 1 | | | | | | |
| Writing/ language | 0.156 | 0.00504 | 0.215 | 0.213 | -0.0529 | 1 | | | | | |
| Customer Service | -0.146 | -0.0790 | -0.103 | -0.138 | 0.0752 | -0.171 | 1 | | | | |
| Management | 0.246 | -0.0279 | 0.256 | 0.116 | -0.141 | 0.0988 | -0.110 | 1 | | | |
| Financial | 0.236 | -0.0627 | 0.295 | 0.145 | -0.125 | 0.157 | -0.0608 | 0.253 | 1 | | |
| Computer (general) | 0.254 | -0.137 | 0.360 | 0.0807 | -0.183 | 0.183 | 0.000133 | 0.166 | 0.238 | 1 | |
| Computer (specific) | 0.247 | -0.143 | 0.263 | 0.0594 | -0.208 | 0.157 | -0.0674 | 0.116 | 0.212 | 0.319 | 1 |

*Note: Observations weighted by full-time equivalents.*

*All correlations are significant at the 0.001-level*

of covariates. Thus, the skill measures appear as suitable regressors in regressions in which similar sets of covariates are included. The next section introduces the regression analyses used to estimate gender differences in returns to skills.

# 4. Regression analyses

In this section, I first outline the regression models used to estimate gender differences in returns to skills. Next, results are presented, and finally, a few robustness checks are reported.

## *4.1. Models*

Regressing hourly wages on skills and their gender interactions will indicate whether women and men with same skill requirements also receive the same wage. The econometric methods I rely on are simple but suitable for the question at hand. I regress ln(hourly wages) on skills and female*skills interactions with extensive sets of control variables and fixed effects. Before writing out the relevant regression models, I outline the four sets of control variables and fixed effects that are applied:

1. Individual controls and and number of keywords: *parent dummy, parent*female interaction, age, age$^2$, years of experience, years of education, immigrant dummy, marriage dummy, part-time dummy, year FEs, start-month FEs, number of keywords*
2. Firm controls: *1-letter industry dummies, firm location, number of employees, a private sector dummy*
3. 3-digit occupation fixed effects
4. Firm fixed effects

The parent dummy equals one if an individual has a child less than 18 years old. The parent*female interaction term is included to control for the gender specific effects of parenthood (cf. Kleven, Landais, and Søgaard, 2018), which may otherwise affect the estimates of gender differences in returns to skills. Controls are additively included in the regression models, so it suffices to write out the full model including all controls and fixed effects:

$$w_{iofym} = \beta_0 + I_{iym}\beta_1 + F_{fy}\beta_2 + S_{ofym}\gamma_1 + S_{ofym} \times g_i \times \gamma_2 + \lambda_o + \ \phi_f + \theta_y + \delta_m + \varepsilon_{iofym}$$

Where the subscript $i$ indicates variation at the individual-level, $f$ at the firm-level, $o$ at the occupation-level, $m$ at the *start*-month-level, and $y$ at the year-level. $w_{iofy}$ is ln(hourly

wage). $I_{iym}$ is a matrix of individual start-month-varying characteristics and includes a female dummy, $F_{fy}$ a matrix of firm year-varying characteristics, $S_{ofym}$ is a matrix of the nine skill measures that vary at the firm*occupation*year*start-month-level. $g_i$ is a female dummy variable, which equals 1 for women only. $\lambda_o$ are the occupation FEs, $\phi_f$ the firm FEs. Finally, $\delta_m$ are start-month FEs and $\theta_y$ are year FEs. Note that the start-month FEs and year FEs are *not* interacted, and thus, they are not year*start-month FEs. The vector $\gamma_1$ gives the coefficients on the nine skill measures for men and $\gamma_1 + \gamma_2$ for women, i.e. $\gamma_2$ is the gender differences in the skill measures' coefficients.

As the nine skill measures only vary at the firm*occupation*start-month levels, the errors $\varepsilon_{iofym}$ are correlated within these cells. Thus, I follow the approach taken by Hersch (1998) and cluster my standard errors at these levels. Such an approach is also recommended by Cameron and Miller (2015). The nine skill measures are perfectly correlated within clusters (they do not vary within), and thus, applying cluster-robust standard errors significantly inflate the estimated errors. After briefly discussing identification in the next section, the estimated coefficients and their cluster-robust standard errors are presented.

## *4.2. Identification*

The full specification with both occupation and firm FEs targets the issue that workers with certain skill compositions may sort into high/low paying occupations and firms. Note that including both occupation and firm FEs is not analogue to including firm*occupation FEs, and thus, variation specific to the firm*occupation interactions remains. The estimated coefficients can be interpreted as *within*-firm*occupation returns to skills and *within*-firm*occupation gender differences in returns.

Although the full specification with various controls, occupation, and firm FEs may identify returns to skills, one should consider omitted variable/ability bias at the individual level. However, not many individuals start more than one job spell in rather short sample period (May 2010-January 2016). Furthermore, there is little variation in skills for the few individuals that do (again due to the short sample period), and thus, a specification with including individual FEs is not feasible, although estimates from such a specification would warrant a more causal interpretation.

## *4.3. Results*

First, estimation results from the linear regression model outlined above are reported, but *excluding* the gender interactions with the nine skill measures. These results are reported in Table 4. The estimates show that when both occupation and firm FEs are included in

the model, the coefficients on the skill measures tend to be insignificant. An exception is jobs categorised as requiring "customer service" skills, which appear is associated with lower wages. The coefficient on "social" skills is also marginally significant and negative. Thus, one could jump to the premature conclusion that task-specific skills generally do not matter for wage formation beyond what can be explained by occupation and firm FEs.

However, after including gender interactions with the nine skills measures in the model, a different picture emerges, see Table 7. In the model including both occupation and firm FEs, consider the coefficients on the skill measures that are not interacted with the female dummy, $\gamma_1$. These coefficients can be interpreted as *within*-firm*occupation returns to skills for men. The coefficients on the "cognitive", "financial", and "computer (specific)" skill measures are positive and significant even after introducing occupation and firm FEs. However, the coefficients on the "social" and "customer service" measure are negative and significant.

In the model including both occupation and firm FEs, the coefficients on the female*skill interactions, $\gamma_2$, can be interpreted as *within*-firm*occupation gender differences in returns. Notice the coefficient on the interaction terms with the "cognitive", "customer service", "financial", "computer (general)", and "computer(specific)" all are negative and highly significant. Thus, for five out of nine skill measures, women face lower returns. The interaction term with the "social" skill measure is positive and significant at the 0.05-level. These estimates are the main take away from this paper. If the gender dimension of returns to skills was ignored, I could have concluded that skill generally do not yield returns beyond what can be explained by occupation and firm FEs. Instead, this specification set the stage for the conclusion that men face positive returns to a number of skills, even within firm*occupation cells, and that women face lower returns than men to most skills, and women's returns to skills are near zero ($\gamma_1 + \gamma_2$). However, before jumping to this conclusion, a couple of robustness checks must be considered.

## *4.4. Robustness*

A couple of robustness checks are performed. First, the full specification with both occupation and firm FEs is re-estimated for a number of subpopulations in order to check whether or not the results from the previous section are driven by a certain group of workers. In the existing literature, computer and cognitive skills have been highlighted as complementing technology and technological change, and thus, yielding positive returns. However, the results from the previous section indicate that this is only the case for men. Social skills have also been emphasised as complementing technological change, but together with cognitive skills (Deming, 2017; Deming and Kahn, 2018). Therefore, potential interactions between

Table 4: ln(hourly wage) regressed on skills, but no gender interaction

| | Dependent variable: ln(hourly wages) | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Individual controls | Occupation FEs | Firm FEs | Both FEs |
| Female=1 | -0.0487*** | -0.0391*** | -0.0465*** | -0.0390*** |
| | [0.00158] | [0.00123] | [0.00136] | [0.00115] |
| Cognitive=1 | 0.0273*** | 0.00586** | 0.0190*** | 0.00221 |
| | [0.00261] | [0.00199] | [0.00245] | [0.00204] |
| Social=1 | -0.0111 | -0.0168** | 0.000516 | -0.00716* |
| | [0.00860] | [0.00612] | [0.00444] | [0.00324] |
| Character=1 | -0.0396*** | -0.00757** | -0.0359*** | -0.00405 |
| | [0.00292] | [0.00248] | [0.00252] | [0.00220] |
| Writing/language=1 | 0.00404 | 0.00152 | 0.00229 | -0.00234 |
| | [0.00295] | [0.00266] | [0.00259] | [0.00222] |
| Customer Service=1 | -0.0843*** | -0.0268*** | -0.0679*** | -0.0259*** |
| | [0.00689] | [0.00419] | [0.00624] | [0.00416] |
| Management=1 | 0.0538*** | 0.00490* | 0.0529*** | 0.000306 |
| | [0.00283] | [0.00245] | [0.00280] | [0.00214] |
| Financial=1 | 0.00793* | 0.000383 | 0.0121*** | 0.00347 |
| | [0.00376] | [0.00267] | [0.00256] | [0.00216] |
| Computer (general)=1 | 0.00695* | 0.00188 | 0.00174 | -0.00373 |
| | [0.00283] | [0.00194] | [0.00283] | [0.00194] |
| Computer (specific)=1 | -0.00451 | 0.00328 | -0.0125*** | 0.00250 |
| | [0.00308] | [0.00251] | [0.00262] | [0.00212] |
| Parent=1 | 0.0706*** | 0.0573*** | 0.0660*** | 0.0541*** |
| | [0.00155] | [0.00122] | [0.00135] | [0.00114] |
| Parent=1 × Female=1 | -0.0739*** | -0.0597*** | -0.0690*** | -0.0561*** |
| | [0.00176] | [0.00141] | [0.00155] | [0.00129] |
| $R^2$ | 0.459 | 0.568 | 0.516 | 0.607 |
| N | 938435 | 938435 | 938432 | 938432 |
| Individual controls | Yes | Yes | Yes | Yes |
| Firm controls | Yes | Yes | Yes | Yes |
| Occupation FEs | No | Yes | No | Yes |
| Firm FEs | No | No | Yes | Yes |

Observations weighted by full-time equivalents.

Cluster-robust standard errors in brackets.

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 5: ln(hourly wage) regressed on skills with gender interaction

| | Dependent variable: ln(hourly wages) | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Individual controls | Occupation FEs | Firm FEs | Both FEs |
| Female=1 | -0.0442*** | -0.0215** | -0.0328*** | -0.0201*** |
| | [0.00807] | [0.00751] | [0.00484] | [0.00453] |
| Cognitive=1 | 0.0399*** | 0.0169*** | 0.0309*** | 0.0127*** |
| | [0.00408] | [0.00297] | [0.00340] | [0.00254] |
| Social=1 | -0.0294** | -0.0239** | -0.0144** | -0.0129** |
| | [0.0104] | [0.00845] | [0.00539] | [0.00409] |
| Character=1 | -0.0402*** | -0.00718* | -0.0353*** | -0.00321 |
| | [0.00387] | [0.00302] | [0.00313] | [0.00237] |
| Writing/language=1 | 0.00211 | 0.00328 | 0.00288 | -0.000309 |
| | [0.00460] | [0.00401] | [0.00356] | [0.00280] |
| Customer Service=1 | -0.0572*** | -0.0149** | -0.0479*** | -0.0132** |
| | [0.00747] | [0.00528] | [0.00662] | [0.00467] |
| Management=1 | 0.0657*** | 0.00791* | 0.0621*** | -0.00125 |
| | [0.00429] | [0.00330] | [0.00379] | [0.00254] |
| Financial=1 | 0.0262*** | 0.0114** | 0.0359*** | 0.0165*** |
| | [0.00549] | [0.00407] | [0.00341] | [0.00248] |
| Computer (general)=1 | 0.00944* | 0.00933** | 0.00378 | 0.00308 |
| | [0.00403] | [0.00292] | [0.00398] | [0.00265] |
| Computer (specific)=1 | 0.00267 | 0.00672 | -0.0112** | 0.00782** |
| | [0.00446] | [0.00347] | [0.00346] | [0.00283] |
| Female=1 × Cognitive=1 | -0.0193*** | -0.0179*** | -0.0182*** | -0.0168*** |
| | [0.00372] | [0.00282] | [0.00307] | [0.00230] |
| Female=1 × Social=1 | 0.0360*** | 0.0135* | 0.0268*** | 0.0104* |
| | [0.00703] | [0.00636] | [0.00438] | [0.00427] |
| Female=1 × Character=1 | 0.000444 | -0.000867 | -0.00116 | -0.00134 |
| | [0.00341] | [0.00279] | [0.00268] | [0.00230] |
| Female=1 × Writing/language=1 | 0.00293 | -0.00279 | -0.000777 | -0.00306 |
| | [0.00409] | [0.00356] | [0.00329] | [0.00265] |
| Female=1 × Customer Service=1 | -0.0559*** | -0.0261*** | -0.0405*** | -0.0276*** |
| | [0.00506] | [0.00488] | [0.00420] | [0.00406] |
| Female=1 × Management=1 | -0.0192*** | -0.00475 | -0.0145*** | 0.00274 |
| | [0.00418] | [0.00332] | [0.00334] | [0.00261] |
| Female=1 × Financial=1 | -0.0294*** | -0.0183*** | -0.0380*** | -0.0209*** |
| | [0.00424] | [0.00352] | [0.00300] | [0.00242] |
| Female=1 × Computer (general)=1 | -0.00450 | -0.0128*** | -0.00428 | -0.0113*** |
| | [0.00336] | [0.00268] | [0.00298] | [0.00227] |
| Female=1 × Computer (specific)=1 | -0.0148*** | -0.00706* | -0.00355 | -0.00983*** |
| | [0.00406] | [0.00305] | [0.00296] | [0.00251] |
| Parent=1 | 0.0692*** | 0.0562*** | 0.0645*** | 0.0532*** |
| | [0.00157] | [0.00121] | [0.00135] | [0.00113] |
| Parent=1 × Female=1 | -0.0726*** | -0.0585*** | -0.0673*** | -0.0550*** |
| | [0.00179] | [0.00142] | [0.00156] | [0.00129] |
| $R^2$ | 0.461 | 0.569 | 0.518 | 0.608 |
| N | 938435 | 938435 | 938432 | 938432 |
| Individual controls | Yes | Yes | Yes | Yes |
| Firm controls | Yes | Yes | Yes | Yes |
| Occupation FEs | No | Yes | No | Yes |
| Firm FEs | No | No | Yes | Yes |

Observations weighted by full-time equivalents.

Cluster-robust standard errors in brackets.

* p < 0.05, ** p < 0.01, *** p < 0.001

these skill measures are also explored.

### *4.4.1. Subpopulations*

The first robustness check focusses on the full specification, which include all occupation and firm FEs as well as various other controls. The model is estimated again on the following subpopulations:

a. Professionals [7]
b. Workers in large firms (with 100 or more employees)
c. Workers in small firms (with fewer than 100 employees)
d. Full-time workers
e. Workers that remain employed for at least 12 months after commencing a job spell

This exercise may reveal that the results from the full datasets may be "driven" by certain subgroups. Results for each subpopulation are reported in Table 6.

Notice that the results from the previous subsection generally hold for all the selected subpopulations. However, the results for relatively small sample of job spells of workers at smaller firms differ to some degree. For example, the coefficient on cognitive skills is insignificant, although the coefficient on its interaction term with the female dummy remains negative and significant. The number of job spells per firms is naturally lower for small firms, and thus, the firm FEs may account for more of the variation in wages at these firms.

### *4.4.2. Interactions*

In the existing literature, interactions between certain skills are often emphasised in the context of technological change. Thus, four sets of interaction terms are explored. First, an interaction term between the "social" and "cognitive" skill measures is included. This interaction term is of particular interest after the recent work by Deming (2017), and Deming and Kahn (2018). Next, interaction terms between "cognitive" and "computer (general)", as well as "cognitive" and "computer (specific)" are included respectively. Lastly, both of the computer skills are interacted with the "cognitive" skill measure. In the earlier literature on the RBTC hypothesis, cognitive and computer skills were emphasised as complementing technology during recent periods of technological change. The four sets of interaction terms are also interacted with the female dummy to check for possible gender differences in returns to these. The estimates are reported in Table 7.

[7] Here, professionals are crudely defined as workers with the 1-digit DISCO-codes "1", "2", or the 3-digit code "321".Deming and Kahn (2018) use SOC-codes, but the crosswalk between SOC- and DISCO-codes include numerous many-to-many walks, so it cannot be applied in this context.

Table 6: ln(hourly wage) regressed on skills with gender interaction for subpopulations

| | Dependent variable: ln(hourly wages) | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | All | Professionals | Large firms | Small firms | Full-time | Whole year |
| Female=1 | -0.0201*** | -0.0295*** | -0.0208*** | -0.00891 | -0.0324*** | -0.0227*** |
| | [0.00453] | [0.00845] | [0.00513] | [0.00639] | [0.00531] | [0.00515] |
| Cognitive=1 | 0.0127*** | 0.0124*** | 0.0120*** | 0.00794 | 0.0124*** | 0.0137*** |
| | [0.00254] | [0.00315] | [0.00262] | [0.00493] | [0.00278] | [0.00278] |
| Social=1 | -0.0129** | -0.0217** | -0.0143** | 0.00705 | -0.0142** | -0.0135** |
| | [0.00409] | [0.00832] | [0.00460] | [0.00472] | [0.00434] | [0.00445] |
| Character=1 | -0.00321 | -0.00594* | -0.00208 | -0.0139** | -0.00520* | -0.00477 |
| | [0.00237] | [0.00291] | [0.00249] | [0.00460] | [0.00261] | [0.00258] |
| Writing/language=1 | -0.000309 | -0.0106* | -0.000635 | 0.00282 | -0.00315 | -0.000999 |
| | [0.00280] | [0.00531] | [0.00295] | [0.00477] | [0.00303] | [0.00304] |
| Customer Service=1 | -0.0132** | -0.00252 | -0.0149** | 0.0126* | -0.00960 | -0.0102* |
| | [0.00467] | [0.0114] | [0.00514] | [0.00561] | [0.00528] | [0.00518] |
| Management=1 | -0.00125 | -0.00198 | -0.00117 | -0.000604 | -0.00355 | -0.00222 |
| | [0.00254] | [0.00317] | [0.00265] | [0.00503] | [0.00269] | [0.00272] |
| Financial=1 | 0.0165*** | 0.0218*** | 0.0172*** | 0.0161** | 0.0187*** | 0.0174*** |
| | [0.00248] | [0.00340] | [0.00253] | [0.00585] | [0.00272] | [0.00267] |
| Computer (general)=1 | 0.00308 | 0.0120*** | 0.00341 | -0.000190 | 0.00257 | 0.00160 |
| | [0.00265] | [0.00315] | [0.00275] | [0.00565] | [0.00285] | [0.00284] |
| Computer (specific)=1 | 0.00782** | 0.00389 | 0.00655* | 0.0177** | 0.00752* | 0.00808** |
| | [0.00283] | [0.00341] | [0.00292] | [0.00655] | [0.00309] | [0.00306] |
| Female=1 × Cognitive=1 | -0.0168*** | -0.0102*** | -0.0163*** | -0.0147* | -0.0157*** | -0.0174*** |
| | [0.00230] | [0.00283] | [0.00237] | [0.00584] | [0.00255] | [0.00251] |
| Female=1 × Social=1 | 0.0104* | 0.0206* | 0.0118* | -0.00565 | 0.0158** | 0.0108* |
| | [0.00427] | [0.00837] | [0.00486] | [0.00535] | [0.00488] | [0.00477] |
| Female=1 × Character=1 | -0.00134 | 0.000444 | -0.00195 | 0.00533 | -0.00198 | -0.000224 |
| | [0.00230] | [0.00262] | [0.00239] | [0.00598] | [0.00263] | [0.00249] |
| Female=1 × Writing/language=1 | -0.00306 | 0.00817 | -0.00300 | -0.00943 | -0.0000893 | -0.00203 |
| | [0.00265] | [0.00454] | [0.00279] | [0.00524] | [0.00305] | [0.00289] |
| Female=1 × Customer Service=1 | -0.0276*** | -0.0710*** | -0.0272*** | -0.0270*** | -0.0371*** | -0.0292*** |
| | [0.00406] | [0.00893] | [0.00437] | [0.00745] | [0.00460] | [0.00458] |
| Female=1 × Management=1 | 0.00274 | 0.00290 | 0.00342 | -0.0192** | 0.00479 | 0.00453 |
| | [0.00261] | [0.00286] | [0.00269] | [0.00639] | [0.00280] | [0.00277] |
| Female=1 × Financial=1 | -0.0209*** | -0.0245*** | -0.0210*** | -0.0352*** | -0.0228*** | -0.0217*** |
| | [0.00242] | [0.00303] | [0.00247] | [0.00730] | [0.00267] | [0.00258] |
| Female=1 × Computer (general)=1 | -0.0113*** | -0.0138*** | -0.0114*** | -0.0211** | -0.0114*** | -0.0110*** |
| | [0.00227] | [0.00282] | [0.00236] | [0.00646] | [0.00249] | [0.00246] |
| Female=1 × Computer (specific)=1 | -0.00983*** | -0.0110*** | -0.00821** | -0.0244** | -0.00929*** | -0.0110*** |
| | [0.00251] | [0.00303] | [0.00258] | [0.00796] | [0.00270] | [0.00271] |
| Parent=1 | 0.0532*** | 0.0525*** | 0.0531*** | 0.0497*** | 0.0497*** | 0.0527*** |
| | [0.00113] | [0.00169] | [0.00117] | [0.00341] | [0.00123] | [0.00126] |
| Parent=1 × Female=1 | -0.0550*** | -0.0550*** | -0.0552*** | -0.0440*** | -0.0481*** | -0.0541*** |
| | [0.00129] | [0.00188] | [0.00134] | [0.00431] | [0.00144] | [0.00142] |
| $R^2$ | 0.608 | 0.602 | 0.607 | 0.653 | 0.646 | 0.629 |
| N | 938432 | 379798 | 877097 | 61254 | 543647 | 603751 |
| Individual controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Occupation FEs | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm FEs | Yes | Yes | Yes | Yes | Yes | Yes |

Observations weighted by full-time equivalents.

Cluster-robust standard errors in brackets.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: ln(hourly wage) regressed on skills, skill interactions, and gender interactions

| | Dependent variable: ln(hourly wages) | | | | |
|---|---|---|---|---|---|
| | (1) No interactions | (2) Cognitive and social | (3) Cognitive and computer(specific) | (4) Cognitive and computer(general) | (5) Cognitive and computer(both) |
| Female=1 | -0.0201*** | -0.0208*** | -0.0205*** | -0.0218*** | -0.0220*** |
| | [0.00453] | [0.00491] | [0.00455] | [0.00450] | [0.00452] |
| Cognitive=1 | 0.0127*** | 0.00259 | 0.0105*** | 0.00631 | 0.00552 |
| | [0.00254] | [0.00786] | [0.00267] | [0.00334] | [0.00340] |
| Social=1 | -0.0129** | -0.0160*** | -0.0128** | -0.0128** | -0.0127** |
| | [0.00409] | [0.00471] | [0.00408] | [0.00407] | [0.00406] |
| Character=1 | -0.00321 | -0.00311 | -0.00310 | -0.00272 | -0.00270 |
| | [0.00237] | [0.00237] | [0.00236] | [0.00237] | [0.00237] |
| Writing/language=1 | -0.000309 | -0.000290 | -0.000227 | -0.000344 | -0.000289 |
| | [0.00280] | [0.00280] | [0.00280] | [0.00279] | [0.00279] |
| Customer Service=1 | -0.0132** | -0.0134** | -0.0131** | -0.0128** | -0.0127** |
| | [0.00467] | [0.00467] | [0.00467] | [0.00465] | [0.00465] |
| Management=1 | -0.00125 | -0.00127 | -0.00112 | -0.00124 | -0.00116 |
| | [0.00254] | [0.00254] | [0.00254] | [0.00253] | [0.00253] |
| Financial=1 | 0.0165*** | 0.0165*** | 0.0165*** | 0.0166*** | 0.0165*** |
| | [0.00248] | [0.00248] | [0.00247] | [0.00247] | [0.00247] |
| Computer (general)=1 | 0.00308 | 0.00310 | 0.00301 | -0.00219 | -0.00168 |
| | [0.00265] | [0.00265] | [0.00265] | [0.00357] | [0.00356] |
| Computer (specific)=1 | 0.00782** | 0.00785** | 0.00249 | 0.00714* | 0.00372 |
| | [0.00283] | [0.00283] | [0.00394] | [0.00285] | [0.00388] |
| Female=1 × Cognitive=1 | -0.0168*** | -0.0149 | -0.0145*** | -0.00753* | -0.00683* |
| | [0.00230] | [0.00903] | [0.00247] | [0.00307] | [0.00313] |
| Female=1 × Social=1 | 0.0104* | 0.0114* | 0.0102* | 0.0102* | 0.0101* |
| | [0.00427] | [0.00498] | [0.00426] | [0.00419] | [0.00418] |
| Female=1 × Character=1 | -0.00134 | -0.00144 | -0.00147 | -0.00208 | -0.00210 |
| | [0.00230] | [0.00231] | [0.00229] | [0.00230] | [0.00230] |
| Female=1 × Writing/language=1 | -0.00306 | -0.00308 | -0.00315 | -0.00303 | -0.00308 |
| | [0.00265] | [0.00265] | [0.00264] | [0.00264] | [0.00264] |
| Female=1 × Customer Service=1 | -0.0276*** | -0.0275*** | -0.0278*** | -0.0287*** | -0.0287*** |
| | [0.00406] | [0.00405] | [0.00407] | [0.00397] | [0.00398] |
| Female=1 × Management=1 | 0.00274 | 0.00273 | 0.00255 | 0.00281 | 0.00271 |
| | [0.00261] | [0.00261] | [0.00262] | [0.00261] | [0.00262] |
| Female=1 × Financial=1 | -0.0209*** | -0.0209*** | -0.0208*** | -0.0210*** | -0.0210*** |
| | [0.00242] | [0.00242] | [0.00241] | [0.00241] | [0.00241] |
| Female=1 × Computer (general)=1 | -0.0113*** | -0.0114*** | -0.0112*** | -0.00272 | -0.00320 |
| | [0.00227] | [0.00227] | [0.00228] | [0.00294] | [0.00293] |
| Female=1 × Computer (specific)=1 | -0.00983*** | -0.00981*** | -0.00364 | -0.00864*** | -0.00537 |
| | [0.00251] | [0.00250] | [0.00342] | [0.00251] | [0.00339] |
| Cognitive=1 × Social=1 | | 0.0110 | | | |
| | | [0.00807] | | | |
| Female=1 × Cognitive=1 × Social=1 | | -0.00234 | | | |
| | | [0.00927] | | | |
| Cognitive=1 × Computer (specific)=1 | | | 0.00885 | | 0.00586 |
| | | | [0.00522] | | [0.00520] |
| Female=1 × Cognitive=1 × Computer (specific)=1 | | | -0.0101* | | -0.00537 |
| | | | [0.00481] | | [0.00486] |
| Cognitive=1 × Computer (general)=1 | | | | 0.0120* | 0.0107* |
| | | | | [0.00487] | [0.00487] |
| Female=1 × Cognitive=1 × Computer (general)=1 | | | | -0.0195*** | -0.0183*** |
| | | | | [0.00431] | [0.00437] |
| Parent=1 | 0.0532*** | 0.0532*** | 0.0532*** | 0.0532*** | 0.0532*** |
| | [0.00113] | [0.00113] | [0.00113] | [0.00112] | [0.00113] |
| Parent=1 × Female=1 | -0.0550*** | -0.0550*** | -0.0550*** | -0.0550*** | -0.0550*** |
| | [0.00129] | [0.00129] | [0.00129] | [0.00129] | [0.00129] |
| $R^2$ | 0.608 | 0.608 | 0.608 | 0.608 | 0.608 |
| N | 938432 | 938432 | 938432 | 938432 | 938432 |
| Individual controls | Yes | Yes | Yes | Yes | Yes |
| Firm controls | Yes | Yes | Yes | Yes | Yes |
| Occupation FEs | Yes | Yes | Yes | Yes | Yes |
| Firm FEs | Yes | Yes | Yes | Yes | Yes |

Observations weighted by full-time equivalents.
Cluster-robust standard errors in brackets.
* p < 0.05, ** p < 0.01, *** p < 0.001

The coefficients on the "social"*"cognitive" term and its interaction with the female dummy are insignificant, and after including these terms, the coefficient on the "cognitive" skill measure and its interaction are also insignificant. However, the interaction terms with the computer skill measures are more interesting. Returns to "computer (general)" skills do not only appear to be gender specific, but they also depend on the interaction with "cognitive" skills. Returns to "computer (specific)" skills, however, does not appear to depend on the interaction with the "cognitive" skill measure. One explanation for this could be that cognitive skills are assumed, when "computer (specific)" skills are demanded. Such an assumption cannot be captured by the job vacancy data, where only the revealed demand for skills can be measured.

# 5. Discussion

This section discusses my results in relation to the existing literature. The combination of Danish job vacancy data and individual-level register data is unique to this study. Internationally, only few studies has merged job vacancy data with individual-level data, but with much lower match rates (e.g.Kettemann et al., 2018). Thus, I am among the first to be able to utilise individual-level variation in characteristics and wages together with data from job vacancies.

I find a positive association between hourly wages and the "cognitive" skill measure, even after including occupation and firm FEs, but only for men. The coefficient on the "social" skill measure alone is negative, and thus, social skills are associated with lower for men. The interaction between "social" skills and the female dummy is positive, but the total effect of social skills is also negative for women, although of a small magnitude. The coefficient on the dual requirement of both social and cognitive skills is insignificant, and so the coefficients with dual requirement's interaction with the female dummy. This contrast a number of studies on returns to skills, which highlight the interaction between social and cognitive skills as particularly important (Deming, 2017; Deming and Kahn, 2018; Weinberger, 2014). This may be due to differences between the US and Danish labour market, but it may also be due to the fact that individual-level data on skills and wages have not been utilised in a US context.

Because of the relatively short sample period, I cannot confirm whether or not changing skills prices has caused a narrowing of the gender pay gap (as pointed out by Bacolod and Blum, 2010; Beaudry and Lewis, 2014; Rendall, 2010; Yamaguchi, 2018). However, my results do indicate that differences in returns to skills contribute to the gender pay gap (cf. Lindley, 2012). More specifically, I generally find negative and significant coefficients

on the female interactions with the "cognitive", "customer service", "financial", "computer (general)", and "computer(specific)" skill measures. Thus, ignoring the gendered dimension of returns to skills would lead to biased results, overestimating the returns to skills for women and underestimating them for men (with the exception of social skills).

Furthermore, I find that interaction between the "cognitive" and "computer (general)" skill measures are associated with higher wages for men, but not for women. The interaction between cognitive and computer skills was highlighted as particularly important in the early literature on the RBTC hypothesis. My results indicate that this skill interaction is still important for wage formation as it is associated with higher wages for men, but not for women, even after controlling for occupation and firm FEs.

# 6. Conclusions

In this paper, a novel combination of Danish job vacancy and Danish matched employeer-employee data is operationalised for the first time. I derive nine task-specific skill measures from job posts by a reading of keywords. The vacancy and register data are matched on the firm*occupation*start-month-level, which involves some aggregation. However, a high degree of gender segregation at these levels preserves variation in skills across genders. With this combination of data, it is possible to show that variation in skills cannot be entirely explained by an extensive set of control variables, occupation FEs, and firm FEs. Keeping this in mind, the skills measures are included in wage regressions. After including an extensive set of control variables and FEs in wage regressions, the coefficients on the nine skill measure are largely insignificant in the models, but only when gender interactions with the skill measures are excluded. When including interactions between gender and skills, a different story emerges. Even after including occupation FEs, firm FEs, and an extensive set of control variables, the coefficients on the female interactions with "cognitive", "customer service", "financial", "computer (general)", and "computer(specific)" skills are negative and significant. "social" skills are associated with lower wages, but more so for men than for women. Thus, ignoring the gendered dimension of returns to skills would lead to biased results and conclusions. Additionally, interactions between the skill measures are considered. After including controls and FEs, the interaction between the "cognitive" and "social" skill measures does not have a significant effect on wages. However, the interaction between the skill measures "cognitive" and "computer (general)" is associated with higher wages for men, but not for women.

# Appendix A. Data details

## *A.1. Register data*

Observations with the following characteristics are sequentially excluded from the sample:

a) With a missing DISCO-code or firm identifier
b) With a total number of hours for a given year below the equivalent of a full-time month (1923.96/12 as defined by Statistics Denmark) or above 3,500 hours. Part-time workers remain included.
c) Aged under 20, or over 65 when their job spell commences.
d) With an hourly wage below 30 DKK or above 5,000 DKK (in 2016-levels)
e) With total wages exceeding 10,000,000 DKK (in 2016-levels)
f) Enrolled at an educational institution when their job spell commences.
g) At firms with less than 5 full-time equivalents in the matched sample (see below).
h) In an 3-digit occupational group with less than 50 full-time equivalents in matched sample (see below).

Criterion a) and b) are the most restrictive. Criterion a) is necessary to construct job spells at the firm*occupation level, missing DISCO-codes are mostly observed in the private sector. I describe the DISCO-coding in detail below. Criterion b) is imposed to avoid observations where hours of work may be misreported, e.g. freelance work. In addition, I believe that jobs spells with fewer hours than the equivalent of a full-time month are less likely to appear in the job vacancy data due to fixed costs of hiring.

## *A.2. Vacancy data*

As pointed out in the section 2.2, all keywords are assigned either a skill category or a noise tag. This is done the following way: 1) The most frequent keywords (approx. 2000) are assigned a skill category or noise tag manually. These words amount to the vast majority of keyword-observations. 2) Using online dictionary APIs each word's synonyms are obtained. Each word's synonyms are assigned the same category. 3) Using online dictionary APIs the each word's definitions are obtained. 4) Using the definition of the words, the remaining non-categorised words are assigned a category using machine-learning methods. The machine-learning methods are described in more detail here.

The training set consists of both the more than 2000 manually categorised words and their categorised synonyms. In order to categorise the remaining words, the dictionary definition of each keyword obtained from two dictionaries, one Danish dictionary and one English

dictionary. To use the English dictionary, the keywords are translated beforehand. Although the translation step may seem tedious, it involves some regularisation of the keywords, which again helps when looking up definitions of the words. Next, the classification exercise is undertaken.

Two approaches to the classification problem is repeated for both Danish and English versions of the keywords' definitions. The first approach is a one-step categorisation, where each keywords is assigned one of 10 categories, i.e. either one of the nine skills or a noise tag. A Random Forest Predictor is used for this exercise. The second approach is a two-step categorisation. In the first step, each keyword is classified as either noise or non-noise. In the second step each non-noise word is assigned to one of the nine skill categories. For both steps a Random Forest Predictor is applied.

Thus, four predicted categorisations are available for each keyword that was not a part of the training set: a one-step and a two-step version for both the Danish and English definitions. If predictions from all four approaches agree on a category, the keyword is assigned to this category. The same step is undertaken if predictions from three out of four approaches agree. Some words' definitions are only available in either the Danish or English dictionary. These words are categorised if the two approaches in the same language agree and if the probability of the predicted class is relatively high. For the few words that have not been categorised after these steps, English predictions with very high probabilities are considered and assigned to keywords. The predictions based on the English definitions are typically more reliable due to longer definitions of the keywords. If keywords are not categorised after this procedure, they are assigned a noise tag.

## *A.3. Data match*

### *A.3.1. DISCO-codes*

Although variables on wages and hours in BFL are automatically imputed from the Danish tax authorities' data, the *6-digit* DISCO-codes are not. As they require some "manual" coding, i.e. placing a worker in a category, they do not appear in the Danish tax authorities' data. Hence, Statistics Denmark collect the DISCO-codes in a separate procedure. For public employees, Statistics Denmark impute DISCO-codes directly from the public wage data where every employee's job title/position is recorded. In the private sector, Statistics Denmark collect data on employees from firms with 100 or more employees every year.[8] Smaller firms are sampled to report DISCO-codes on their employees from year to year.

[8]For more details, see:
https://www.dst.dk/ext/loen/Vejl_Lon_ligeaar–pdf

Private employers are supplied with a correspondence table between job titles/positions and DISCO-codes in order to secure consistent reporting.[9] If a private firm is not sampled, Statistics Denmark impute an individual's DISCO-code from the previous year given that changes no in the individual's employment are observed. Otherwise, they estimate a DISCO-code from register data on each individual's education, the industry of the individual's employer, and the individual's membership of an unemployment insurance fund (these funds are often occupation-specific).[10]

In the case of the job vacancy data, HBS first extract a job title from each job post. Using a correspondence table between job titles/positions and DISCO-codes similar to that supplied by Statistics Denmark to DISCO-reporting firms, HBS can then identify the *6-digit* DISCO-code which corresponds to the extracted job title.[11] Thus, both the register data's and the job vacancy data's *6-digit* DISCO-codes are imputed from detailed job titles/positions.

### *A.3.2. Occupational and industrial distributions*

Lists of 1-digit occupations and 1-letter industries with their titles are included in Table 8 and Table 9. Figure 4 and Figure 5 shows the occupational and industrial distribution of workers in the matched and the full sample respectively. Notice the overrepresentation of the occupations "2 Professionals" and "5 Services and Sales Workers" as well as of the industries "O Public administration and defence; compulsory social security", "P Education", and "Q Human health and social work activities". These occupations are dominated by large groups of public employees, namely teachers, nurses and care assistant.

[9] For more details, see:
https://www.dst.dk/da/Indberet/oplysningssider/loenstatistik/stillingsbetegnelser-disco-08-i-loenstatistikken

[10] For more details, see:
https://www.dst.dk/Site/Dst/SingleFiles/hojkvalbilag.aspx?varid=107187&bilagid=183191

[11] For more details, see:
http://www.hbseconomics.dk/wp-content/uploads/2017/09/Eftersp%C3%B8rgslen-efter-sproglige-kompetencer.pdf

Table 8: 1-digit occupations

| 1-digit code | Occupational title |
|---|---|
| 0 | Armed Forces Occupations |
| 1 | Managers |
| 2 | Professionals |
| 3 | Technicians and Associate Professionals |
| 4 | Clerical Support Workers |
| 5 | Services and Sales Workers |
| 6 | Skilled Agricultural, Forestry and Fishery Workers |
| 7 | Craft and Related Trades Workers |
| 8 | Plant and Machine Operators and Assemblers |
| 9 | Elementary Occupations |

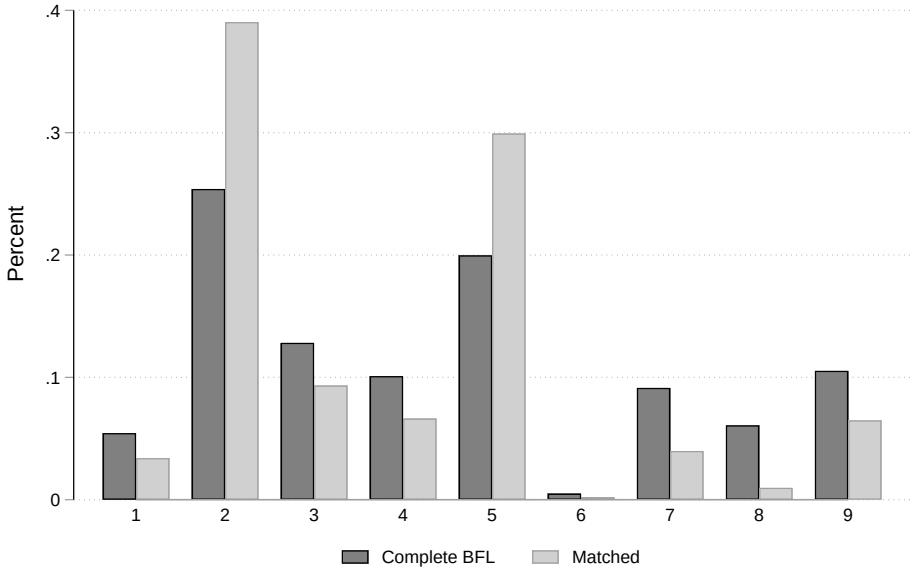
*Source: http://www.ilo.org/public/english/bureau/stat/isco/docs/structure08.docx*

Table 9: 1-letter industries

| 1-letter code | Industry title |
|---|---|
| A | Agriculture, forestry and fishing |
| B | Mining and quarrying |
| C | Manufacturing |
| D | Electricity, gas, steam and air conditioning supply |
| E | Water supply; sewerage; waste management and remediation activities |
| F | Construction |
| G | Wholesale and retail trade; repair of motor vehicles and motorcycles |
| H | Transporting and storage |
| I | Accommodation and food service activities |
| J | Information and communication |
| K | Financial and insurance activities |
| L | Real estate activities |
| M | Professional, scientific and technical activities |
| N | Administrative and support service activities |
| O | Public administration and defence; compulsory social security |
| P | Education |
| Q | Human health and social work activities |
| R | Arts, entertainment and recreation |
| S | Other services activities |

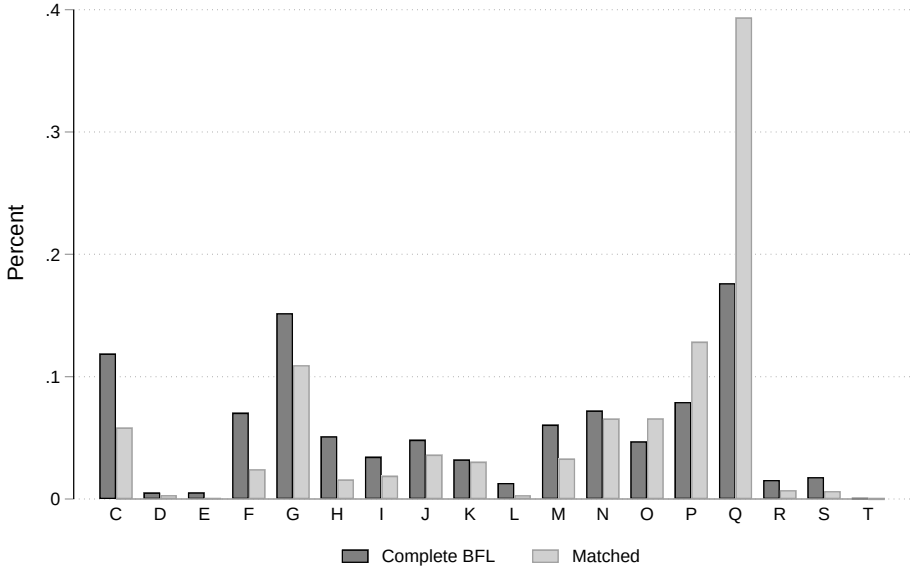
*Source: http://ec.europa.eu/competition/mergers/cases/index/nace_all.html*

Figure 4: Distribution of occupations in BFL and matched data




*Source: BFL, HBS-Jobindex 2010-2016.*
*Note: Observations weighted by full-time equivalents.*

Figure 5: Distribution of industries in BFL and matched data




*Source: BFL, HBS-Jobindex 2010-2016.*
*Note: Observations weighted by full-time equivalents.*

# References


Acemoglu, D., Autor, D. H., 2011. Skills, tasks and technologies: Implications for employment and earnings. In: *Handbook of Labor Economics*, Elsevier Inc., vol. 4, chap. 12, pp. 1043–1171.

Autor, D. H., Levy, F., Murnane, R. J., 2003. The Skill Content of Recent Technological Change: An Empirical Exploration. The Quarterly Journal of Economics 118, 1279–1333.

Bacolod, M. P., Blum, B. S., 2010. Two Sides of the Same Coin: U.S. "Residual" Inequality and the Gender Gap. The Journal of Human Resources 45, 197–242.

Beaudry, P., Lewis, E., 2014. Do male-female wage differentials reflect differences in the return to skill? Cross-city evidence from 1980-2000. American Economic Journal: Applied Economics 6, 178–194.

Black, S. E., Spitz-Oener, A., 2010. Explaining Women's Success: Technological Change and the Skill Content of Women's Work. The Review of Economics and Statistics 120, 187–194.

Blau, F. D., Kahn, L. M., 2017. The Gender Wage Gap: Extent, Trends, and Explanations. Journal of Economic Literature 55, 178–865.

Cameron, A. C., Miller, D. L., 2015. A Practitioner ' s Guide to Cluster- Robust Inference. Journal of Human Resources 50, 317–372.

Cerina, F., Moro, A., Rendall, M. P., 2017. The Role of Gender in Employment Polarization. Ssrn pp. 1–39.

Cortes, G. M., Jaimovich, N., Siu, H., 2018. The "End of Men" and Rise of Women in the High-Skilled Labor Market .

Deming, D., Kahn, L. B., 2018. Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals. Journal of Labor Economics 36, S337–S369.

Deming, D. J., 2017. The growing importance of social skills in the labor market. Quarterly Journal of Economics 132, 1593–1640.

Goldin, C., 2014. A grand gender convergence: Its last chapter. American Economic Review 104, 1091–1119.

Goos, M., Manning, A., Salomons, A., 2009. Job Polarization in Europe. American Economic Review 99, 59–63.

Goos, M., Manning, A., Salomons, A., 2014. Explaining Job Polarization: Routine-Biased Technological Change and Offshoring. American Economic Review 104, 2509–2526.

Hersch, J., 1998. Compensating Differentials for Gender-Specific Job Injury Risks. American Economic Review 88, 598–607.

Hershbein, B., Kahn, L. B., 2018. Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings. American Economic Review 108, 1737–1772.

Jarman, J., Blackburn, R. M., Racko, G., 2012. The Dimensions of Occupational Gender Segregation in Industrial Countries. Sociology 46, 1003–1019.

Kettemann, A., Mueller, A. I., Zweimüller, J., 2018. Vacancy Durations and Entry Wages: Evidence from Linked Vacancy-Employer- Employee Data. IZA Discussion Paper Series .

Kleven, H., Landais, C., Søgaard, J. E., 2018. Children and Gender Inequality: Evidence from Denmark.

Levanon, A., Grusky, D. B., 2016. The Persistence of Extreme Gender Segregation in the Twenty-first Century. American Journal of Sociology 122, 573–619.

Lindley, J., Machin, S., 2012. The Quest for More and More Education: Implications for Social Mobility. Fiscal Studies 33, 265–286.

Lindley, J. K., 2012. The gender dimension of technical change and the role of task inputs. Labour Economics 19, 516–526.

Lindley, J. K., 2016. Lousy pay with lousy conditions: The role of occupational desegregation in explaining the UK gender pay and work intensity gaps. Oxford Economic Papers 68, 152–173.

Modestino, A. S., Shoag, D., Ballance, J., 2016a. Downskilling: changes in employer skill requirements over the business cycle. Labour Economics 41, 333–347.

Modestino, A. S., Shoag, D., Ballance, J., 2016b. Upskilling: do employers demand greater skill when workers are plentiful? Working Paper pp. 1–46.

Olivetti, C., Petrongolo, B., 2014. Gender gaps across countries and skills: Demand, supply and the industry structure. Review of Economic Dynamics 17, 842–859.

Olivetti, C., Petrongolo, B., 2016. The Evolution of Gender Gaps in Industrialized Countries. Annual Review of Economics pp. 405–434.

Rendall, M., 2010. Brain Versus Brawn: The Realization of Women's Comparative Advantage. SSRN Electronic Journal pp. 1–26.

Weinberger, C. J., 2014. The Increasing Complementarity between Cognitive and Social Skills. The Review of Economics and Statistics 96, 849–861.

Yamaguchi, S., 2018. Changes in Returns to Task-Specific Skills and Gender Wage Gap. Journal of Human Resources 53, 32–70.