Online Appendix

Families' Career Investments and Firms' Promotion Decisions

Frederik Almar, Benjamin Friedrich, Ana Reynoso, Bastian Schulz and Rune Vejlin

Appendix OA Data

This Appendix provides a comprehensive overview of the data sources used in our paper and it documents how we process them. Section OA.1 provides an overview of the data sources we use. We provide details on sample selection in Section OA.2. The key variables are discussed in Section OA.3.

OA.1 Data Sources

We use administrative data provided by Statistics Denmark, which records information for the universe of Danish residents. We use a selection of registers from this comprehensive source.

Our starting point is to create a *baseline sample* with yearly observations for the entire Danish population in the years 1980–2018 aged 19 to 55. This data set is created from two registers provided by Statistics Denmark: the register PERSONER covers the years 1980–1984, and the register BEF covers the remaining years from 1985–2018. Both registers cover all individuals living in Denmark at the end of each calendar year.[1] From these sources, we obtain basic individual characteristics such as age, gender, identifiers for parents, identifiers for cohabiting partners and spouses, identifiers for children, municipality of residence, and country of origin.

We add educational variables from the registers UDDA and VEUV. UDDA is a panel data set with yearly observations.[2] This gives us detailed information on completed educational programs across all levels and fields and the date of graduation. VEUV contains information on postgraduate courses/continuing education, which we use to characterize some forms of management training, see Section OA.3.

Next, we sequentially add labor market variables from a range of different registers provided by Statistics Denmark. First, we add variables on labor market income from either employment or self-employment from the register IND. Second, we add hourly wages, hours worked, and employer- and industry identifiers from the register IDAN. IDAN contains information on multiple employment relationships of different types undertaken by an individual during a year. Individuals may have more than one employment record per year. In our case, we want to identify the most "relevant" employment relationship for each individual. To do this, we rank the different types of employment and keep the variables from the highest-ranked employment relationship. The ranking is as follows:

[1] In the data provided by Statistics Denmark, all variables in PERSONER and BEF are measured at the start of the calendar year. Because most other data sets refer to the end of the calendar year or the second half of the year, we adjust these data sets by lagging the year variable by one. This ensures that all variables are measured as close in time to each other as possible.

[2] From 1980–2007, the information is measured by the end of the year, while from 2008 and forward it is measured in October.

1. Main job in November
2. Most important non-November job
3. Self-employed worker with employees in November
4. Self-employed workers without employees in November
5. Helping spouse,

where 1 is the highest rank. We drop employment observations (but not individuals) that do not belong to any of these five ranked categories. Accordingly, the individuals are recorded as non-employed (zero labor supply).

Third, we add variables on labor supply and the accumulation of human capital. We extract the yearly labor supply status in the main job from the register RAS. This is used directly in the model estimation, see more details on the variables in Section OA.3. Furthermore, we extract labor market experience. We use it to deflate wages properly, taking into account the aging population and, relatedly, increasing productivity and wages due to more experienced workers. To measure experience, we combine the registers EXPYEAR and IDAP.[3] We accumulate all individual work experience gained between 1964–1979 from EXPYEAR.[4] This information is then merged with individuals observed in IDAP in 1980 to ensure that we include their work experience prior to 1980. Next, we accumulate experience up to 2018 to get the accumulated work experience for each individual at the end of a particular year. We correct for breaks in experience spells by writing experience forward.[5]

Fourth, to identify the occupation in which an individual is employed, we use the register AKM. The 6-digit occupation variable only covers the 1991–2018 period. After fixing some missing data problems in the original variable,[6] we create a new aggregate variable containing the first 3 digits of the 6-digit code, resulting in 149 occupational categories. Moreover, we fix the break in 2009–2010 for the 3-digit code by mapping forward.[7]

Finally, we use the dataset AKU, which contains the Danish part of the European Labor Force Survey (LFS). AKU is a comprehensive survey covering the 2000–2018 period providing detailed responses on hours worked and other characteristics of the respondents' working conditions. The survey respondents can be linked to the administrative registers. We use AKU to cross-check the

[3]Labor supply variables from Statistics Denmark prior to 2008 are based on data from a mandatory Danish pension fund (ATP). In the ATP data, full-time work corresponds to around 30 hours or more per week, i.e., we cannot distinguish between an individual working 35 or 50 hours per week prior to 2008. After 2008, we have contractual work hours.

[4]EXPYEAR builds on mandatory pension payments and transforms the pension payments into hours worked using an algorithm developed by Statistics Denmark.

[5]These breaks occur when individuals are missing in the data for some years, which is usually due to stays abroad. Without this correction, their accumulated experience would be set to zero upon their return.

[6]One of the original variables already covers the entire 1991–2018 span, but we have detected some cases with missing coverage. We fix this as much as possible by using two other occupation variables covering the 1993–2009 and 2010–2013 time spans, respectively (reflecting a break in the nomenclature between 2009–2010). In cases where the 1991–2018 variable is missing, we substitute for one of the other variables, depending on the particular year, if these have non-missing observations instead. We remove duplicates to ensure that we have one observation per individual per year.

[7]This means that we compare codes of individuals just before and after the break working in the same establishment, and then we use the most prevalent changes in codes to update the pre-2010 codes. For example, if a group of individuals has the occupation code '123' in 2009, and a majority of them have changed to '321' in 2010, then we map forward by changing all pre-2010 '123' codes to '321'.

variables on labor supply from RAS and to create additional moments for an especially demanding labor supply status, super-full-time work, which we define in Section OA.3.6.

OA.2 Samples

We start from the *baseline sample* described above. It consists of the entire population of Danish residents in the age range of 19 to 55 in the period 1980–2018.

For each individual, we keep all observations from the first time they appear in the baseline sample until 2018 and select the cohort who graduated from their highest educational program in the years 1991–1995. We call this our *inflow sample*. To follow the career of these individuals over time, we apply the following steps to get from *baseline sample* to *inflow sample*:

1. We use the register UDDA to compute the *highest education* achieved by individuals over the period they are observed. Moreover, we use the register IND to compute whether an individual is ever self-employed.
2. We create an *interim sample* by selecting those individuals who achieved their highest education between 1991 and 2008 and who have never been self-employed. This interim sample is used solely to estimate the ambition types and career ladders, see Section OA.3.
3. We use the registers PERSONER and BEF to match each individual to the identity of their *decisive domestic partner*, see Section OA.3.2 for the definition. 46% of individuals in the interim sample are matched to such a partner. 40% of these individuals are married to a decisive domestic partner who is in the baseline sample but did not graduate between 1991 and 2008. In this case, we reintroduce the observations of these partners.
4. Finally, we exclude individuals who graduated between 1996 and 2008 and their partners. We also exclude individuals who are neither married according to our definition of partnership nor single. We further exclude individuals who had their first child before graduating from their highest education and who cannot be assigned an ambition type, career ladder, training- or manager status. The remaining sample is our *inflow sample*. Taken together, the inflow sample contains individuals who graduated between 1991 and 1995 and their identified partners. 60% of individuals who graduated between 1991 and 1995 are matched to a partner of whom 74% are in the baseline sample but did not graduate between 1991 and 1995.

Because we use administrative population records, attrition is very infrequent: 84% of individuals are observed for at least 20 years. Our inflow sample is an unbalanced panel of 152,390 individuals who achieved their highest education between 1991 and 1995 and their partners. The reason for starting in 1991 is that we do not observe occupation before 1991—a key variable in our analysis. Moreover, we decide to end our sample in 1995 to focus our analysis on a single cohort of individuals who we observe for at least 24 years. 83% of individuals in our sample are born between 1963 and 1976. In total, our data consists of 52,231 couples, 23,024 single women, and 24,905 single men.

OA.3 Key variables

OA.3.1 Educational programs

We assign individuals the highest educational program (excluding programs related to on-the-job training, see below) achieved by the age of 35. We define an educational program as the four-digit education code (variable HFAUDD from the UDDA register). There are 1,108 unique codes in the inflow sample. In the case of compulsory schooling, we further divide by region of graduation to capture variation in the value of a compulsory education across rural and urban areas. A program within secondary education would be carpenter or care worker (both vocational training programs). At the bachelor level, examples of programs include teacher and nurse. Finally, at the graduate level, business with focus on marketing is an example of a program.

OA.3.2 Partnership and marital status

We start from all individuals in the *baseline sample* to identify each individual's *decisive domestic partner*. We consider both legally married and cohabiting couples.[8] The idea is to identify the person with whom an individual makes joint decisions about fertility and career investments, which are the key choices in our model. To operationalize this idea in the data, we consider partners who are attached to the individual for at least five consecutive years, with the spell starting when the wife is 28 years old or younger and the husband 32 years old or younger. In cases where multiple partners meet these criteria, we keep the partner from the longest-lasting relationship. The asymmetric age thresholds for men and women are chosen to balance the single shares by sex. Age 28 for women is set equal to the age threshold between periods t_1 and t_2 in the model (see Appendix D.1), while age 32 for men is chosen by targeting the single share of women who got together with their decisive domestic partner partner at age 28 or earlier. These different age thresholds reflect that women, on average, get married at an earlier age and to a slightly older partner.

Our definition of marriage is having a *decisive domestic partner*. The non-married group in the data consists of two groups of individuals: 1) some who will not make joint decisions about fertility and career investments with a partner, see our definition above, and 2) some who would be classified as married had the age thresholds been slightly higher. Through the lens of our model, this second group of individuals who marry "too late" can be viewed as neither married nor single. Thus, we define a four-year "buffer zone" of couples who are not classified as married but would have been classified as married had the age thresholds been 32 for women and 36 for men. Subsequently, we classify individuals as *single* if they are neither married nor in the buffer zone. In our inflow sample, 69% of individuals have a decisive domestic partner. For comparison, in the baseline sample 64% of individuals who are married by the age 30 have a partner for at least 5 years. By the age of 50, this share is 81%.[9]

[8] Cohabiting couples are defined by Statistics Denmark as two opposite-sex individuals who share the same address, exhibit an age difference of less than 15 years, have no family relationship, and do not share housing with adults other than their partner. Our data do not allow us to identify cohabiting same-sex couples.

[9] Here we only include individuals who turn 19 between 1980-1995 to make the numbers more comparable to the

OA.3.3 Hourly wages, wage growth, and starting wages

To define our ambition types and career ladders, we have to define starting wages and wage growth. We base these calculations on the (larger) *interim sample* to get precise estimates of starting wages and wage growth at the levels of educational programs and occupation-firm combinations.

Our starting point is the *hourly wage* as provided by Statistics Denmark in its narrow definition, which excludes benefits and various contributions. This narrow definition is most reflective of marginal productivity. We deflate hourly wages by running a regression of log wages on year dummies with 2000 as the base year. We further control for differences in wage-experience profiles by education by including interactions of educational levels and accumulated labor market experience, see Section OA.1. This ensures that log hourly wages are comparable over time even as the education and experience composition of the sample changes. We then subtract the year dummies from the hourly wage, thereby constructing an hourly wage measure that controls for wage inflation and aggregate changes in education and experience.

We define individual *wage growth* in our *interim sample* as the difference between an individual's average hourly wage in years 1 to 5 in the sample, which we define as their *starting wage*, and their average hourly wage in years 9 to 11 in the sample. We focus on wage growth in the early career because wage growth is highest during that period and, thus, most indicative of human capital investments. Whenever we aggregate individual wage growth, e.g., at the educational program or occupation-firm level, we trim the top 1% of individual wage growth to exclude outliers.

OA.3.4 Ambition types

First, we aggregate the individual observations of starting wages and wage growth in the interim sample to the average at the educational program level. Second, we standardize the educational program averages of starting wages and wage growth. Third, we use k-means clustering, with the standardized variables as inputs, to assign educational programs (and thereby their graduates) to four clusters, which we label *ambition types*. Intuitively, the k-means clustering algorithm (Steinley, 2006) minimizes the within-cluster variation in standardized averages of starting wages and wage growth across the four categories. In other words, each of the four categories is internally homogeneous in terms of the starting wages and wage growth obtained by the graduates of the educational programs within that category. In Almar et al. (2024) we study marital sorting based on ambition and how it relates to changes in household inequality.

OA.3.5 Career Ladders

To define career ladders in the data, we first aggregate the individual observations of wage growth in the interim sample to the average of coworkers who begin their careers in the same occupation-firm combination. Second, we define occupation-firm combinations that exhibit hourly wage growth at or above the 80th percentile as *steep ladders* and *flat ladders* otherwise.

inflow sample.

We assign individuals to their most frequent occupation-firm combination over the first five years in the interim sample. We condition the cell size of these combinations to consist of at least five coworkers to smooth out individual contributions. The flip side of this restriction is that firm-occupation combinations defined by very fine occupational codes might not include at least five coworkers. Thus, we proceed in five iterations starting with finely defined occupational codes and then gradually moving to coarser versions. If an individual is not assigned to an occupation-firm combination in one iteration, we try to assign them in the next iteration. First, we assign occupation-firm combinations based on three-digit occupational codes and firm ID. Second, we use two-digit occupational codes and firm ID. Third, we use detailed four-digit educational program codes and firm ID. Fourth, we use a coarse educational level code (with four levels) and firm ID. Fifth, we only use firm ID.

The assignment of either steep or flat ladders has so far concerned the first period t_1 in the model (see Section D.1). For the subsequent periods t_2 and t_3, the assignment of occupation-firm combinations to ladders is based on the initial t_1 assignments, i.e., an occupation-firm combination that is defined as steep in t_1 is similarly defined as steep in t_2 and t_3. In cases where an individual has both flat and steep ladder positions within periods t_2 and t_3, we use the most prevalent type to characterize the entire period.

20% of women and 27% of men sort into the steep ladder in t_1. There is significant variation across workers within ambition types (Appendix OA.3.10). For example, approximately 25% and 50% of individuals of ambition type θ_1 and θ_4, respectively, sort into the steep ladder (Figure OA.1). Our interpretation is that a worker's ambition type reflects the expected career path of the worker based on the educational degree from an ex ante perspective. In contrast, being on a step or a flat ladder is a choice (made jointly with the spouse if married according to our definition) that is part of a particular career-life balance plan. For example, a law graduate can be on a steep career ladder at a private law firm or a flat ladder in the public sector.

OA.3.6 Labor supply

We use the RAS register and the AKU survey dataset to construct four labor supply states: non-participation, part-time, full-time, and super-full-time work. The three former states are directly determined by the part-time/full-time variable available in RAS. Non-participation refers to not being employed, e.g., not having a highest-ranked employment relationship, by the end of November in a given year. Those who have a highest-ranked employment relationship in a given year are characterized as either part-time or full-time employed depending on the hours worked per week. The threshold between part-time and full-time is 30 hours (1980-1992), 27 hours (1993-2007), or 32 hours (2008-2018) (Lund and Vejlin, 2016).[10]

We assign the mode of non-participation, part-time, and full-time within a period (see Section D.1) as the labor supply status of the corresponding period. In cases where more than one mode exists, we always assign part-time as the labor supply status for this given period.

[10]Please see Section OA.1 for a detailed description of how weekly working hours are measured.

Prior to 2008, we do not observe contracted hours in RAS, i.e., we do not observe to what extent individuals work above the part-time/full-time threshold, e.g., 35 hours or 50 hours. However, with detailed information on hours worked and time worked during a week from AKU, we can identify surveyed individuals working *super-full-time*. Hence, this labor supply status is only observed for a subset of those working full-time. For comparability, we apply the same definition based on AKU throughout our sample period, e.g., both prior to and after 2008. We define super-full-time relative to the Danish standard full-time working week corresponding to 37 hours. Hence, a surveyed individual works super-full-time in two cases: 1) the individual reports that a usual working week is 38 hours or more; 2) the individual reports that a usual working week is 37 hours, and they either report sometimes working in the evening, at home, on Saturdays, on Sundays, at night, or sometimes working overtime.[11]

OA.3.7 Managers and promotions

We consider workers to be promoted if they are observed holding a managerial occupational code for at least two consecutive years. That is, the first digit in the (D)ISCO code as provided by Statistics Denmark has to be equal to 1 for two consecutive years.

OA.3.8 On-the-job management training

Our strategy to measure whether a worker has received on-the-job management training is based on two data sources: (i) information on continuing education programs—MBA degrees and, specific to Denmark, HD degrees[12]—from the education register VEUV; (ii) information on worker transitions between different occupational codes, which we observe in the AKM register, see Section OA.1.

The idea is that firms can use both external education programs (HD, MBA) and specific roles within the firm (reflected in occupational codes) to train workers and prepare them for managerial positions. Our data show a clear negative relationship between firm size and the likelihood of workers completing external continuing education degrees. This suggests that larger firms tend to train their workers internally, which we seek to identify by tracking workers' progression through different occupational codes before being promoted into management (see, e.g., Frederiksen and Kato, 2017).

We combine information on external education programs and internal management training to predict the probability of becoming a manager.

Let ET_i be a dummy for whether a worker i has completed such an external training program. Let IT_{it}^k be a dummy for whether worker i is observed with a specific non-managerial (2-digit) occupational code k during period t. These periods refer to career stages in line with our model (e.g., early career), see Sections 3.6 and D.1.

[11] Importantly, we disregard shift workers who report working in the evening, at night, on Saturdays, or on Sundays as these are standard working conditions for shift workers and not a sign of extraordinarily high labor supply.

[12] HD stands for "Handelshøjskolens Diplomuddannelse", i.e., business school graduate education. HD programs are flexible, business economics diploma programs for employed individuals seeking management training. HD programs are offered both as individual courses and as a complete education program at two levels. HD1: Basic education in business economics. HD2: Specialized education dedicated to a specific subject area within business economics. HD1 and HD2 together are typically a four-year part-time educational program that corresponds to a bachelor's level education.

To capture how both external and internal on-the-job management training contribute to the probability of being promoted into management, we estimate the following binary response model. The outcome mg_i is a dummy for whether individual i is ever observed as a manager:

$$P(mg_i = 1) = G\left(\alpha + \sum_k \sum_t \alpha_{kt}^{IT} IT_{it}^k + \alpha^{ET} ET_i\right) \tag{2}$$

where α is a constant, α_{kt}^{IT} is the effect of a specific occupation k in career stage t, α^{ET} is the effect of having completed external management training, and G is the logistic CDF. Estimating this model reveals that external management training is quantitatively the most important predictor of a promotion into management because α^{ET} is an order of magnitude larger than the positive and significant occupation-specific effects α_{kt}^{IT}.

We use a "receiver operating characteristic" (ROC) curve to assess the predictive power of model (2). The ROC curve is a graphical tool that illustrates the performance of a binary classifier model at varying threshold values. At a probability threshold of 0.05, our estimated model correctly classifies 85.98% of individuals in our data (managers with training and non-managers without training). 50.80% of managers previously received training.

Finally, to construct the training variable, we select the occupation $\times$ career stage interactions IT_{it}^k that are positive and statistically significant predictors of becoming a manager. Moreover, we include ET_i, the dummy for having completed a continuing education managerial training program.

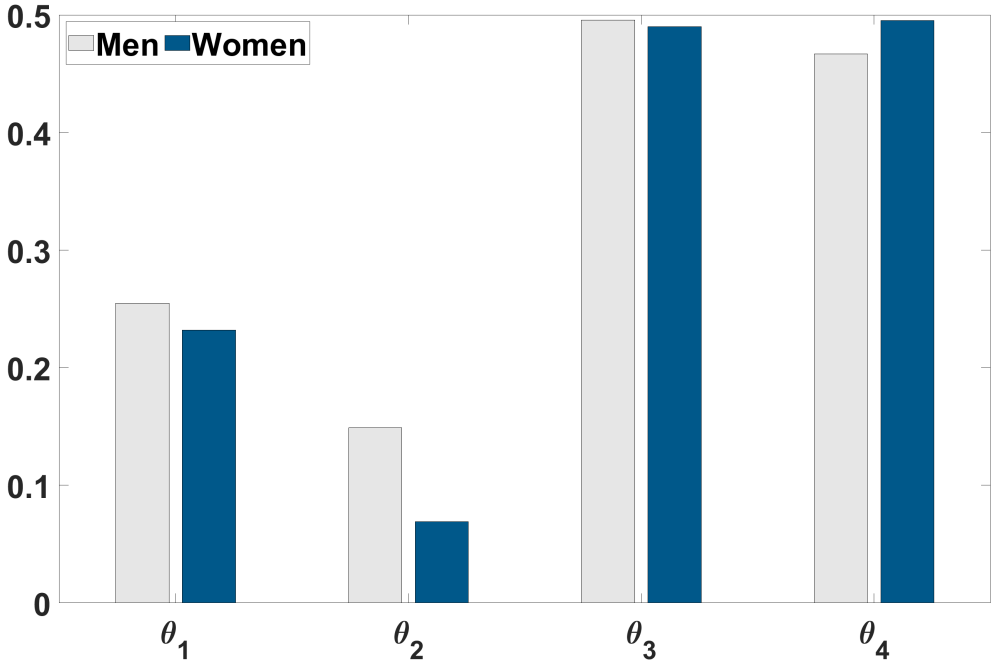
OA.3.9 Annual earnings

To construct moments of earnings, we use the annual taxable salary variable from the IND register. We first deflate earnings by subtracting the year effects described above for hourly wages. Then, we assign earnings by period in the model (see Section D.1) as the by-period mean of deflated earnings.

OA.3.10 Descriptive statistics

Table OA.1 provides a summary of descriptive statistics for the inflow sample, which is the sample we use for estimating the model. The total number of observations is 3,491,849 covering the years 1991 to 2018, yielding an average of 124,709 observations per year. The average birth year is 1970 and the average share of females is 48% across years. 90% of the sample consists of native Danes. The most common level of education across ages and years (not the highest degree ever achieved) is secondary education (54%) followed by an equal share with primary education or a bachelor's degree (19%). Master's and Ph.D. degrees make up the lowest share (8%). In Panel B, we show variables relevant to the marriage market part of the model. 71% are married, where married means married with a decisive domestic partner according to our definition above. The average age of marriage is less than 25 years. Recall that our definition of marriage includes both legal marriage and cohabitation. 63% of individuals have children and the average age of having the first child is just above 29 years. Panel C on labor market outcomes shows sizable gender gaps. Men generally work more hours in the

Figure OA.1: Fraction of men and women in the steep ladder by ambition type



Notes: θ refers to the *ambition type* defined and constructed as explained in Section 2.2: $\theta_1 = (low, low)$, $\theta_2 = (high, low)$, $\theta_3 = (low, high)$, and $\theta_4 = (high, high)$.

labor market, are more likely to be on a steep career ladder, and are more likely to receive on-the-job managerial training and promotions.

Moreover, Figure OA.1 shows that the fraction of men and women in the steep ladder varies by ambition type. Specifically, individuals or higher ambition are more likely but sort into more demanding career paths, but not all graduates from the same types of programs select into the same type of ladder. For example, only about 50% of individuals of the highest ambition types θ_3 and θ_4 sort into the steep ladder.

Table OA.1: Descriptive Statistics for the inflow sample

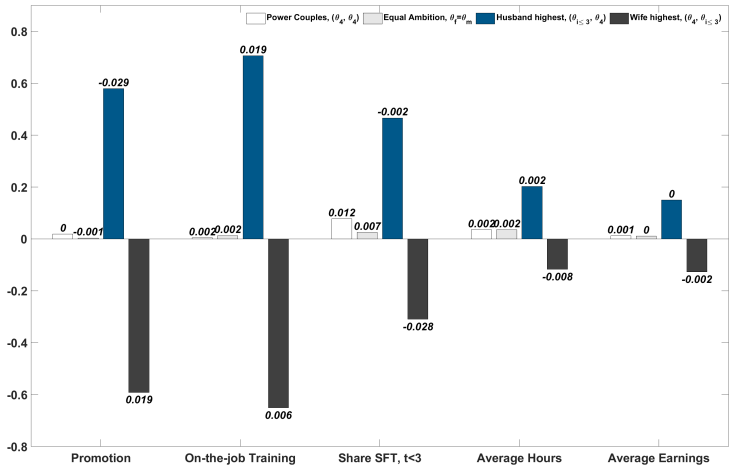
Sample	Inflow	
Total observations	3,491,849	
Observations per year	124,709	
	Mean	SD
Year	2,005	8.23
Panel A: Demographics		
Birth year	1,970	4.81
Sex ratio	0.48	0.03
Immigrant status	0.10	0.30
Primary school	0.19	0.39
Secondary school	0.54	0.50
Bachelor	0.19	0.39
Master & Ph.D.	0.08	0.27
Panel B: Marriage market and family		
Married	0.71	0.45
Age at marriage	24.62	3.35
Has children	0.63	0.48
Age at first child	29.27	4.77
Panel C: Labor market outcomes		
Women		
Non-participation	0.15	0.36
Part-time work	0.10	0.30
Full-time work (incl. super full-time)	0.74	0.44
Initial ladder steep	0.20	0.40
On-the-job training	0.16	0.37
Promotion	0.04	0.21
Men		
Non-participation	0.13	0.33
Part-time work	0.05	0.22
Full-time work (incl. super full-time)	0.82	0.39
Initial ladder steep	0.27	0.44
On-the-job training	0.25	0.43
Promotion	0.10	0.30

Notes: based on the inflow sample (see OA.2). The sex ratio denotes the number of women to men in a given year. The variable immigrant status takes on value 1 if an individual is either an immigrant or child of an immigrant and 0 otherwise. Marriage is defined as in OA.3 (here we pool all observations across years). All variables in panel C are defined in OA.3.

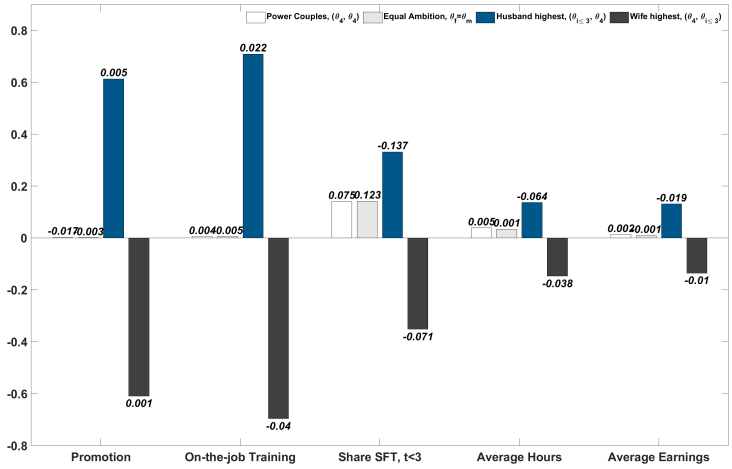
Appendix OB Supporting figures for alternative specifications

Figure OA.2: Gender gaps in model with alternative specifications by couple type

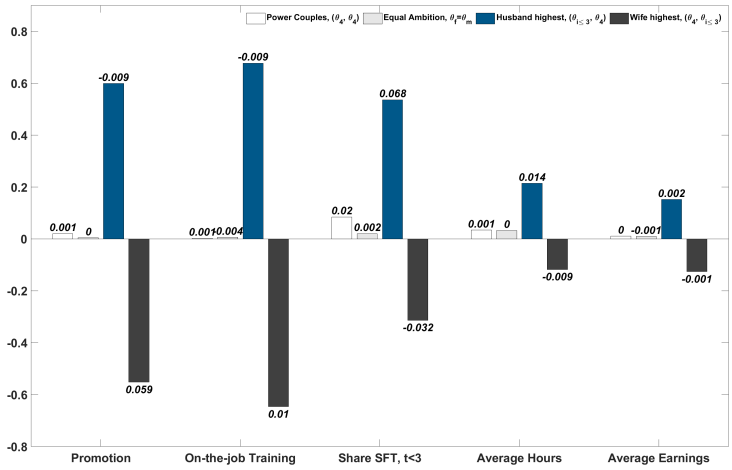
Panel A: Full information



Panel B: $\kappa = 1$



Panel C: History based



Note: numbers in italics represent the change relative to baseline.