Does White Collar Exemption Work? Evidence from A Hybrid Matching Estimator*

Takuya Hasebe
Sophia University

Yoshifumi Konishi
University of Tsukuba

Shunsuke Managi
School of Engineering, Kyushu University

Abstract: White collar exemption (WCE) exempts workers from overtime pay (OP) and work-hour schedules (WS). We estimate the causal effects of WCE on full-time employees in Japan, distinguishing the two exemptions based on the theory of compensating wage differentials. We combine coarsened exact matching with subclassification on propensity scores to optimize on the trade-offs between unbiasedness versus precision of estimates. We find that the WS exemption is shown to decrease wages and increase work hours, yet paradoxically decrease work satisfaction. There is some indication that in Japan, WCE has not lead to discretion at work, which may partly explain this puzzle.

JEL Codes: J33, J38, K31
Key Words: White collar exemption, Overtime pay, Work-hour schedules, Matching, Subclassification

*Our special thanks goes to Shin Hong Joo, who provided valuable assistance in preparing the dataset for this study. The study was in part supported by the fiancial support from Japan Society for the Promotion of Science (JSPS), Grant-in-aid for Specially Promoted Research (*Grant number: 26000001B*).

1. Introduction

Overtime pay (OP) regulation and white collar exemption (WCE) are two commonly used labor-market instruments designed to hit a balance between protecting employment and health and providing incentives for motivation and performance. The OP regulation requires employers to monitor work hours and pay premiums for overtime hours. WCE, on the other hand, allows employers to exempt certain types of non-manual office workers from the OP regulation. WCE is often subject to a heated political debate not only because it can potentially affect a number of firms and workers, but also because there is a great deal of uncertainty regarding its economic consequences.[1]

By design, WCE is intended to influence two types of economic margins. First, WCE exempts eligible workers from the OP premium. Hence, the exempt workers and the non-exempt workers face different price incentives for their labor supply decisions, which in turn also affects employers' demand for the two types of labor. Therefore, WCE may induce changes in equilibrium wages or labor hours or both for the exempt workers, purely through the price mechanism. Second, the exempt status also comes with the work arrangement that allows eligible workers to work flexibly without monitoring and with more discretion as to how and when to work to complete assigned tasks. This flexibility by itself is thought to increase labor productivity, at least for certain types of jobs that involve non-manual, non-repetitive tasks. Thus, in the U.S., the largest group of employees covered under the WCE are administrative, executive, and professional workers (Trejo, 1991).

An ample body of literature empirically investigated the impacts of labor-market regulations [e.g., Costa (2000), Hamermesh and Trejo (2000), and Trejo (1991, 2003)]. However, economic research to date has primarily focused on the effect of OP regulation, but largely ignoring the effect of work-hour schedule (WS) regulation despite its economic significance. Trejo (1991), for example, examines the effect of the OP regulation on hours of work and wages. But his study focuses on the sample of workers who are paid on an hourly basis. By definition, these workers are subject to monitoring of work hours, and hence, do not enjoy the benefits of flexible work arrangements. Our paper attempts to disentangle these two effects of WCE.

To conceptualize the economic impacts of WCE, we build upon Rosen's theory of equalizing wage differentials (1974, 1985, 1986). On one hand, for workers who value flexibility

[1]For example, the U.S. Department of Labor (DOL) updated the eligibility rules for the OP exempt status in May, 2016 and increased the salary threshold from the current $455 per week to $913 per week. In preparing for this increase, the DOL conducted an impact study, which concluded that "4.2 million workers will be directly affected by the rule, and 8.9 million currently overtime eligible workers will get strengthened overtime protections (p.2)... the rule will result in an average annual increase in pay to workers of $1.2 billion per year P.4)" (U.S. DOL, 2016).

at work, the WS-exempt status should be taken as an added 'amenity' to their jobs. On the other hand, the WS-exempt status is costly for firms, at least in the short run, because they must incur administrative and legal costs for institutionalizing the status. Hence, the equilibrium wages must be lower for the WS-exempt status than the WS-regulated status, holding the OP-exempt status. The effect on work hours is unclear, however. On one hand, the WS-exemption may allow workers to be more productive, which may reduce work hours. On the other hand, different worker types self-select into different work arrangements. Workers who value flexibility at work may be those who like to work longer hours. Which effect is stronger is a priori uncertain. Rosen's theory also provides an important prediction for work satisfaction: WS-exempt workers must have higher satisfaction from work because they must have willingly chosen the WS-exempt status despite its lower wage. This study attempts to test these economic predictions empirically, exploiting a unique regulatory setup in the Japanese labor market.

The Japanese labor market is known for its rigidity in personnel management: i.e., strict work-hour control, seniority-based salary system, and pre-scheduled promotion and tenure scheme. The rigidity is often criticized as one of the primary causes of two unfortunate characteristics of the full-time employees in Japan: i.e., long working hours and low earnings per hour of labor input. In response to rising concerns with this rigidity, the government made amendments to the LSA in 1988, effectively creating an analogue of WCE status to the OP regulation. In the Japanese regulatory context, this WCE analogue is called the *discretionary work arrangement* (DWA). What makes the Japanese labor market suitable for our analysis is that there is another OP-exempt arrangement, defined more informally as a traditional labor arrangement, in addition to the official DWA: i.e., An employer appoints an employee to an OP-exempt position that would involve supervisory tasks such as managing a large number of subordinates. We call this type of traditional labor arrangement the *supervisory work arrangement* (SWA). Employees under the SWA typically hold no discretionary power over their work hours or workload, and hence, they often do not meet the DWA eligibility. Nonetheless, Japanese firms often form this informal labor contract with their employees without approval from the labor office, and exempt themselves from paying the overtime premiums. Consequently, we have three labor arrangements: regular OP workers, informally defined non-OP workers without merit of DWA, and official OP-exempt workers with full merit of DWA. This allows us to evaluate the potential outcomes of these alternative arrangements to disentangle the two effects of WCE mentioned above.

The empirical challenge here is that our treatment, assignment to an alternative work arrangement, is clearly endogenous: i.e., workers self-select into different work arrangements. Added to this challenge is that workers rarely change their work arrangements, and when

they do, many other aspects of jobs also change concurrently. This makes conventional quasi-experimental methods highly unreliable. We, instead, employ a matching estimator, exploiting the richness of pre-treatment covariates in our dataset. This dataset comes from a large nation-wide internet survey we conducted in 2016. The survey offers two advantages over other labor-market surveys in Japan. First, it contains rich information on worker's occupational characteristics that are not available in the conventional surveys: i.e., work arrangements, occupational rank, the number of subordinates, department size (in addition to firm size), and on-the-job skill requirements. Second, its sample size is large enough to yield a sufficiently large number of matched individuals who work under different work arrangements yet work on identical jobs.

This richness of our pre-treatment covariates, however, poses the curse-of-dimensionality problem in matching. Even with our relatively large sample size, exact matching leads to no matches. One way to overcome this problem is to employ some dimension-reduction approaches such as propensity score and Mahalanobis matching. Recent advances in the matching literature have shown, however, that these conventional approaches often fail in balancing covariate distributions (Iacus *et al.*, 2011). This was indeed the case in our context. As a better alternative, we devise a hybrid method, which combines coarsened exact matching (CEM) method of Iacus *et al.* (2011, 2012), with the subclassification estimator based on propensity scores (Imbens and Rubin, 2015). The method starts by estimating the propensity scores using all pre-treatment covariates as regressors, and then creates subclasses (or blocks) by stratifying on the estimated propensity scores as well as the key covariates that we deem necessary to meet the uncounfoundedness of assignment. The average treatment effect (ATE) is then estimated as the weighted average of differences in the observed outcomes of matches over these subclasses. The hybrid method is designed to help us optimize on the essential trade-off between balancing covariate distributions for unbiasedness of estimates versus having large enough sample for precision of estimates. In Section 6, we demonstrate the success of this hybrid matching method. As we vary the set of pre-treatment covariates for use in (coarsened) exact matching on the pre-trimmed subsample (based on estimated propensity scores), the estimates of ATEs start to change substantially from OLS and PSM estimates. Yet, the signs of the estimates tend to move unequivocally toward the same direction, with larger standard errors due to smaller sample sizes. The matched sample (in our preferred specification) consists of workers who conduct the same types of non-manual, non-repetitive tasks with similar probabilities to be assigned to either DWA or SWA. Hence, these final matches well represent the target population of interest (see Section 4 for more discussion on this point).

Our results are, in some outcomes, consistent with Rosen's theory, but in others, con-

tradictory to the theory. First, we find that hourly wage is lower for an arrangement that legally permits flexible work scheduling (DWA) than an arrangement that does not (SWA). Hence, this finding is consistent with the idea that workers value flexible work scheduling. Second, work hours are estimated to be higher with DWA than with SWA. This finding is suggestive of little or no gain in terms of efficiency of time use, at least in the short run. This, however, should not be necessarily taken as the evidence of little or no gain in labor productivity. For the latter, we need data on direct measures of output at work. Third, we find that flexible work scheduling is associated with the lower level of work satisfaction and the higher frequency of daily stress than inflexible work scheduling. Although the estimates on work satisfaction are not statistically significant, the signs of the estimates seem robust to different specifications. To explore the potential reasons for this puzzle, we estimate the effects of DWA on hours of meeting (per week) and the percentage of unfruitful meetings. We find there is no significant differences in either of these variables between DWA and SWA. Hence, one plausible explanation for these mixed results is that workers generally value flexible work scheduling, and hence, willingly accept lower wages and longer work hours, yet DWA does not come with the full merit of flexibility and discretion at work. As a result, DWA workers feel more stress and less satisfaction from work. This may also partly explain why work hours are longer with DWA than with SWA. All taken together, our results suggest a potential for large gain from aligning the interests of firm and worker — *allowing DWA workers to enjoy more flexibility and discretion at work* may improve worker's productivity and mental health at the same time.

2. Institutional Background

All employees in Japan, either full-time or part-time, are subject to the Labor Standards Act (LSA) of Japan. Much like the Federal Labor Standards Act in the U.S., the LSA imposes two kinds of wage regulation on the Japanese labor market: a minimum wage and an overtime pay (OP). In principle, the OP regulation requires an overtime wage premium for work hours exceeding 40 hours per week or 8 hours per day. The wage premium is typically 25% of the regular-hour wage rate (though it can be increased up to 50% for work hours performed on pre-scheduled holidays). To comply with the OP regulation, employers must set starting time and ending time of regular hours, and record working hours of their employees for each day.

This rigidity in applying the OP regulation created highly inflexible work-hour management practices among the Japanese firms. In response to rising concerns with this rigidity, the government made amendments to the LSA in 1988, effectively creating two forms of

exception to the OP regulation. The first was to allow for more flexible implementation of the OP regulation. Under the amendments, employers are allowed to average out work hours over a certain period (other than per day/week), and be exempt from paying the overtime premium as long as the averaged work hours do not exceed the legal limit of 40 hours per week (the system is known as the *variable hour arrangement*). Furthermore, under the amendments, employees can choose their own starting time and ending time, deviating from regular hours set by employers (known as the *flexible time arrangement*). Both arrangements are still OP-regulated, yet require a formal agreement with employees.

The second form of exception, which is the focus of this study, was to allow several types of professions to be exempt from the OP regulation. Initially, two types of professions were made exempt from the OP regulation. The first type is those who regularly engage in work outside business offices such as sales personnel and travel attendants (referred to as *off-site work* under the LSA). The second type is those conducting professional or expert services such as researchers, product developers, IT system engineers, and fashion designers (referred to as *professional work* under the LSA). In 1997, the list was expanded to include other types of professionals such as lawyers and accountants. In 2000, further amendments were made to the LSA, and those engaging in *management-related work* were also made exempt from the OP regulation. For either type of profession, an important pre-condition for the OP-exemption is that employees hold flexibility and discretion as to their own work schedules, work hours, and workload.

With the exempt status, employees are deemed to have worked for agreed-upon work hours (often, the legal limit of 40 hours per week) regardless of their actual work hours, and employers are also exempt from monitoring their employees' work hours and paying for OP compensations (they still need to pay the compensation for work hours during night or on holidays).[2] One important aspect of the regulation is that the exempt status for these professions is not granted automatically: Each employer must reach an agreement with each of their employees (or a representative of employees) and obtain an approval from the labor inspection office. Consequently, many of the professional-work and management-work positions do not receive the legal exempt status in practice. To distinguish from non-exempt positions, the positions that are officially OP-exempt are called *discretionary work arrangement* (DWA) positions in the Japanese regulatory language. According to the General Survey on Wages and Working Conditions (GSWC) in 2012 (Ministry of Health, Labour and Welfare), only 3% of firms with 30 employees or more adopt the DWA, whereas

[2]In April 2019, the labor regulations were amended so that firms must 'monitor' even OP-exempt employees' work hours, after a hysteric debate over concerns with long working hours. This is quite unfortunate as it is likely to increase, rather than decrease, work hours of OP-exempt employees while discouraging flexibility in employees' work schedules.

51.3% of firms with 30 employees or more adopt either the variable hour system or the flexible time system.

What is unique about the Japanese labor market is that there exists another informally defined arrangement, which we call the *supervisory work arrangement* (SWA). Traditionally, Japanese firms appoint employees conducting supervisory work to this arrangement. Employees under the SWA typically work on management-related duties such as managing subordinates, yet they are *managed workers* themselves, having virtually no discretionary power over their work schedules, work hours, or workload. Therefore, they often do not meet the eligibility for the DWA. Nonetheless, Japanese firms form this informal labor arrangement with their employees without an approval from the labor office, and exempt themselves from the overtime pay schedule. In the past, employers lost virtually all lawsuits concerning this labor arrangement. Despite that, employees rarely file lawsuits against such firms for fear of losing job security, and hence, such arrangements are still very common in Japan. Table 1 displays the distribution of employees working under alternative labor arrangements by industry and occupation. The data for the table come from a national survey we conducted in 2016, the details of which shall be discussed in Section 4. Of the 40,418 sample of employees, 19,828 (49.1%) were regular full-time employees. Of these full-time employees, 22% work under variable/flexible hour system, 3.7% work under the DWA, and 6.4% work under the SWA.[3] As shown, a non-negligible share of full-time employees still work under the SWA despite its non-legal status.[4]

[3]In the dataset, we cannot distinguish between off-site work system and DWA. Hence, the DWA counts include those who might be working under the off-site work system. We expect that the off-site workers mostly concentrate in the sales/service occupation.

[4]Indeed, the existence of this informal labor arrangement is what fuels the heated debate on the labor-market reform in Japan. The Abe administration recently approved the action plan for the Realization of the Work Style Reform (WSR). Some called the WSR "a major reform in the history of postwar Japan's labor laws and regulations" (The Cabinet Office, March 28, 2017). At the heart of the WSR lies the debate concerning the White Collar Exemption (WCE) rule. Roughly speaking, the WCE rule (in its current debate) attempts to expand the coverage of the existing DWA for virtually all full-time employees earning the pre-tax annual income of 10.75 million yen or higher. Currently, the DWA regulation concerning the management-related work is more stringent than that on the professional work — the former must secure at least 4/5 votes from a labor-management committee for approval and 1/2 of members of the committee must be employee representatives. Meeting this requirement is quite hard and costly for firms in the Japanese context. Often, committee members are selected from the members of a pre-existing labor union. But only a small fraction of companies have their own labor unions in Japan, and the smaller the companies are, the more unlikely they have their own labor unions (MLHW, 2017). Consequently, many of the supervisory positions have not yet been granted the exempt status. The WCE is, therefore, expected to primarily increase the number of management-work positions that may be covered under the DWA. Central to the debate is a tension between the intended versus unintended consequences of the DWA. While the DWA is intended to encourage flexible and efficient use of work hours (and thus labor productivity), there is also a concern that the DWA may simply offer a loophole that would allow employers to avoid paying for overtime work. Thus, some critics argue that the expansion of the DWA is a way to legalize the unlawful practice of SWA: Management-related positions, despite having no discretionary power at work, would be legally exempt from

3. Alternative Models of Labor Supply

Different models of labor markets produce different predictions about the impact of overtime pay regulation. Trejo (1991) discusses a sharp contrast between the fixed-wage model of Ehrenberg (1971) versus the fixed-job model of Lewis (1969). Roughly, the former can be cast as a model that assumes perfectly elastic supply of labor whereas the latter as a model that assumes perfectly inelastic supply. Consequently, the former model predicts that when the OP regulation places an exogenous increase in the cost of hiring labor, the equilibrium labor hours be shrunk while the equilibrium wage remains fixed. In contrast, the latter model predicts that the equilibrium wage be reduced so as to completely offset the OP regulation while the equilibrium labor hours remain fixed. Taken together, the OP regulation would reduce either labor hours or wages, or both when labor supply is less than perfectly elastic or inelastic. In other words, these models predict that assignment of a job to the OP-exempt status should increase either the labor hours or the wages or both for workers performing that job.[5]

The problem, however, is that these models are silent as to the potential impacts of the DWA relative to the SWA. Neither follows the overtime pay schedule, and hence, the economic margin the above models capture stays constant between the two arrangements.[6] Nonetheless, there is an important difference between the DWA and the SWA. The DWA rules require that prior to its implementation, an agreement must be reached as to the expected task/workload and the hours of work (usually the legal limit of 40 hours per week). The employees are then allowed to work freely, without employer's monitoring, as they see it desirable for completing the assigned task/workload.

Given this, the DWA may be better modeled as a labor-market contract over job attributes, and hence, is amenable to Rosen's model of labor markets. In Rosen's framework, the labor markets are similar to a 'marriage' market, matching employees with various individual attributes (talents, skills, preferences) and jobs with different work attributes (wages, tasks, work environments). In this market, a worker sells her labor service along with her

the overtime-pay regulation.

[5]A number of previous studies have tested fixed-wage and fixed-job models using data from various countries. For example, evidences are from Canada (Friesen, 2001; Skuterud, 2007), the United Kingdom (Bell and Hart, 2003) and Japan (Kuroda and Yamamoto, 2012), and many from the United States (Barkume, 2010; Costa, 2000; Hamermesh and Trejo, 2000; Trejo, 1991, 1993, 2003). The findings are mixed in that neither of these two models are found to be completely invalid in these studies. Which model are more relevant depends on countries and types of workers.

[6]Strictly speaking, the SWA is not granted a legal OP-exempt status, and hence, the risk of facing a lawsuit may affect the firm behavior. However, there is a reason to believe that is unlikely the key driver for observable differences, if any, in labor hours or wages between the two arrangements.

talents and skills and buys a job that comes with a package of various job attributes. Equalizing or compensating wage differentials arise naturally so as to compensate for differences in work attributes — jobs that offer unfavorable working conditions such as risky activities and onerous tasks must pay more than average wages to attract employees. We build on Rosen's framework to arrive at some testable predictions about the economic impacts of alternative work arrangements.[7]

Consider a labor market in which workers have homogenous tastes for work hours l and wages w, but differ in skills or productivity φ. Heterogeneity in tastes can be incorporated into the model, but adds very little to what we discuss below at the expense of notational ease. Workers perform tasks, whose outputs are measured conveniently by t, through a production technology $t = h(l, \varphi)$. To focus on essentials, consider two types of jobs, one that requires demanding tasks t_H and another that requires much less t_L. The t_H jobs may be supervisory work or professional work that can be potentially covered by the DWA. Suppose for the moment that two types of jobs are offered at given wages w_L and w_H, respectively.

Assume that h is uniquely invertible with respect to l, and that l increases with t and decreases with φ. It follows then that preference relations over the (l, w) space can be uniquely mapped to those over the (t, w) space, and that workers will have heterogeneous preferences in the (t, w) space even if they have homogenous preferences over (l, w). Figure 1 depicts an example of such preference relation in the (t, w) space. This worker is currently indifferent between a job package A versus another job package C. Hence, if the worker is given two job packages A and B (instead of C), she would prefer to choose A because the compensating wage differential z required for this worker to take up more onerous task t_H instead of t_L is larger than the actual wage differential Δw.

A firm's demand for labor can be also modeled almost symmetrically. Suppose that firm's profit per worker π is given by $\pi = f(t, \nu) - w$, where ν describes firm's productivity in generating profits from output t. Assuming sufficient concavity of f in t, the iso-profit curve in the (t, w) space is convex, as in Figure 1. Given the market wages w_L and w_H, this firm would attain higher profits from offering job type B than job type A. This firm is willing to compensate as much as z, yet the actual wage differential is Δw. Hence, if the market wages remain as they are, the firm would fetch an economic rent equaling $z - \Delta w$. Rosen (1986) has shown how the equilibrium wages arise through the market clearing condition, which can be defined explicitly by assuming some distributions for φ and ν.

Now let us introduce another job characteristic s, which equals 1 if a job is converted

[7]As discussed in Trejo, the fixed-job model also has some roots in Rosen's framework. However, the fixed-job model leaves out one important aspect of Rosen's model — both labor hours and wages are part of job attributes that must be negotiated and determined via market equilibrium, and hence, neither attributes should stay "fixed" over different regulatory arrangements.

to the DWA. Assume that only the t_H job can be converted. The promise of the DWA is that it gives more freedom and discretionary authority as to how a worker may complete the assigned task t_H. Because workers can always choose to work the same way without adjusting their work hours, all workers should weakly prefer this added work 'amenity'. Then this added work amenity would shift the indifference curve down for every worker who strictly prefers it. On the other hand, as discussed in Section 2, converting a position to the DWA is costly to firms. Hence, if there is no productivity gain from the arrangement, this added job feature might also shift down any firm's iso-profit curve. These changes are illustrated with IC' and IP' (red curves) in Figure 1. One important promise of the DWA, as its proponents argue, is that this added job amenity may also increase worker productivity — workers know their best in what they do and how to achieve the given task, and hence, they work most efficiently if they are given more authority as to allocation of their own labor. If the productivity gain is sufficiently larger than the cost, firm's iso-profit curve might indeed shift upward (IP'') instead of downward (IP').

The reasoning suggests, unfortunately, that the economic impacts of this added work attribute are not necessarily clear. Because virtually all workers prefer having this attribute, their indifference curves would shift downward. But the iso-profit curves might shift up for some firms and down for other firms. Hence, the compensation required to accept the onerous task is lower for all workers while the compensations firms are willing to offer for it may be higher or lower. This should bring the equilibrium wage differential smaller, though the magnitude of the impact may not be large since firms may enjoy productivity gains. Moreover, the impact on the labor hours is indeterminate. On one hand, the DWA allows workers to work flexibly and efficiently, and hence, in some cases, workers may be able to cut back hours of work. On the other hand, the added feature will attract more inefficient workers to take up the onerous task, and hence, this selection may lead to longer average work hours. This gives us a good reason to take this to an empirical study, which we shall turn to below.

4. Data

Our study relies on a large cross-sectional sample of employees from a nationwide internet survey conducted in December 2016. In designing the survey, we had two goals in mind. The first is to obtain a sufficiently large sample for the target population of interest. By design, WCE is intended only for workers who conduct non-manual, non-repetitive tasks. While it may be certainly interesting to see its effect on, for example, construction workers or call-center clerks, more practical policy implications can be gained on workers that are directly

targeted by WCE. Hence, our target population consists of full-time workers who conduct non-manual, non-repetitive tasks who are equally likely to be assigned to alternative work arrangements. Naturally, these workers are typically male, older, more educated, and more concentrated in certain types of occupations such as managerial and technical jobs. Second, we aim for collecting more and better information on worker's occupational characteristics such as occupational rank, department size, the number of subordinates, on-the-job skill requirements, and work arrangement status. Information on these attributes is often not available in national surveys in Japan (see Table 3 below), yet is critical in ensuring the unconfoundedness of the treatment assignment.

The survey was administered under the contract with Nikkei Research Inc. to the pool of registered internet monitors who had some jobs at the time of the survey. The survey resulted in 40,418 responses, of whom 19,828 were regular full-time employees. To focus on the population of interest, we restrict our sample further to full-time employees whose ages are between 25 and 65 years old and who work at domestic companies under either the traditional work arrangements or two OP-exempt arrangements, SWA and DWA, excluding those under other flexible work arrangements. We further trim our sample to employees in construction, manufacturing, IT/telecommunication, whole trade/retail trade, finance/insurance, and real estate. The numbers of employees under either SWA or DWA are not sufficient in other industries. For the same reason, we restrict our sample to employees whose occupation is categorized as management/supervisory, professional/technical, administrative, clerical, sales, or service. Our working sample then consists of 7,308 observations after dropping observations with missing information.

Table 2 compares our sample to the samples from two well-known surveys in Japan, the Employment Status Survey (ESS) and the Japan-Keio Household Panel Survey (JHPS/KHPS).[8] As expected, our sample contains higher shares of male, older, more educated workers than the ESS sample. We also have higher shares of administrative, managerial, and professional occupations. Clearly, ours is not a nationally representative sample, and instead is skewed toward our target population of interest. This skewness, however, should not be necessarily taken as the weakness of the internet survey. For instance, the JHPS/KHPS sample also shares similar tendencies — higher shares of male, older, and administrative and managerial workers — despite the fact that it is administered via mail with stratified random sampling.

The question is, does this skewed sample representation pose problems in estimating the causal effect of interest? Our answer is, No. Our causal inference will be based on an

[8]ESS is a nationally-representative labor survey conducted every five year, covering roughly 110 million individuals at age 15 or older. One drawback of ESS prevents us from exploiting ESS for our analysis: it lacks information on important occupational characteristics (see Table 3).

informative subsample, where we only use matched workers based on their pre-treatment covariates so as to ensure the unconfoundedness of treatment assignment. What matters is that this subsample, to be constructed later, is representative of the target population of interest. As discussed earlier in this section, our population of interest is the marginal workers whose jobs involve non-manual, non-repetitive tasks who are equally likely to be assigned to different work arrangements. In this sense, whether our sample is representative of the population at large is irrelevant. However, there is indeed another, closely-related concern. That is, if there is endogenous sample selection and if that cannot be corrected for by accounting for all observables. To gauge the degree of endogenous sampling due to unobservables, we report the sample shares of industries after adjusting for the joint distributions of age, education, gender, and occupation in the ESS sample in the last two columns of Table 2.[9] The idea is that if these observables are the primary factors for the skewed sample representation, then adjusting for the joint distribution of these variables should make our sample distribution come closer to the ESS sample. The adjusted sample distribution of industries indeed comes closer to the ESS sample. However, some difference still remains.[10] We shall return to this issue in Section 6.

Table 3 reports the descriptive statistics of our working sample by work arrangement status. Of the sample of 7,308 observations, there are 552 SWA workers (7.6%) and 404 DWA workers (5.5%). As discussed in Section 2, the OP-exempt status is not yet prevalent in Japan, but our large-scale survey results in a relatively large number of the OP-exempt workers in the sample. Panel A displays the means and the standard deviations of the two outcome variables of interest, weekly work hours and hourly wage. In collecting information on work hours, we ask each respondent what time s/he starts and finishes working on a typical day.[11] Using this information, we calculate hours of work per day. Annual incomes are reported with an interval of 1 million yen. We compute an hourly wage for each observation by dividing the mid-point of income interval by weekly work hours times 48 (= 4 weeks $\times$ 12 months). Panel B displays the same for pre-treatment variables used in estimation. Virtually all variables in Panel B conform to the standard labor surveys, and hence, their

[9]The adjusted share S_i^{Adj} is computed as

$$S_i^{Adj} = \sum_{\mathbf{x}} S_i(\mathbf{x}) P^{EES}(\mathbf{x})$$

where $S_i(\mathbf{x})$ is the share of industry i conditional on the adjusted variables $\mathbf{x}$ computed using our sample, and $P^{EES}(\mathbf{x})$ is the joint distributions of $\mathbf{x}$ computed from the EES sample.

[10]The sum of squared differences in industry shares decreases from 0.035 before adjustment to 0.011 after adjustment.

[11]In contrast, the ESS and the JHPS/KHPS surveys simply ask each respondent to record average hours of work per week over a year.

detailed descriptions are left to the Appendix A. Two exceptions are the job rank and the use of logical reasoning skills at work. The former is defined as an employment position that a respondent hold at his/her workplace: Division heads (*bucho*), section chiefs (*kakaricho, kacho, or jicho*), and ordinary employees. For the latter, the survey asked a respondent to allocate 100 points to five types of skills that s/he finds necessary to complete his/her usual tasks at work. The four categories are: logical reasoning, interpersonal, leadership, creative, and paperwork. A preliminary analysis suggests that point allocations over these skills are highly correlated. Hence, we use only the logical reasoning skill for ease of interpretation as well as to avoid multicollinearity.

We see that both SWA and DWA are associated with *longer* work hours and *higher* hourly wages than the OP-regulated status. However, these work arrangements are also associated with longer years of schooling, higher occupational rank, and larger department/firm size. Interestingly, DWA has a higher rate of ordinary personnel having no direct subordinate whereas SWA has a higher rate of division head managing a large number of subordinates. This has to do with the fact that DWA is more prevalent in professional and technical occupations whereas SWA is more prevalent in managerial and supervisory occupations. Despite these differences, however, the use of logical reasoning skill at work does not seem to differ substantially between DWA and SWA, although we do see a difference between OP versus non-OP arrangements. These important, yet subtle, differences in the distributions of job-level characteristics across alternative work arrangements highlight the need to control these important covariates. Yet, as shown in the last two columns, the ESS and the JHPS/KHPS surveys do not contain information on these variables.

Table 3 also hints us an important empirical challenge. Changes in work arrangements do not occur frequently, and when they do, they tend to concur with simultaneous changes in job characteristics (e.g., occupational rank and task) as well as worker characteristics (e.g., marital status). This makes conventional panel-data methods highly unreliable. For instance, as a precursor to this study, we compiled a subsample of 107 full-time workers from the JHPS/KHPS dataset who are sufficiently informative for our purpose. Of these individuals, only 7 workers changed, either from OP to non-OP, or from non-OP to OP between 2015 and 2016. A fixed-effect regression of work hours has returned highly insignificant coefficients on work arrangements dummies, only after controlling for a small number of job and worker characteristics. For this reason, we employ an alternative (possibly more viable) approach, which we turn to in the next section.

5. Empirical Strategy

Our quantity of interest, for each of the outcome variables of interest y (i.e., work hours l or wages w), is the population average treatment effect (ATE)

$$ATE = E[y_{it} - y_{is}],$$

where y_{it} and y_{is} denote potential outcomes under alternative labor arrangements $t, s \in \{0, 1, 2\}$. For ease of interpretation, we let $t = 0$ denote regular OP status, which serves as the base for evaluating the impact of the SWA ($t = 1$) and the DWA ($t = 2$).

A number of empirical strategies are available for estimating these ATEs. The empirical challenge in our context, however, is that as discussed in Section 4, only a handful of changes in labor arrangements occur each year in a moderately-sized panel dataset, and when they do occur, such changes often concur with simultaneous changes in occupational tasks and ranks. Panel-based methods such as fixed-effect and difference-in-differences models, therefore, cannot plausibly control for the endogeneity of assignments to alternative labor arrangements. A potentially better approach in this case is to exploit cross-sectional variation and find an informative subsample consisting of identical workers who conduct identical tasks at identical firms, yet happen to work in different work arrangements. This is the venue we pursue.

For each unit i, let D_{it} denote the assignment to labor arrangement t, with $D_{it} = 1$ if i works in arrangement t and $D_{it} = 0$ otherwise. Since the assignments to labor arrangements are mutually exclusive, we observe the outcome of only one assignment. Hence, by definition, $y_i = D_{i1}y_{i1} + D_{i2}y_{i2} + (1 - D_{i1} - D_{i2})y_{i0}$. Our task is to choose an informative subsample M_{ts} for each pair of treatments t and s from the sample of n workers, so as to consistently estimate the estimand of interest

$$\tau_{mts} = \frac{1}{m_{ts}} \sum_{i \in M_{ts}} (y_{it} - y_{is}), \tag{1}$$

where $m_{ts} = \#M_{ts}$. The state of the art in the literature (e.g., Imbens and Rubin, 2015) suggests that we break down this task into three steps: (1) selection of pre-treatment variables X, so that the unconfoundedness (or ignorability) assumption holds, (2) choice of a matching method, so as to ensure 'closeness' of X for these subsamples, and (3) choice of an estimator, so as to implement the empirical analogue of the estimand. The literature also suggests that of these, (1) and (2) are the most important because the estimates are more sensitive to the choice of estimators when the covariate distributions are substantially different than when they are similar [see Chapter 14 of Imbens and Rubin (2015)].

Selecting Pre-treatment variables

The key here is to select the set of covariates X_i such that the unconfoundedness assump-

tion holds:

$$y_{is} \perp D_{it}|X_i \quad \text{for} \quad s,t \in \{0,1,2\}.$$

That is, the assignment to a work arrangement is independent of the potential outcomes, once conditioned on X_i. This assumption is also termed as selection on observables.

Our theory predicts that labor hours and wages may differ across different work arrangements, even with identical jobs and tasks. The problem, however, is that workers under different work arrangements are typically assigned to different jobs and tasks. Hence, the most important conditioning variables are those that characterize their jobs and tasks: department size, occupation classification, occupational rank, on-the-job skill requirements, and the number of subordinates to supervise. Fortunately, our dataset contains information on these job characteristics.

As with other studies, we also use variables that describe workers' socioeconomic characteristics: age, education, gender, marital status, and the number of children in 3rd grade or younger. Note that Rosen's framework predicts that there is generally a sample selection — workers self-select into tasks that match their preferences and skill sets. Our choice of pre-treatment variables is *not* meant to control for this type of self-selection. What we meant to control for instead is the unobservable aspects of jobs and tasks that correlate with work arrangements. For example, a female worker with children who is assigned to a discretionary work arrangement may be permitted to work on a less demanding task. Observable individual-level characteristics can serve as controls for such unobservable aspects of jobs and tasks.

Our discussion in Section 2 also highlights the need to control for firm-level characteristics. On one hand, the likelihood of DWA is higher at firms that are larger in size or located in urban areas, presumably because the cost per worker of converting a position to DWA can vary due to scale economies and search frictions. On the other hand, work hours and wages are generally higher at these firms since their labor demand are also high. This means that the unconfoundedness assumption would be violated if these firm-level characteristics are not controlled for. We thus control for the following set of firm-level characteristics: firm size, industry classification, and population density.

A Hybrid Matching Method

The next important step is to select a matching method to improve balance in covariate distributions between treatments. The most straightforward way to ensure the covariate balance is *exact* matching: i.e., choose M_{ts} such that for all $i \in M_{ts}$, $X_{it} = X_{is}$. With the large number of covariates, however, exact matching often leads to no or insufficient matches. As a result, analysts often rely on *approximate* matching methods such as the

Mahalanobis matching and the propensity score matching. These methods find matches using some metrics to measure the 'distance' of covariate vectors: $d(X_{it}, X_{is})$. Recently, however, studies have revealed an important problem with these existing methods (e.g., Iacus *et al.*, 2011) — these methods do not necessarily ensure balance in covariate distributions, and matched units using these metrics can be indeed quite 'far' from each other in the multidimensional space. In our analysis section, we demonstrate that this is indeed true in our dataset as well.

Iacus *et al.* (2011) proposes an alternative matching method known as coarsened exact matching (CEM). The idea behind CEM is quite simple. Analysts often have substantive knowledge on the empirical problem at hand, so they can use the knowledge to coarsen each variable without losing its crucial information. For example, worker's age may be coarsened into 20s, 30s, 40s etc. Then exact matching is applied to the coarsened covariates. The method is shown to work well in a variety of empirical contexts for improving balance in covariate distributions without losing the size of matched subsample [Iacus *et al.* (2011; 2012)].

CEM, however, is not free from the curse of dimensionality. In labor market contexts, analysts often need to control for a large number of categorical variables, which can be coarsened only so as not to violate the ignorability assumption. This, unfortunately, may lead to insufficient matches. This turns out to be the case with our study — CEM, with 13 covariates and sufficient coarsening, resulted in only 17 matched units. In this regard, we are back again to the essential trade-off in choice of a matching metric, long recognized in the literature, between balancing the covariate distributions (to ensure unconfoundedness) versus finding sufficient matches for precision of estimates.

Given this trade-off, we propose a *hybrid* matching algorithm, which takes the best of both worlds. To see the idea, note first that the idea of coarsening approach can be applied to the propensity score itself. That is, find matches based on the coarsened values of the propensity score. Indeed, this method is known as a subclassification (or blocking) on the propensity score (Imbens and Rubin, 2015). But the subclassification method per se can result in highly unbalanced covariate distributions (as shown later) as it still relies on the propensity score as a distance metric. We, however, recognize here that this propensity score can be thought of as another conditioning variable, which extracts much of the important information from all variables in X. We can take a few variables from X and this new variable and apply CEM on this new subset of conditioning variables.

Formally, we first estimate the propensity score $P(X)$ using all variables in X. Then, take a subset of variables from X and call it X^c. We coarsen P and X^c into $C(P, X^c)$. Then, apply exact matching on $C(P, X^c)$. By construction, this hybrid method uses all of

the information in X. Hence, it operates under the same unconfoundedness assumption. It has the advantage of propensity score matching as a dimension-reduction algorithm, so this method is guaranteed to produce a sufficient number of matches as long as the dimension of X^c is small. Yet, it applies (coarsened) exact matching on some of the key variables X^c, so that the covariate balance for these important variables is ensured.

Except for our explicit use of CEM approach, hybrid matching algorithms have been well recognized in the literature [see p.343 of Imbens and Rubin (2015)]. We, however, exploit the merits of our hybrid method more explicitly. We vary the partition of X in implementing this method, passing variables one by one to the set X^c. Importantly, since we continue to use the same propensity score (which uses all of X), expanding the set X^c can only improve the covariate balance. Hence, this method allows us to visualize the essential trade-off in real empirical contexts — as we pass more variables to X^c, the covariate balance improves, reducing the bias of the estimate (in expectation), yet the size of matched sample gets smaller, making the estimate less precise. Our analysis section substantiates the importance of assessing this trade-off in our context. Indeed, CEM and subclassification estimators can be thought of as two special cases of this hybrid method.

Implementation with Multiple Treatments

One important complication arises in implementing this hybrid method in our context. Instead of binary treatment, we have three (mutually exclusive) treatment levels: non-OP ($t = 0$), SWA ($t = 1$), and DWA ($t = 2$). We closely follow Cattaneo (2010) to modify the subclassification estimator amenable to our context. He has shown how the basic idea of inverse probability weighting (IPW) estimator (Horvitz and Thompson, 1952) can be extended to the case of multiple treatments. This estimator proceeds in two steps. First, estimate, using the full sample, the generalized propensity score, $p_t(X) = \Pr(T = t|X)$, for each treatment level $t = 0, 1$, and 2. This can be done, for example, by estimating multinomial logit. Second, use the set of empirical moments to arrive at Horvitz-Thompson-type IPW estimator. The subclassification estimator can be indeed interpreted as a version of this IPW estimator.

The complication here, however, is how one might construct a stratum (or subclass) for this (and our hybrid) estimator. Cottaneo's arguments imply that an appropriate balancing score for units under each treatment t is the propensity score for *that* treatment p_t. In principle, then one should define a stratum by a set of multidimensional bins (as opposed to one dimensional bins). This, however, adds to the curse of dimensionality problem discussed above, and undermines the merit of the subclassification estimator (and our hybrid estimator).

We consider the solution to this, using the economic argument as follows. Note, first,

that SWA and DWA, are two OP-exempt work arrangements, and are economically quite different from the non-OP arrangement. The primary difference between SWA and DWA is their legal status: the former needs approval from the labor bureau and the latter does not. Hence, employees under the two arrangements tend to work on similar tasks. In the Japanese context, employers, most likely, consider assignment of their workers in a two-step sequence: first whether to assign to OP versus non-OP, and then given the non-OP status, whether to assign to SWA versus DWA. The logic suggests that the probability of non-OP status, $P(X) = p_1(X) + p_2(X)$, is a sufficient balancing statistic for all three treatments.

Given this, we implement our hybrid method in five steps.

(1) Estimate the propensity score $P_i(X_i) \equiv \Pr(D_{1i} = 1 \text{ or } D_{2i} = 1|X_i)$, using the full sample. This can be done by either a logit model estimating the probability of the single non-OP status or a multinomial logit and summing the two estimated probabilities $\hat{p}_1$ and $\hat{p}_2$. Both produce quantitatively similar results.

(2) Select the subset of pre-treatment variables in X^c. This is the set of variables for which analysts would like perfect balancing of covariate distributions. In our case, these are the job and task characteristics, which tend to co-move with work arrangement status.

(3) Then, coarsen each of the variables X^c as well as $\hat{P}$ into $C(\hat{P}, X^c)$. This effectively creates the set of multi-dimensional bins in the space $(\hat{P}, X^c)$.

(4) Apply exact matching to $C(\hat{P}, X^c)$. Strata that contain units in *all* treatments are retained, and strata that contain units in only one or two treatments are discarded. This procedure produces subsample M_t for each treatment level.

(5) Estimate the ATE for each pair of treatments, using a version of subclassification-weighting estimator, which we shall explain in detail below. For ease of interpretation, we use the OP status as the base treatment (which may be termed as 'control' analogous to the case of binary treatment).

In the analysis section, we vary the set X^c from an empty set (which corresponds to the classic subclassification estimator) to the full covariates (which corresponds to CEM). This serves as a way to visualize the trade-off between covariate balance versus precision of estimates and, by doing so, to assess the direction of bias due to imbalance in covariate distributions. This is particularly useful for evaluating the robustness of estimates to covariate balance as well as evaluating the direction of bias due to covariate imbalance in case estimates are not robust.

Estimation and Inference

For estimation, we closely follow the classic subclassification estimator outlined in Imbens and Rubin (Ch. 17, 2015). The estimator has the form:

$$\hat{\tau}_{mts} = \sum_{b \in B} q^b \left(\bar{y}_{it}^b - \bar{y}_{is}^b \right),$$

where B is the set of blocks (or 'strata') and q^b is the relative size of the subsample m^b in block b in the matched subsample m: $q^b = m^b/m$. That is, we take a simple difference in means of the outcome variables between the units that received different treatments and then average these means using weights q^b's. It is straightforward to see that this estimator essentially amounts to the Horvitz-Thompson type weighting estimator:

$$\hat{\tau}_{mts} = \frac{1}{m_t} \sum_{i \in M_t} w_i \cdot D_{it} \cdot y_i - \frac{1}{m_s} \sum_{i \in M_s} w_i \cdot D_{is} \cdot y_i \tag{2}$$

where the weight w_i is the relative frequency of units that received each treatment status within each stratum.

The weights w_i's are defined as follows per Iacus *et al.* (2011). Let M_t^b denote the units with treatment t in stratum b, with count $m_t^b = \#M_t^b$. We emphasize here that each stratum is defined by the Cartesian product of multi-dimensional bins on space $(\hat{P}, X^c)$. The number of total matches for each treatment t is $m_t = \#M_t = \sum_b m_t^b$. We then assign to each *matched* unit i the following weight w_i:

$$w_i = \frac{m^b}{m_t^b} \cdot \frac{m_t}{m} \quad \text{for } i \in M_t^b.$$

By definition, a unit receives w_i if that unit belongs to a stratum for which units are observed for all treatment levels.

Then as with the subclassification estimator, we can implement this estimator using the weighted least squares (WLS) regression:

$$y_i = \alpha + \beta_1 D_{i1} + \beta_2 D_{i2} + \epsilon_i,$$

with weights w_i. This WLS regression yields $\hat{\tau}_{m10} = \hat{\beta}_1$ and $\hat{\tau}_{m20} = \hat{\beta}_2$, and we can estimate τ_{m21} as $\hat{\beta}_2 - \hat{\beta}_1$. Since the sampling variance of our estimator is the sum of the sampling variances of the two terms of equation (2), the standard errors of our estimators can be obtained from this WLS regression.[12] Imbens and Rubin (2015) discuss several desirable

[12]The fact that we use the estimated propensity score causes a technical difficulty of computing standard

properties of the classic subclassification estimator over Horvitz-Thompson's weighting estimator, which uses the inverse of the estimated propensity scores as weights. Essentially the same arguments apply for our hybrid estimator, except that ours should, in principle, reduce bias further by improving the covariate balance.

One caveat is in order. By design, a matching estimator is used to form an 'informative' subsample in which covariate distributions are well balanced between treated and control units. In the process, the original sample is trimmed down to a (typically much smaller) subsample. Because covariates are balanced for this subsample, under the unconfoundedness, we obtain an unbiased estimate of the causal effect from this subsample. However, we need to be careful about the *population* for which the estimate from this subsample is intended for. By trimming the sample, we loose observations, and in many cases, we end up with no observation for some subpopulations. If treatment effects are heterogeneous across different subpopulations (that is, treatment effects depend on covariates), losing observations for some subpopulations may indeed increase the bias for the population ATE. A subclassification estimator (and their variants such as ours) takes care of this issue partially by weighting estimates according to the size of matches in each subclass. This, however, is no help when we have no observation for some subclasses that we deem are important for estimating the population ATE. The problem essentially boils down to whether or not the treatment effects are expected to differ between matched versus non-matched units. Solon *et al.* (2015) offer an excellent guide on this point. If we expect heterogeneous treatment effects, we should explicitly 'study the heterogeneity' instead of 'averaging it out' (by weighted least squares). In the case of matching estimator, however, we cannot take up this advice directly because the problem lies with some subclasses that have no informative subsample. For these subclasses, estimates based on the original unmatched sample are not useful for assessing the heterogeneity. We instead take the following view — we do not attempt to infer from these subclasses, and instead, we appropriately define the target population for which the matched sample is well suited for statistical inference. We shall return to this issue in Subsection 6.D where we examine the matched sample.

Measures of Covariate Balance

A commonly used measure of balance in the covariate distribution between units with

errors. Abadie and Imbens (2016) show how the first-stage estimate of the propensity score may affect the asymptotic variance of the propensity score matching estimators. Their argument can apply to our estimator as we match observations based on the estimated propensity score. The standard error estimates from the WLS regression do not take into account these estimation errors. Unfortunately, commonly used bootstrapped standard errors are not a legitimate solution to this problem in the case of subclassification or matching estimators. Indeed, bootstrapping can make the problem even worse (see Abadie and Imbens, 2008; Imbens and Rubin, 2015). As discussed in Abadie and Imbens (2016), however, ignoring the estimation errors does not necessarily understate the standard errors.

different treatment levels is a normalized difference (also known as normal distance). The normalized difference takes the difference in the average values of a covariate by treatment status, scaled by their standard deviations. As a rule of thumb, larger than one quarter on this measure would be taken as a sign of imbalance (Imbens and Rubin, 2015). In such a case, regression-based approach is unlikely to be effective in removing bias. To assess the degree of multivariate imbalance, one could simply average the values of this measure (in absolute terms) over all covariates. We call this average normalized difference (AND). Formally, for a pair of treatment statuses t and s,

$$AND_{t,s} = \frac{1}{K} \sum_{k=1}^{K} \left| \frac{\bar{X}_{k,t} - \bar{X}_{k,s}}{\sqrt{\left(s_{k,t}^2 + s_{k,s}^2\right)/2}} \right|.$$

There are several problems with this measure of imbalance, however. First, this measure can assess only mean imbalance, but says nothing about other moments. Second, in many real applications, improving balance on one variable often leads to imbalance in other variables (Iacus *et al.*, 2011). The AND measure can be highly insensitive to such trade-offs. To address both of these problems, Iacus *et al.* (2011, 2012) propose an alternative measure of imbalance, known as the L1 measure of multivariate imbalance.

The measure starts by choosing bins of each variable. Define $H(X)$ to be the Cartesian product of such bins over the set of covariates X. In our case, a natural choice for H would be those that correspond to the coarsened set $C(X)$. Let h index elements of H, which denotes each multidimensional bin. Let f_t and f_s be the relative empirical frequency distributions for a pair of treatments t and s defined over these multidimensional bins H. Analogously, $f_{h,t}$ is the relative empirical frequency for observations in the bin h for treatment t. Then, the L1 measure of multivariate imbalance between treatments t and s is

$$\mathcal{L}_1(f_t, f_s; H) = \frac{1}{2} \sum_{h \in H(X)} |f_{h,t} - f_{h,s}|.$$

That is, the L1 measure calculates the L^1 norm of the distance between the two multivariate histograms, one for units with treatment t and another for units with treatment s, over the multidimensional bins H. By definition, this measure incorporates imbalance in all moments and of all covariates included in X.

6. Results

A. Work Hours and Hourly Wage

We now discuss our main results. In Table 4, we report the estimates of ATEs from four regressions: (1) ordinary least squares (OLS), (2) subclassification on propensity scores (PS blocking), (3) our hybrid matching estimator, and (4) coarsened exact matching (CEM). All estimators start from the same sample consisting of 7,308 full-time male employees. Except for OLS, however, the sample is further trimmed to those with estimated propensity scores $\hat{P}$ ranging between 0.20 and 0.50. This ensures that our working sample satisfies the overlap assumption, thus, forming a comparable basis, for all three treatment levels. The top panel reports the estimates of ATEs for two outcome variables, weekly working hours and logged hourly wage. For all estimators, we use the same set of covariates: i.e., thirteen pre-treatment variables that we believe are essential for ensuring the unconfoundedness assumption. The bottom panel reports two measures of covariate distributions, L1 measure and average normal distance (AND), for each of three treatment pairs.

The table demonstrates the stark influence of alternative matching methods on covariate distributions. In the original sample, L1 measures are close to 1 for all treatment pairs (see the column under OLS). This means that in the original sample, all treatment pairs have highly unbalanced covariate distributions. This is also confirmed with AND measures (as a rule of thumb, the normal distance of close to 0.25 is taken as the sign of imbalance). As expected, PS blocking improves the differences in the means of covariates, but does not quite improve the overall covariate balance — L1 measures remain above 0.9 for all pairs. This result is consistent with Iacus *et al.* (2011). Our hybrid matching method, using 8 out of 13 pre-treatment variables for exact matching, substantially improves the covariate balance — L1 measures are now around 0.7 and AND measures are far below 0.1. CEM, using all thirteen covariates, leads to (close to) perfect covariate balance. However, this balancing of covariate distributions comes at the cost of further decline in sample size. Our hybrid method yields 89 matches whereas full CEM leads to only 17 matches.

The table also demonstrates the substantive impacts of covariate balance on the ATE estimates. With OLS and PS blocking, where covariates are highly unbalanced, OP-exempt arrangements (i.e., SWA and DWA) are estimated to significantly increase work hours relative to the OP-regulated arrangement (see Panel A). This is consistent with earlier empirical findings (Kuroda and Yamamoto, 2012).[13] With our hybrid method, where the sample consists of only those who have identical occupational ranks and tasks, only the DWA arrangement is statistically significant. As a result, the difference in work hours between DWA and SWA turns significantly positive. This difference should be attributed to the WS exemp-

[13]Using KHPS, Kuroda and Yamamoto (2012) compared OP-exempt workers with regular OP workers and find that the OP-exempt tend to work longer than the non-OP-exempt while hourly wages do not differ significantly. However, their discussion mainly focuses on workers assigned to supervisory/managerial positions, who have been traditionally exempt from OP regulations without full merit of WCE.

tion because DWA is both OP-exempt and WS-exempt, but SWA is only OP-exempt. Thus, we conclude that the OP exemption (i.e., SWA) per se does not necessarily increase work hours, yet the WS exemption does, at least in the Japanese context. This result is indeed consistent with our theoretical prediction. Notice that the estimates from CEM have the same signs, but *larger* in magnitude, than the estimates from our hybrid method. Nonetheless, the difference in work hours between DWA and SWA turns insignificant because CEM estimates have larger standard errors due to its small sample size.

The effects of covariate distributions on hourly wage are even more stark (see Panel B). OLS and PS blocking estimates imply that OP-exempt arrangements increase hourly wage relative to the OP-regulated arrangement. With OLS and PS blocking, units are not matched on occupational ranks and tasks, at least directly. Hence, these estimates are likely picking up the confounding effects of these characteristics. With our hybrid method, where units are directly matched on these attributes, DWA is estimated to *decrease* hourly wage thought insignificant, whereas SWA has no or negligible effect, relative to the OP-regulated arrangement. As a result, DWA is estimated to significantly decrease hourly wage relative to SWA. This is consistent with Rosen's theory of equalizing wage differentials: workers prefer having flexibility in work scheduling, and hence are willing to take on flexible work arrangements for lower wages.

While our hybrid approach does seem to yield sensible estimates, there may be a concern with the precision of estimates. The sample size may be too small to make credible statistical inferences. The advantage of our approach arises in this context. As more variables are passed onto exact matching (while we keep using the same PS blocking), the sample size necessarily shrinks. This surely increases the sampling variance. However, the improved covariate balance, as a result, lead to less bias, and potentially, less variance within each stratum. One needs to weigh the imprecision (or insignificance) of estimates against the bias of estimates. In this regard, looking at a single estimate is probably not quite helpful, as the estimate itself is a random variable. Hence, we attempt to access the direction of bias by examining how the estimates change as more variables are passed onto exact matching.[14]

Figures 2 displays how the sample size and the measures of covariate imbalance change with the number of covariates used in exact matching. As the covariate balance improves, the sample size shrinks quite sharply. Importantly, it is relatively harder to improve on L1

[14]Obviously, there is some order dependency: The order in which covariates are passed to the exact matching changes the paths for both the measures of imbalance and the estimates. The results presented here use economic reasoning on the order: Match job characteristics first, worker characteristics second, and firm characteristics the last. We also experimented with an algorithm, which seeks to minimize the L1 measure for each number of covariates in exact matching. Not surprisingly, the hybrid method converges to CEM faster with this algorithm (i.e., at a smaller number of covariates in exact matching), but virtually all other aspects of the results stay essentially the same. See Online Appendix C.

measure than on AND. This is expected, as discussed in Section 5, since the L1 measure is sensitive to the multivariate imbalance whereas AND only considers the variable-by-variable imbalance. Figure 3 displays the ATE estimates along with 95% confidence intervals, in response to the changes in sample size and covariate distribution. We see some patterns that seem suggestive of the direction of bias. As we increase the number of covariates used in exact matching, (a) the gap in working hours between SWA and OP tends toward zero, (b) the gap in working hours between DWA and OP tends to increase, (c) the gap in hourly wage (in log) between SWA and OP tends to stay roughly the same, and (d) the gap in hourly wage (in log) between DWA and OP tends toward a negative value. These patterns are consistent with our discussion on Table 4, and hence, support our inferences from the estimates reported in Table 4.

In sum, we can summarize our findings as follows. The OP-exempt arrangement may have a tendency to increase work hours and hourly wage, relative to the OP-regulated arrangement, yet the impacts may not be large. The quantitatively large impacts found in earlier studies may be picking up the confounding effects of simultaneous changes in occupational ranks and tasks. On the other hand, the WS-exempt arrangement tends to increase work hours, but decrease hourly wage, relative to the WS-regulated arrangement. This result is not in line with the Japanese government's sentiments toward the WCE, but is indeed consistent with the theory of equalizing wage differentials.

B. Satisfaction from Work

There is a growing interest among economists and policy practitioners on the relationship between mental health and labor productivity (e.g., Bubonya *et al.*, 2017; Chatterji *et al.,* 2011; Stevenson, 2017). In this context, there is a potential for aligning the interests of firm and worker — flexible work arrangements may improve worker's productivity and mental health at the same time. Bloom *et al.* (2015), for example, report that in the Chinese context, the Work from Home (WFH) arrangement has led to improved work satisfaction while substantially increasing work performance and retention rate. Similar gains may be expected from the discretionary work arrangement.

In this regard, Rosen's theory of equalizing wage differentials indeed offers a clear economic prediction: Workers who work under the WS-exempt arrangement (i.e., DWA) *willingly* chose to do so despite its negative wage premium, and hence, for these workers, an increase in work satisfaction must be high enough to compensate for the decrease in hourly wage. Fortunately, we do have some data on measures of work satisfaction.[15] We exploit

[15]Unfortunately, we do not have data on measures of labor productivity at work.

our survey responses to the following two questions: (a) How satisfied are you currently with your work? and (b) How often do you feel stress from daily life? Both responses are recorded in a five-point scale, with 1 indicating 'not at all satisfied' ('none') and 5 indicating 'highly satisfied' ('very often'). Question (a) is directly related to satisfaction from work whereas question (b) asks frequency of feeling stress from daily life. We take daily life stress as a proxy for stress at work since work constitutes a substantial part of daily life at least in the Japanese context. A more detailed description of the variables is available in the Online Appendix A.

As before, Table 5 reports the estimates of ATEs from four regressions: OLS, PS-blocking, our hybrid method, and CEM. Panel A and Panel B, respectively, display the results on work satisfaction and frequency of daily stress. As discussed earlier, we prefer the estimates from our hybrid method (in column 3) over others, and hence, focus our discussion on these estimates.

Two surprising observations are in order. First, relative to the OP-regulated arrangement, DWA is estimated to *decrease* work satisfaction, after directly matching on occupational and worker characteristics. Surprisingly, this decrease in work satisfaction is larger for DWA than for SWA, although the difference between the two estimates is not statistically significant. Second, DWA is estimated to *increase* frequency of feeling daily stress both relative to OP and SWA arrangements. These positive estimates are statistically significant at conventional levels. Furthermore, as Figure 4 demonstrates, the estimates from our hybrid method essentially have consistent signs (although their magnitudes change) across alternative sets of covariates used in exact matching. Hence, we are relatively confident about the signs of these estimates.

These two results reinforce each other, pointing to the same important puzzle: Those working under DWA are less satisfied with their work and more likely to feel stress in daily life than those under SWA despite that DWA is also associated with lower hourly wage. This appears in direct contradiction to both the theory and the findings from related studies [e.g., Bloom *et al.,* (2015)]. We shall investigate this puzzle in the next section.

C. Flexibility and Discretion at Work

To resolve the puzzle, we start by the following observation. If workers take DWA positions at their own will despite their lower wages, they must like at least some aspects of the positions (*other than* those directly explained by the included covariates). We know that DWA legally exempts *firms* from strict monitoring of work hours, and hence, from paying for overtime wage premiums. Since DWA is based on the explicit contract between firm and

worker, DWA workers must know this at the time of choosing their positions, and hence, they should like this aspect of DWA. Yet, given what we found in the preceding subsection, there must be some other aspects of DWA that these workers must feel unsatisfied with. What are these undesirable aspects of DWA that are not part of SWA?

As one plausible explanation, we explore the following conjecture. An important premise of DWA is that its legal exemption status also translates into more flexibility and discretion at work, particularly with respect to how to allocate their work hours. What if DWA workers hold the same expectation, yet the reality does not meet their expectation? DWA workers may feel unsatisfied if they work on the same task as SWA workers do for lower hourly wages while they do not enjoy its full merit of DWA as much as they hoped for. In other words, the results of the preceding section may be picking up the *net effect* of DWA on worker utility.

Since we cannot directly observe discretion at work, we consider two survey responses as proxies: (a) hours of meeting per week and (b) the percentage of unfruitful meetings. The first variable is based on the following question: How many hours of meeting do you have on average per week at your company? The percentage of unfruitful meetings is the response to the following question: What percentage of meetings do not result in clear conclusions, outcomes, or instructions? These two variables are intended to measure related, yet different responses at work. The former allows us to evaluate if DWA changes the task content: i.e., allocation of work hours over different tasks at work. Meetings are an integral part of job responsibility, at least in the Japanese context. Hence, the changes in hours spent on meetings may be an important indicator of changes in the task content. The latter, on the other hand, is intended to evaluate the impact of DWA on the degree of flexibility and discretion at work. If workers have indeed flexibility and discretion over how/when to achieve their given task, they must be able to avoid unfruitful meetings either by not attending such meetings, by making such meetings shorter, or by making these meetings more productive. We do not necessarily expect DWA to decrease hours of meeting per se, but we do expect DWA to decrease unfruitful meetings if DWA indeed fulfills its expectation.

We apply the same hybrid method as in the previous subsections, using hours of meeting and the percents of unfruitful meetings as our outcome variables. The results are reported in Table 6, again along with estimates from OLS, PS-blocking, and CEM. As before, we prefer the estimates from our hybrid method because OLS and PS-blocking tend to pick up confounding effects of unbalanced covariates. Our results point to the same message: relative to SWA, DWA is estimated to decrease both the hours of meeting and the percent of unfruitful meetings, but the estimated impacts are not statistically highly insignificant. As shown in Figure 5, the estimates seem highly robust to alternative sets of controls used in exact matching. Interestingly, relative to OP, DWA has essentially no effect on the percent

of unfruitful meetings, in virtually all specifications.

These results seem to support our conjecture: DWA did not deliver the full merit of flexibility and discretion to workers beyond its legal exemption from strict work-hour control. DWA workers may be facing other constraints at work since the frequency of inefficient meetings is just one indicator of flexibility/discretion at work. The possibility that DWA does not quite come with productive use of time is problematic and undermines the important premise of the DWA. If this is indeed true, the advice for the Japanese government is clear: give DWA workers more flexibility and discretion at work for more labor productivity.

The results thus far are, however, not conclusive as there may be other equally plausible explanations. For example, one possibility is the self-selection of workers. Workers who have high demand for flexible work environments may be those who are sensitive to stress, and hence, have tendency to under-report work satisfaction and over-report daily stress.[16] We do not have sufficient data/information to eliminate other possible hypotheses. This may be an important direction for future research.

D. Bias and Heterogeneity in Original versus Matched Samples

The strength of our hybrid estimator is that the estimated treatment effect is more likely to be unconfounded for each subclass (or block), as we construct the subclass by directly matching on key covariates as well as estimated propensity scores using all covariates. The weakness, however, is that we tend to lose more observations. As discussed earlier, the loss of sample size can be harmful not only for the precision but also for the unbiasedness of the estimate for the *population* ATE. It can lead to bias if we lose all observations for some subclasses *and* if the treatment effects in these omitted subclasses differ from those of the retained subclasses. This is a classic problem of external validity. The same problem exists for virtually all quasi-experimental studies — we can estimate the causal effects only on the subsample of the population where natural experiments create an informative subsample; nothing can be inferred on the remaining population if the causal effects are expected to be heterogeneous. The question then is, to which part of the population our matched sample is likely to deliver a credible ATE estimate?

To begin, the matched sample consists only of full-time employees in the manufacturing or finance sector. In addition, the matches are found only in managerial/supervisory or professional/technical occupations. This is expected because the algorithm discards a subclass

[16]Note that our empirical strategy is not designed to control this type of selection on unobservable preferences. Selection on unobservables of this type is fine for our main analysis on hours of work and hourly wage because it is part of Rosen's equilibrium sorting. What matters for these outcomes is that we control unobservables that affect task content t.

if that subclass does not contain units in all treatment levels. Furthermore, we also see that matched workers are either section chiefs or division heads, but no ordinary personnel. Interestingly, matched workers manage a varying number of subordinates; some work alone while others manage a large number of subordinates. We expect the former to hold professional/technical positions and the latter to hold managerial/supervisory positions. Thus, the matched sample is likely to represent our target population of interest, full-time workers (in Japan) who conduct non-manual and non-repetitive tasks. The estimated impacts are useful for this population, but probably not for others.

Figure 6 helps us decompose the effects of matching (and hence, the sources of bias). These figures plot the OLS estimates (with 95% CIs) using the original sample (left) and the WLS estimates using only the matched sample (right) for subclasses constructed from simply tabulating observations by occupational type and rank.[17] Panel A plots the estimates on work hours and panel B on (logged) hourly wage. To focus on essentials, we only discuss the impact of DWA relative to SWA. To begin, we see the OLS estimates tend to vary across subclasses: Some are positively significant, others are negatively significant. This variability should not be taken as evidence for heterogeneous treatment effects since they are likely to be biased due to unbalanced confounding covariates. The extent of the bias can be gleaned by comparing the estimates for each subclass — we see that within *each* subclass, the estimates on the original sample differ substantially from the estimates on the matched sample. Hence, our hybrid method removes bias in two ways, by removing uninformative subclasses altogether and by removing uninformative observations within the remaining subclasses. As a result, however, we can only extrapolate our estimates to the population of workers who work in the remaining occupation types and ranks.

Interestingly, the figures also seems to indicate some important heterogeneity within the matched sample. Out of four subclasses on the matched units, three show positive impacts of DWA on work hours, but one shows a negative impact (all relative to SWA). On hourly wages, the results are reversed: Three show the negative impacts while one shows a positive effect. On both outcomes, the anomaly occurs on the subclass consisting of division heads with managerial and supervisory occupations. For this subclass, DWA is estimated to reduce work hours and increase hourly wage. Had we excluded this subclass, our ATE estimate would have been more statistically significant. The question then is, what is the source of this heterogeneity? In fact, the result may seem quite intuitive to those familiar with the Japanese labor market. In Japan, division heads in managerial positions are managers who hold authority over personnel affairs and thus discretionary power over their own work. For this

[17]Note that these subclasses are not the actual subclasses used for estimation. They are used to help us facilitate the discussion.

particular subpopulation, DWA effectively reduces work hours and increases hourly wage, possibly because it allows them to enjoy the full merit of DWA. Indeed, the supplementary analysis in the Online Appendix B shows that DWA is estimated to significantly reduce hours of meeting (relative to SWA) for this subclass, but not for others. On the other hand, in the matched sample, division heads in professional positions are directors who do not hold much discretionary power over their own work. For this subpopulation, DWA increases work hours and decreases hourly wage. The analysis in the Appendix B seems to confirm this point: DWA is estimated to significantly increase the rate of unfruitful meetings (relative to SWA) for this subclass, with no effect on hours of meeting. These heterogeneous treatment effects suggest that the economic impact of DWA may critically depend on how much flexibility and discretion come with that status. This seems also consistent with our explanation of the puzzle below. Our analysis on heterogeneous effects is limited, however, because we have a very small sample in each subclass. A larger sample might give us a more conclusive verdict on this issue.

7. Conclusion

OP regulation and WCE have been the center of a recent debate on labor market reforms in Japan. In this debate, WCE is cast as a means to ameliorate long work hours and low earnings per hour of full-time employees in Japan. Despite its economic and policy significance, little empirical evidence exists on the economic impacts of WCE (in Japan or worldwide). In addition to the price incentives created by an exemption to pay overtime pay premium, WCE is indented to provide an additional economic incentive by allowing eligible workers to work flexibly without rigid work-hour control. We also examine the impact of this margin by making use of a unique feature of the Japanese labor market, where two OP-exempt work arrangements exist: *discretionary work arrangement* (DWA) and *supervisory work arrangement* (SWA). While DWA legally exempts workers from pre-determined work schedules, SWA is only an informal work arrangement under which workers hold little discretionary power over their work schedules. Hence, DWA is taken as an official WCE status while SWA is an (unofficial) OP-exempt status.

Our empirical study relies on the cross-sectional survey we conducted in 2016 on Japanese full-time workers. The survey is designed to obtain detailed information on occupational characteristics (such as occupational rank, skill requirements, and the number of subordinates, work arrangements, hours of meeting, satisfaction from work) that are often not available in the conventional labor survey. We apply a (new) hybrid matching estimator, exploiting the unique outcome as well as pre-treatment variables from the survey. The hybrid

matching method combines the conventional subclassification estimator based on (estimated) propensity scores (Imbens-Rubin, 2015) and the coarsened exact matching developed in Iacus *et al.* (2011, 2012). The method is designed to optimize on the trade-offs between balancing covariate distributions for unbiasedness of estimates versus maintaining large enough sample size for precision of estimates. The method is simple to implement and is shown to work quite well for our sample.

Our results signify the importance of carefully delineating the two types of exemptions under WCE. We find that the OP-exempt status itself *does not* induce longer work hours or higher hourly wage. This is in sharp contrast to not only the textbook theory of labor markets but also the findings from earlier empirical studies (e.g., Kuroda and Yamamoto, 2012). Hence, our result does not support critics' argument that the OP exemption leads to long working hours. However, our result also indicates that the DWA, a legal OP-exempt status, leads to longer work hours and lower hourly wages, relative to OP and SWA arrangements. Therefore, our finding does not support the proponents of the WCE. However, these results are indeed consistent with Rosen's theory of compensating wage differentials — workers who value flexible work schedules are matched with such work arrangements for lower equilibrium wages.

Interestingly, we also find some puzzling results that are at odds with the theory. Despite that they willingly choose flexible work schedules for lower wages, DWA workers tend to feel *lower* satisfaction from work than SWA workers do. To disentangle this puzzle, we examine two measures of flexibility/discretion at work: hours of meeting per week and the rate of unfruitful meetings. Though results are not conclusive, we find some indication that DWA does not enhance flexibility/discretion at work — neither hours of meeting nor the rate of unfruitful meetings are not statistically different between DWA and SWA. Our interpretation thus far is that although DWA workers value flexible work schedules, and hence, willingly accept lower wages, they are not truly satisfied with the work arrangement since they do not enjoy the full merit of flexibility and discretion at work. Examining the matched sample further at the subclass level seems to reinforce this interpretation. Only in the subclass consisting of division heads in the managerial positions, DWA is estimated to significantly reduce hours of meeting whereas in the subclass consisting of division heads in the professional positions, DWA is estimated to significantly increase the rate of unfruitful meetings.

These results seem to imply an important policy implication for Japan and, more generally, for WCE rules around the world. Labor-law practitioners often pay attention to three eligibility tests: the salary basis test, the salary level test, and the job type test. The current debate about WCE in Japan also focuses on the salary level test — the new rule

covers salary-based workers earning more than 10.5 million yen per year. In contrast, our results signify that flexibility/discretion may be the most important factor in determining the economic consequences of WCE. The economic outcomes of WCE depend crucially on whether or not the WCE status comes with the real flexibility and authority to manage their own work, particularly as to how to complete tasks at work. If WCE comes only with the legal exemption from work schedules, it may only results in longer work hours and lower wages, without much prospect for improving labor productivity or innovation at workplace. In this sense, we concur with the sentiment offered in Bloom *et al.* (2015) — give workers more flexibility and discretion; it will likely enhance workers' satisfaction from work while improving their productivity.

References

[1] Abadie, Alberto and Guido W. Imbens. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6), 1537-1557.

[2] Abadie, Alberto and Guido W. Imbens. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2), 781-807.

[3] Barkume, Anthony. (2010). The structure of labor costs with overtime work in US jobs. *Industrial and Labor Relations Review*, 64(1), 128-142.

[4] Bell, David NF. and Robert A. Hart. (2003). Wages, hours, and overtime premia: Evidence from the British labor market. *Industrial and Labor Relations Review*, 56(3), 470-480.

[5] Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying. (2014). Does working from home work? Evidence from a Chinese experiment. *The Quarterly Journal of Economics*, 130(1), 165-218.

[6] Bubonya, Melisa, Deborah A. Cobb-Clark, and Mark Wooden. (2017). Mental health and productivity at work: Does what you do matter? *Labour Economics* 46, 150-165.

[7] Cabinet Office. (2017). The Action Plan for the Realization of Work Style Reform. http://www.kantei.go.jp/jp/singi/hatarakikata/pdf/The_Action_Plan_for_the_Realization_of_Work_Style_Reform.pdf

[8] Cattaneo, Matias D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2), 138-154.

[9] Chatterji, Pinka, Margarita Alegria, and David Takeuchi. (2011). Psychiatric disorders and labor market outcomes: Evidence from the National Comorbidity Survey-Replication. *Journal of Health Economics*, 30(5), 858-868.

[10] Costa, Dora L. (2000). Hours of work and the Fair Labor Standards Act: A study of retail and wholesale trade, 1938–1950. *Industrial and Labor Relation Review* 53(4), 648-664.

[11] Ehrenberg, Ronald G. (1971). Fringe benefits and overtime behavior. Massachusetts, Heath.

[12] Friesen, Jane. (2001). Overtime pay regulation and weekly hours of work in Canada. *Labour Economics* 8(6), 691-720.

[13] Hamermesh, Daniel S., and Stephen J Trejo. (2000). The demand for hours of labor: Direct evidence from California. *The Review of Economic Statistics* 82(1), 38-47.

[14] Horvitz, Daniel G., and Donovan J. Thompson. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47(260). 663-685.

[15] Iacus, Stefano M., Gary King, and Giuseppe Porro. (2011). Multivariate matching methods that are monotone imbalance bounding. *Journal of the American Statistical Association* 106(493), 345-361.

[16] Iacus, Stefano M., Gary King, and Giuseppe Porro. (2012). Causal inference without balance: Coarsened exact matching. *Political Analysis* 20(1), 1-24.

[17] Imbens, Guido W., and Donald B. Rubin. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge: Cambridge University Press.

[18] Kuroda, Sachiko, and Yamamoto Isamu. (2012). Impact of overtime regulations on wages and work hours. *Journal of the Japanese and International Economies* 26(2), 249-262.

[19] Lewis, Gregg H. (1969). Employer interests in employee hours of work. Mimeo, Department of Economics, University of Chicago.

[20] Ministry of Health, Labour and Welfare. (2012). General Survey on Working Conditions, https://www.mhlw.go.jp/english/database/db-l/general-survey.html

[21] Ministry of Health, Labour and Welfare. (2017). Basic Survey on Labour Unions, https://www.mhlw.go.jp/english/database/db-l/labour_unions.html

[22] Rosen, Sherwin. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* 82(1), 34-55.

[23] Rosen, Sherwin. (1985). Implicit contracts: A survey. *Journal of Economic Literature* 23(3), 1144-1175.

[24] Rosen, Sherwin. (1986). The theory of equalizing differences. *Handbook of labor economics* 1, 641-692.

[25] Skuterud, Mikal. (2007). Identifying the potential of work–sharing as a job-creation strategy. *Journal of Labor Economics* 25(2), 265-287.

[26] Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. (2015). What are we weighting for? *Journal of Human Resources* 50(2), 301-316.

[27] Stevenson, D. (2017). Thriving at work: The Stevenson/Farmer review of mental health and employers. London: Department for Work and Pensions, Department of Health and Social Care

[28] Trejo, Stephen J. (1991). The effects of overtime pay regulation on worker compensation. *The American Economic Review* 81(4), 719-740.

[29] Trejo, Stephen J. (1993). Overtime pay, overtime hours, and labor unions. *Journal of Labor Economics* 11(2), 253-278.

[30] Trejo, Stephen J. (2003). Does the statutory overtime premium discourage long workweeks? *Industrial and Labor Relations Review* 56(3), 530-551.

[31] United States Department of Labor. (2016). Overtime Final Rule: Summary of the Economic Impact Study, http://hrcsuite.com/wp-content/uploads/2017/01/Economic-Impact-OT.pdf

Figure 1. Impact of the Discretionary Work Arrangement
in Rosen's Model of Labor Market Contracts

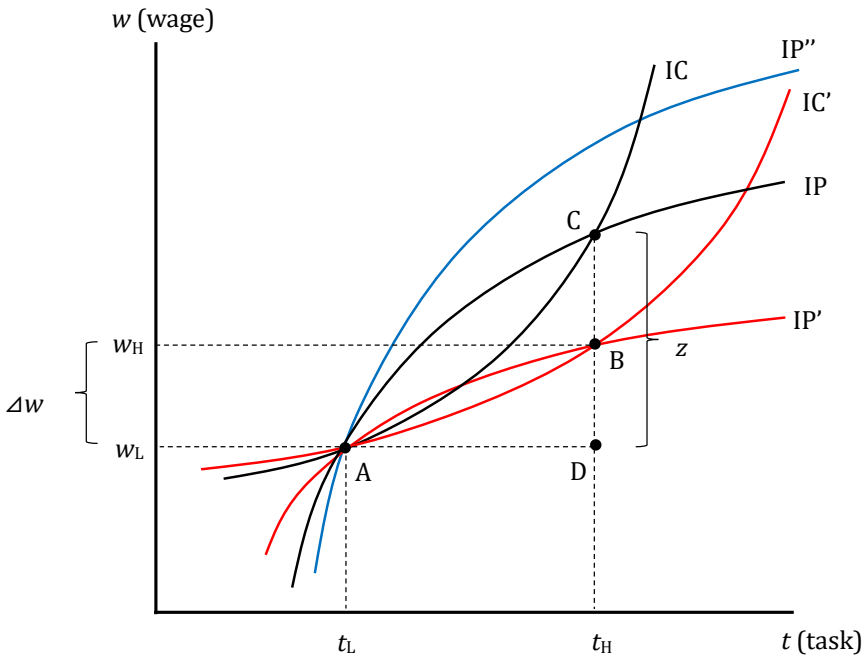


Figure 2. Trade-Off between Sample Size and Covariate Balance

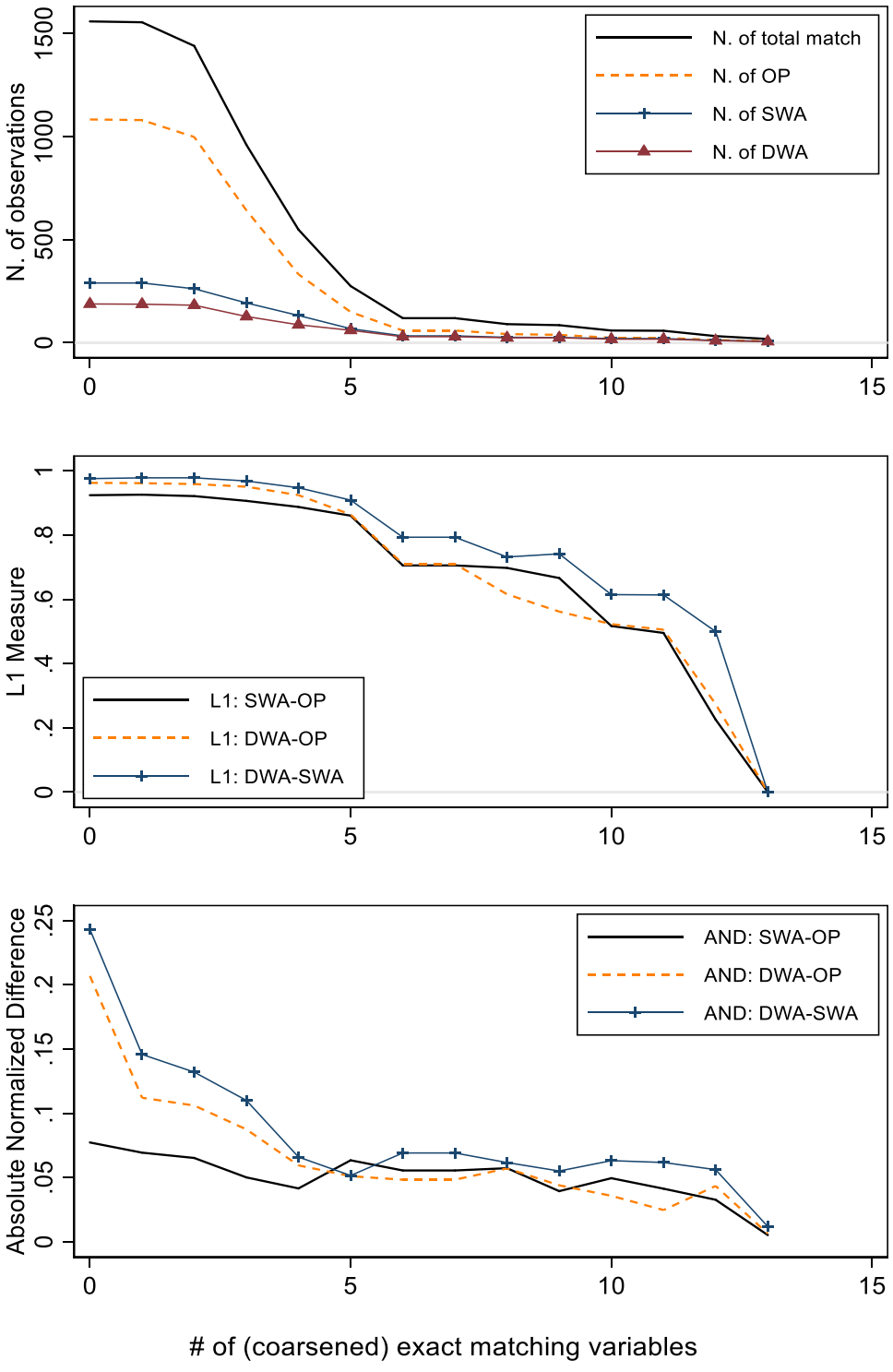
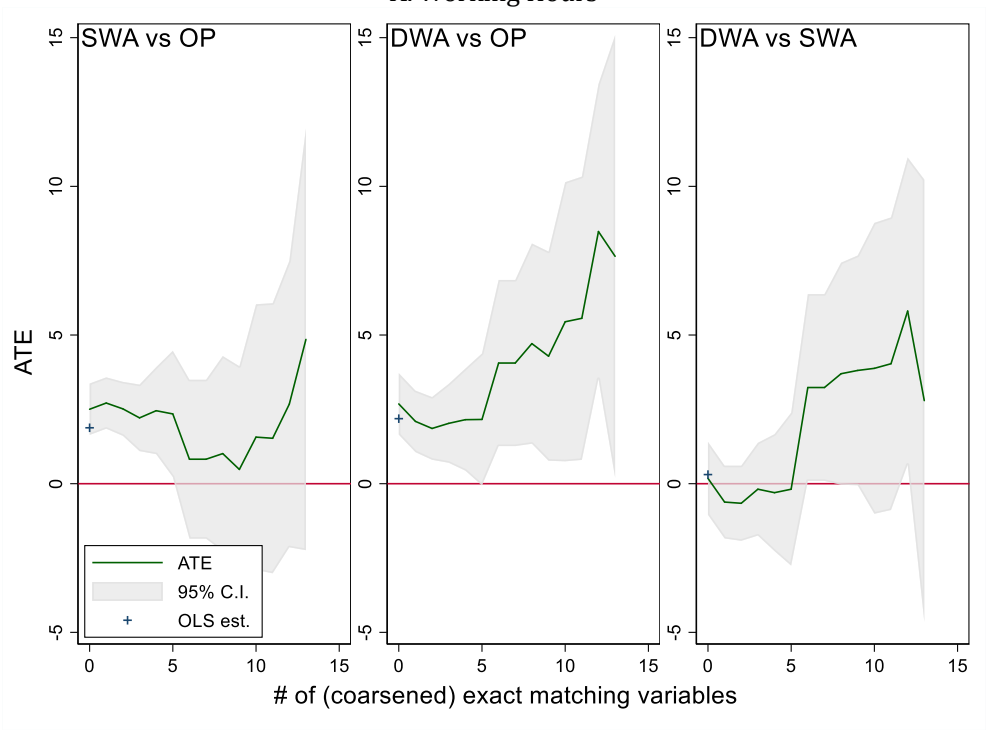


Figure 3. Effect of CEM on ATEs
Estimated Propensity Scores = 0.2 – 0.5

A. Working Hours



B. Logged Wage

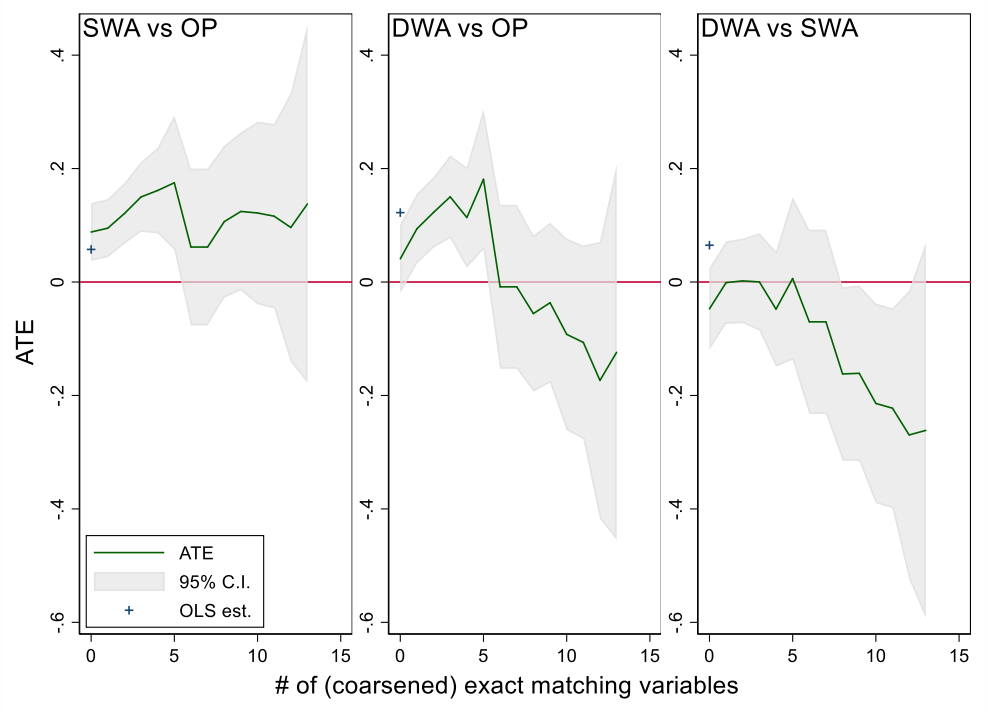


Figure 4. Effect on Happiness, Work Satisfaction, Stress
Estimated Propensity Scores = 0.2 – 0.5

A. Satisfaction from Work

B. Frequency of Stress

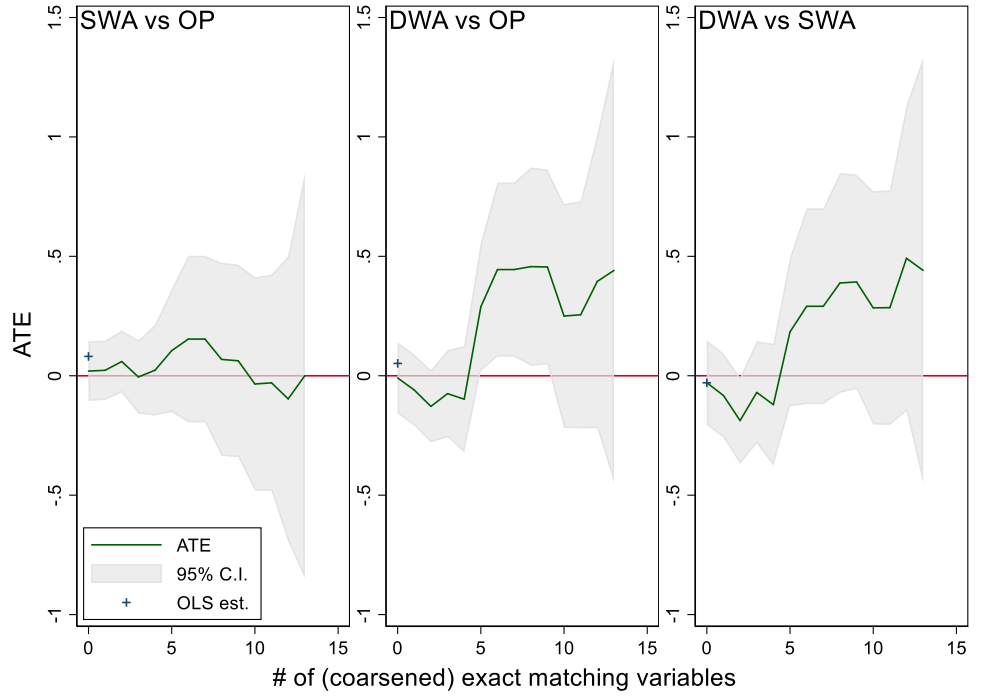
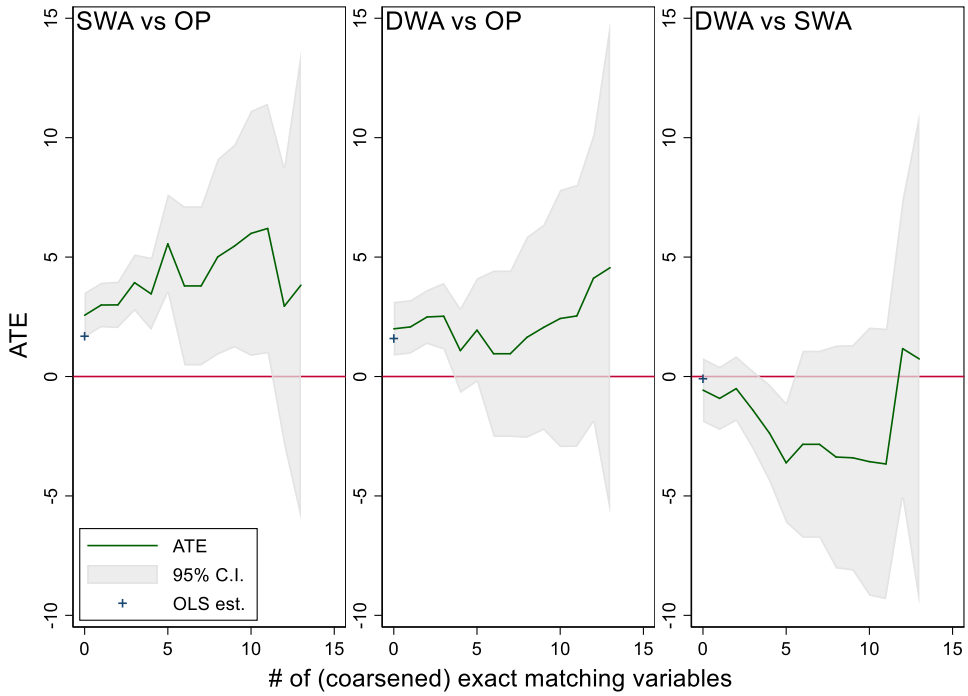


Figure 5. Effect on Work Content
Estimated Propensity Scores = 0.2 – 0.5

A. Hours of Meeting



B. Percent of Unfruitful Meeting

Figure 6. Bias and Heterogeneity in Original versus Matched Samples

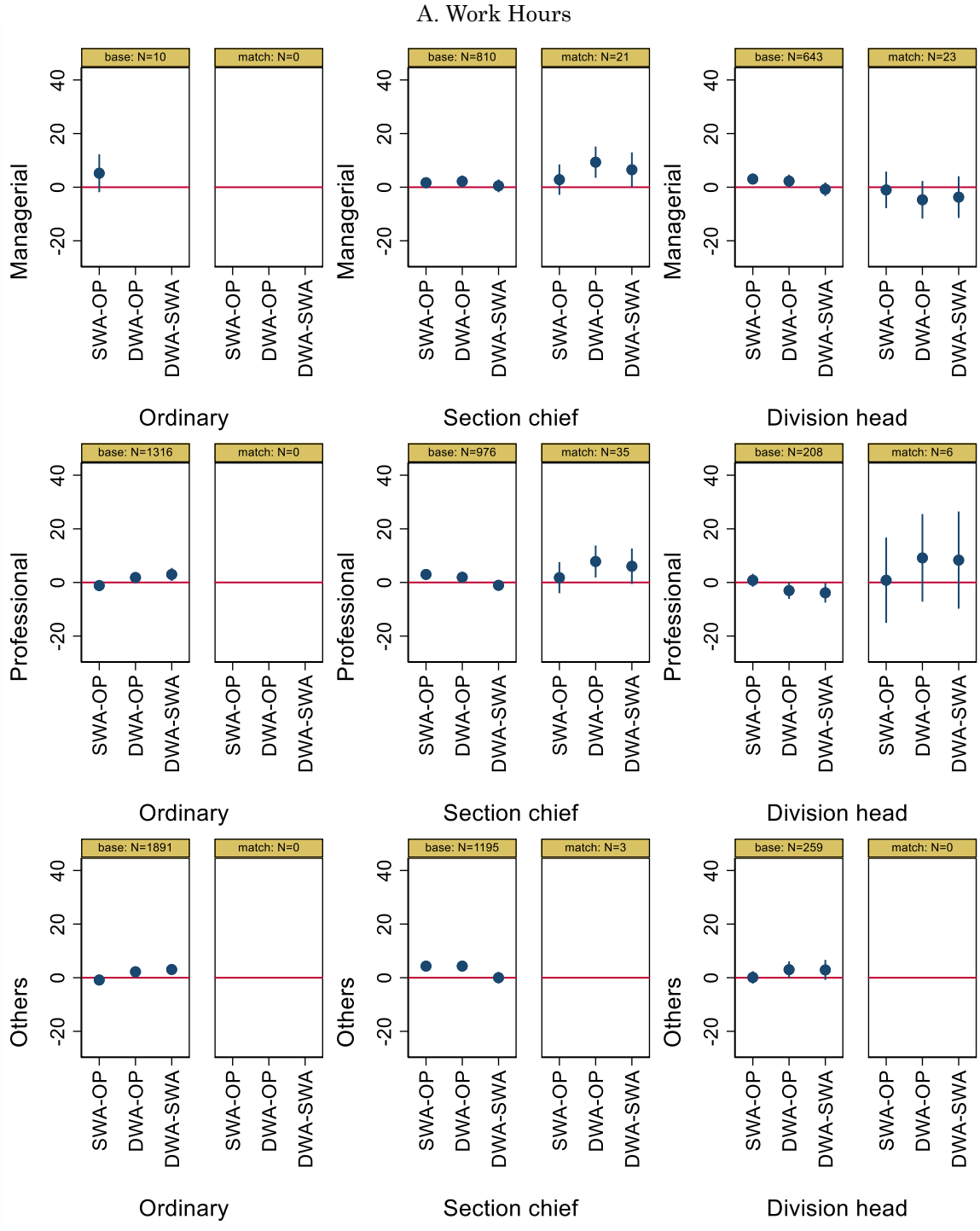


Figure 6. Bias and Heterogeneity in Original versus Matched Samples (Cont'd)

B. Hourly Wage

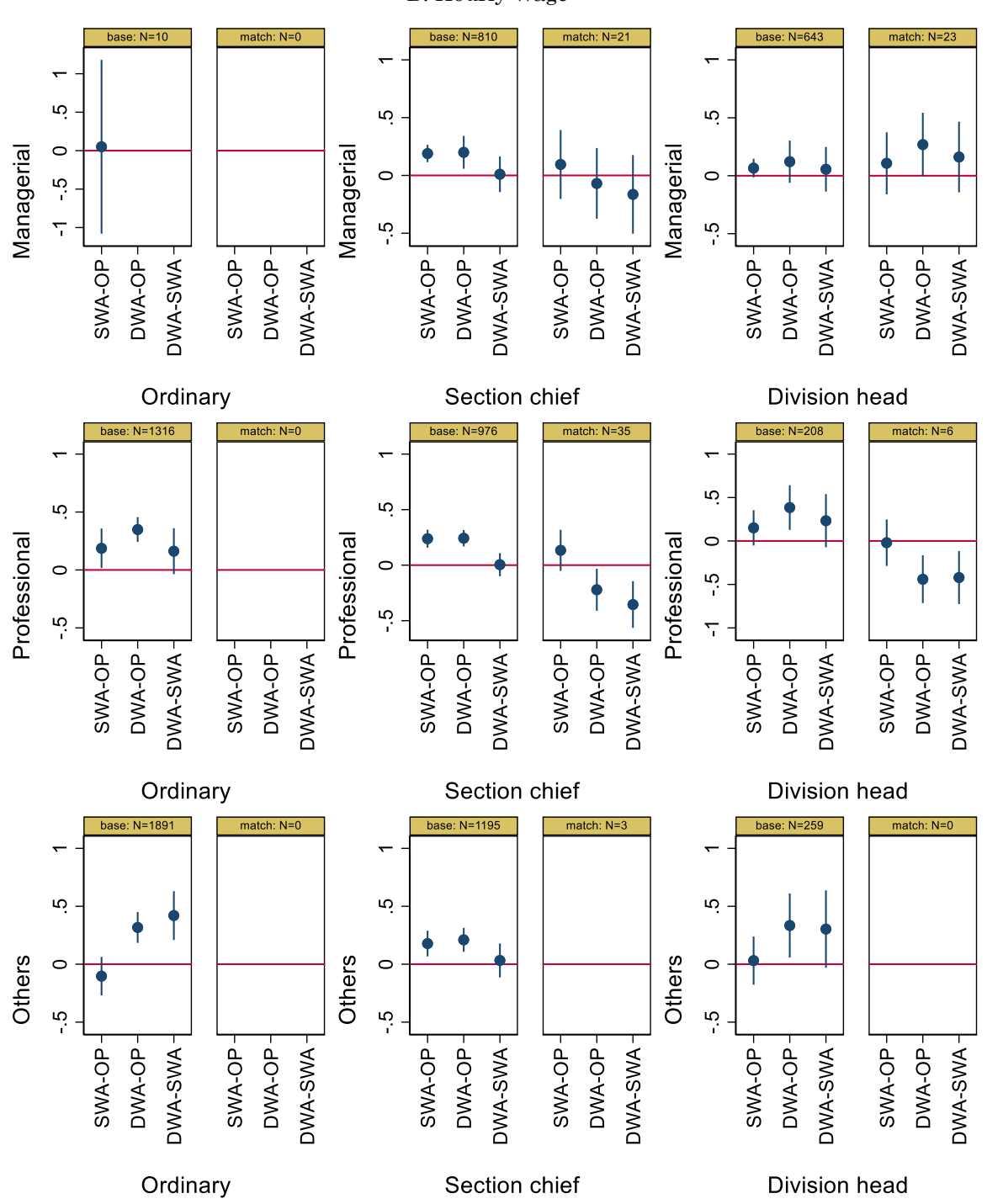


Table 1. Distribution of Employees over Alternative Work Arrangements by Industry and Occupation

	Overtime-pay Positions		No Overtime-pay Positions		Total
	Regular Hours	Flexible/Variable Hours	DWA (OP-exempt)	SWA (OP-exempt)	
All	13,483	4,357	729	1,259	19,828
	(68.0)	(22.0)	(3.7)	(6.4)	(100.0)
By industry					
Construction	1,163	119	20	106	1,408
	(82.6)	(8.5)	(1.4)	(7.5)	(100.0)
Manufacturing	3,744	1,869	305	431	6,349
	(59.0)	(29.4)	(4.8)	(6.8)	(100.0)
IT/Telecomminications	1,250	497	144	96	1,987
	(62.9)	(25.0)	(7.3)	(4.8)	(100.0)
Wholesale and Retail trade	1,450	260	37	107	1,854
	(78.2)	(14.0)	(2.0)	(5.8)	(100.0)
Fiance/Insurance	784	103	55	78	1,020
	(76.9)	(10.1)	(5.4)	(7.7)	(100.0)
Real estate	399	45	7	34	485
	(76.9)	(10.1)	(5.4)	(7.7)	(100.0)
Others	4,693	1,464	161	407	6,725
	(69.8)	(21.8)	(2.4)	(6.1)	(100.0)
By occupation					
Management/Supervisory	2,280	632	100	606	3,618
	(63.0)	(17.5)	(2.8)	(16.8)	(100.0)
Professional/Technical	3,914	1,771	329	292	6,306
	(62.1)	(28.1)	(5.2)	(4.6)	(100.0)
Clerical	4,107	815	123	144	5,189
	(79.2)	(15.7)	(2.4)	(2.8)	(100.0)
Sales	1,340	293	129	105	1,867
	(71.8)	(15.7)	(6.9)	(5.6)	(100.0)
Service	578	292	12	43	925
	(62.5)	(31.6)	(1.3)	(4.7)	(100.0)
Others	1,264	554	36	69	1,923
	(65.7)	(28.8)	(1.9)	(3.6)	(100.0)

Note: Off-site work contracts are counted toward DWA.

Table 2. Comparison of Our Sample and Employment Status Survey

	Employment Status Survey (ESS, 2017)	KHPS/JHPS (2016)	Our Sample (2016)		
			Unadjusted	Adjusted for	
				Age, Edu, Gender	Age, Edu, Gender, Occupation
# of obs.	65,092,103	2,774	21,468		
Male	0.68	0.77	0.83		
Age					
10's	0.01	0.00	0.00		
20's	0.18	0.03	0.03		
30's	0.25	0.18	0.16		
40's	0.29	0.34	0.39		
50's	0.21	0.32	0.36		
60'	0.06	0.11	0.06		
over 70's	0.01	0.02	0.00		
Education					
High school or less	0.44	0.47	0.24		
2-year College	0.19	0.13	0.16		
College	0.34	0.36	0.52		
Graduate School	0.04	0.04	0.08		
Occupation					
Administrative and managerial	0.009	0.122	0.178	0.111	
Professional and engineering	0.224	0.204	0.308	0.298	
Clerical workers	0.258	0.178	0.255	0.307	
Sales workers	0.118	0.106	0.093	0.086	
Service workers	0.068	0.073	0.048	0.065	
Other	0.323	0.317	0.118	0.132	
Industry					
Consturction	0.077	0.093	0.068	0.072	0.059
Manufacturing	0.206	0.219	0.313	0.282	0.279
IT and Communications	0.049	0.033	0.096	0.087	0.070
Wholesale and Retail Trade	0.125	0.106	0.094	0.095	0.099
Finance and Insurance	0.035	0.046	0.051	0.045	0.040
Real Estate	0.016	0.013	0.024	0.026	0.022
Other	0.493	0.490	0.353	0.392	0.430
Work Arrangements					
Overtime Pay	--	0.756	0.675	0.707	0.688
SWA	--	0.064	0.063	0.047	0.033
DWA	--	0.019	0.036	0.027	0.027
Other		0.161	0.225	0.220	0.252

Table 3. Descriptive Statistics by Work Arrangement Status

		Work Arrangement Status			Available in:	
	All	OP	SWA	DWA	ESS	JHPS/ KHPS
N. of observations	7,308	6,352	552	404		
A. Outcome Variables						
Weekly work hours	45.72	45.43	47.58	47.89	Yes	Yes
	(5.9)	(5.5)	(7.9)	(7.9)		
Hourly wage	7.87	7.82	8.21	8.14	Yes	Yes
	(.6)	(.6)	(.5)	(.5)		
B. Pre-treatment Variables						
Worker characteristics						
Age	47.44	47.18	50.67	47.11	No	Yes
	(8.1)	(8.3)	(6.2)	(7.6)		
Female	0.149	0.164	0.040	0.062	Yes	Yes
Num. of Children less than 3rd grade	0.22	0.22	0.18	0.25	Yes	Yes
	(.6)	(.6)	(.5)	(.6)		
Married	0.725	0.709	0.879	0.762	Yes	Yes
Years of Schooling	14.96	14.86	15.46	15.80	Yes	Yes
	(1.9)	(1.9)	(1.8)	(1.7)		
Job characteristics						
Occupation type					Yes	Yes
Management/Supervisory	0.200	0.178	0.498	0.136		
Professional/Technical	0.342	0.340	0.266	0.485		
Clerical	0.309	0.333	0.123	0.193		
Sales	0.138	0.138	0.105	0.178		
Service	0.011	0.011	0.007	0.007		
Job Rank					No	No†
Ordinary	0.440	0.473	0.129	0.344		
Section Chief	0.408	0.391	0.522	0.520		
Division Head	0.152	0.136	0.350	0.136		
Number of Subordinates					No	No
0 employee	0.467	0.486	0.217	0.498		
1-4 employees	0.280	0.282	0.266	0.257		
5 or more employees	0.254	0.232	0.516	0.245		

Note: Reported in the table are either frequencies or means with standard deviations in parenthesis. †: JHPS/KHPS asks whether a respondent holds a managerial position or not, but it is subject to multiple interpretations in the Japanese context.

Table 3. Descriptive Statistics by Work Arrangement Status (Cont'd)

		Work Arrangement Status			Available in:	
	All	OP	SWA	DWA	ESS	JHPS/ KHPS
N. of observations	7,308	6,352	552	404		
A. Outcome Variables						
Weekly work hours	45.72	45.43	47.58	47.89	Yes	Yes
	(5.9)	(5.5)	(7.9)	(7.9)		
Hourly wage	7.87	7.82	8.21	8.14	Yes	Yes
	(.6)	(.6)	(.5)	(.5)		
B. Pre-treatment Variables						
Job characteristics (cont'd)						
Department Size					No	No
1-4 employees	0.254	0.264	0.216	0.146		
5-9 employees	0.274	0.281	0.223	0.235		
10-19 employees	0.209	0.207	0.203	0.260		
20 or more employees	0.263	0.248	0.359	0.359		
Use of Logical Skill					No	No
0-10	0.422	0.439	0.304	0.329		
10-25	0.426	0.412	0.542	0.495		
25-	0.152	0.150	0.154	0.176		
Firm-level characteristics						
Industry type					Yes	Yes
Construction	0.125	0.132	0.118	0.037		
Manufacturing	0.419	0.404	0.516	0.532		
IT/Communcations	0.153	0.150	0.118	0.250		
Wholesale Trade/Retail Trade	0.158	0.168	0.116	0.067		
Finance/Insurance	0.099	0.098	0.105	0.101		
Real Estate	0.045	0.048	0.027	0.012		
Firm Size					Yes	Yes
1-99 employees	0.344	0.365	0.226	0.176		
100-999 employees	0.300	0.308	0.274	0.200		
1000 or more employees	0.356	0.327	0.500	0.624		
Population Density					No	No
1st Tertile	0.315	0.325	0.268	0.208		
2nd Tertile	0.332	0.331	0.341	0.329		
3rd Tertile	0.354	0.343	0.391	0.463		

Note: Reported in the table are either frequencies or means with standard deviations in parenthesis.

Table 4. Effects of Alternative Work Arrangements on Work Hours and Wage

	OLS	Blocking PS	Hybrid	CEM
Estimates of ATEs				
A. Weekly Working Hours				
SWA - OP	1.8804 ***	2.5036 ***	1.0122	4.8529
	(0.2667)	(0.4413)	(1.6732)	(3.6166)
DWA - OP	2.1872 ***	2.6835 ***	4.7107 ***	7.6471 *
	(0.3038)	(0.5278)	(1.7178)	(3.7932)
DWA - SWA	0.3068	0.1799	3.6985 *	2.7941
	(0.3871)	(0.6255)	(1.9052)	(3.7932)
B. Logged Wage				
SWA - OP	0.0575 ***	0.0881 ***	0.1064	0.1377
	(0.0181)	(0.0258)	(0.0683)	(0.1606)
DWA - OP	0.1224 ***	0.0409	-0.0557	-0.1239
	(0.0207)	(0.0309)	(0.0701)	(0.1684)
DWA - SWA	0.0649 **	-0.0472	-0.1621 **	-0.2615 †
	(0.0263)	(0.0366)	(0.0778)	(0.1684)
L1-measure				
SWA - OP	0.983	0.925	0.698	0.000
DWA - OP	0.992	0.963	0.616	0.000
SWA - DWA	0.982	0.976	0.732	0.000
Avg. Normal Distance				
SWA - OP	0.383	0.077	0.057	0.005
DWA - OP	0.306	0.207	0.057	0.006
SWA - DWA	0.272	0.243	0.062	0.012
N. of obs.	7,308	1,559	89	17

Note: In parenthesis are standard errors from WLS regression. Observations are restricted to those with estimated propensity scores = 0.2 - 0.5. Covariates used for exact matching in the hybrid method are: occupational rank, the number of subordinates, industry type, occupational type, use of logical skill on the job, age, gender, the number of children in 3rd grade or younger.

Table 5. Effects of Alternative Work Arrangements on Satisfaction and Stress

	OLS	Blocking PS	Hybrid	CEM
Estimates of ATEs				
A. Satisfaction from Work				
SWA − OP	0.1119 **	0.2171 ***	-0.1416	0.1176
	(0.0565)	(0.0775)	(0.2852)	(0.6619)
DWA − OP	0.0249	0.1084	-0.3625	-0.4706
	(0.0644)	(0.0927)	(0.2928)	(0.6942)
DWA - SWA	-0.0869	-0.1086	-0.2210	-0.5882
	(0.0821)	(0.1099)	(0.3247)	(0.6942)
B. Frequency of Stress				
SWA − OP	0.0811 *	0.0197	0.0684	0.0000
	(0.0436)	(0.0636)	(0.2059)	(0.4317)
DWA − OP	0.0518	-0.0091	0.4570 **	0.4412
	(0.0497)	(0.0761)	(0.2114)	(0.4528)
DWA - SWA	-0.0294	-0.0288	0.3886 †	0.4412
	(0.0633)	(0.0902)	(0.2345)	(0.4528)
N. of obs.	7308	1559	89	17

Note: In parenthesis are standard errors from WLS regression. Observations are restricted to those with estimated propensity scores = 0.2 - 0.5. Covariates used for exact matching in the hybrid method are: occupational rank, the number of subordinates, industry type, occupational type, use of logical skill on the job, age, gender, the number of children in 3rd grade or younger.

Table 6. Effects of Alternative Work Arrangements on Use of Time

	OLS	Blocking PS	Hybrid	CEM
Estimates of ATEs				
A. Hours of Meeting				
SWA − OP	1.6864 ***	2.5656 ***	5.0094 **	3.8235
	(0.2269)	(0.4792)	(2.0892)	(4.9939)
DWA − OP	1.5926 ***	1.9974 ***	1.6386	4.5588
	(0.2585)	(0.5731)	(2.1449)	(5.2377)
DWA - SWA	-0.0939	-0.5682	-3.3708	0.7353
	(0.3294)	(0.6792)	(2.3788)	(5.2377)
B. Percent of Unfruitful Meetings				
SWA − OP	-2.5013 *	-2.3377 †	6.9288	-2.9412
	(1.2928)	(1.6190)	(6.2432)	(13.2662)
DWA − OP	1.3391	1.1899	2.8371	-7.6471
	(1.4727)	(1.9364)	(6.4096)	(13.9137)
DWA - SWA	3.8405 **	3.5276 †	-4.0918	-4.7059
	(1.8768)	(2.2949)	(7.1085)	(13.9137)
N. of obs.	7,308	1,559	89	17

Note: In parenthesis are standard errors from WLS regression. Observations are restricted to those with estimated propensity scores = 0.2 - 0.5. Covariates used for exact matching in the hybrid method are: occupational rank, the number of subordinates, industry type, occupational type, use of logical skill on the job, age, gender, the number of children in 3rd grade or younger.

Online Appendix

Appendix A: Definition of Variables

The appendix explains the variables used in estimation.

Weekly working hour: Hours of work per day are obtained from the question asking each respondent what time s/he starts and finishes working on a typical day. We multiply this number by 5 to obtain weekly working hour.

Hourly wage: Annual before-tax incomes are reported with an interval of 1 million yen from 2 million yen up to 10 million yen, and then 10 to 15 million yen, 15 to 20 million yen, and 20 to 30 million yen. We compute an hourly wage for each observation by dividing the mid-point of income interval by weekly work hours times 48 (= 4 weeks $\times$ 12 months).

Satisfaction from work: The survey asks "How satisfied are you currently with your work?" as of the year of 2015. The possible answers are a 5-point scale from 1 "not at all satisfied" to 5 "very satisfied". Then, the survey asks whether a response's satisfaction with his/her work increases, decrease, or does not change from the previous year. We add 1 scale point to the satisfaction scale as of 2015 if the answer to this question is increase, and subtract 1 if decrease. The resulting scale of satisfaction is from 0 to 6.

Frequency of stress: The variable is based on the two questions: How often do you feel high level of stress from daily life? and How often do you feel low level of stress from daily life? with a possible answer of a 5-point scale from 1 "none" to 5 "very often". We generate the frequency of stress to assigning 0.75 to the scale from the first question and 0.25 the second question.

Hours of meeting: The survey asks "How many hours of meeting do you have on average per week at your company?" Possible responses are "0 hour", "1 to 5 hours", "5 to 10 hours", "10 to 20 hours", "20 to 40 hours", or "40 hours or longer". We use the mid-point of the chosen interval as hours of meeting per week except for the last response. We assign 45 hours for the last response.

Percent of unfruitful meetings: The survey asks "What percentage of meetings do not result in clear conclusions, outcomes, or instructions?" Respondents are asked to report a percentage in integers between 0 and 100.

Age: A respondent's age is measured in years as of December 2016.

Female: A binary variable equals 1 if an observation is female and equals 0 otherwise.

Number of Children in 3rd grade or younger: The variable indicates the number of children in 3rd grade or younger in a household.

Married: A binary variable equals 1 if an observation has a spouse and equals 0 otherwise.

Years of schooling: The survey asks the highest education attainment. We generate a numerical variable of years of schooling by assigning 6 years to "elementary school", 9 years to "junior high school", 12 years to "high school", and 14 years to "vocational school" and "junior college", 16 years to "college", and 18 years to "graduate school".

Occupation type: This categorical variables follows Japan Standard Occupational Classification. The working sample in this study includes observations working as "Administrative and managerial workers", "Professional and engineering workers", "Clerical workers", "Sales workers", and "Service workers".

Job rank: The variable indicates a respondent's job rank. In the survey, possible choices are "Chairman/President (*Kaicho/Shacho*)", "Managing Director (*Senmu/Jomu*)", "Division Head (*Bucho*)", "Deputy head (*Jicho*)", "Section chief (*Kacho*)", "Subsection chief (*Kakaricho*)", "Ordinary (*Ippan-shain*)", and "Other". We exclude "Chairman/President", "Managing Director", and "Other" from the analysis. We combine "Deputy head", "Section chief", and "Subsection chief" into a single category "Section chief".

Number of Subordinates: The variable indicates the interval of the number of subordinates that a respondent supervises. In our survey, possible choices are "0", "1", "2 to 4", "5 to 9", "10 to 19", and "20 or 29", "30 to 49", "50 to 99", "100 or more", and "Not belong to any department, or there is no department". In the analysis, we re-categorize these into "0", "1 to 4", and "5 or more". We treat "Not belong to any department, or there is no department" as "0".

Department size: The variable indicates the interval of the number of employees at a department a respondent belongs to. In the survey, possible choices are "1", "2 to 4", "5 to 9", "10 to 19", "20 to 29", "30 to 49", "50 to 99", "100 or more", and "Not belong to any department, or there is no department". In the analysis, we re-categorize these into "1 to 4", "5 to 9", "10 to 19", and "20 or more". We treat "Not belong to any department, or there is no department" as "1 to 4".

Use of logical skill: The survey asked a respondent to allocate 100 points to five types of skills that s/he finds necessary to complete his/her usual tasks at work. The four categories are: logical reasoning, interpersonal, leadership, creative, and paperwork. A preliminary analysis suggests that point allocations over these skills are highly correlated. We use only the logical reasoning skill for ease of interpretation as well as to avoid multicollinearity. We break down into the intervals as "0 to 10", "11 to 25", and "26 and more".

Industry type: This categorical variable follows Japan Standard Industrial Classification. The working sample in this study includes observations working in "Construction", "Manufacturing", "IT and Communications", "Wholesale and Retail Trade", "Finance and Insurance", and "Real Estate and Goods Rental and Leasing".

Firm size: The variable indicates the interval of the total number of employees at a firm a respondent works for. In our survey, possible choices are "1", "2 to 4", "5 to 9", "10 to 19", "20 to 29", "30 to 49", "50 to 99", "100 to 299", "300 to 499", "500 to 999", "1000 or more", and "Don't know". In the analysis, we re-categorize these into "1 to 99", "100 to 999", and "1000 or more". We treat "Don't know" as missing.

Population density: This variable is the population density of municipality of residence. We then divide the sample into three groups by the tertiles.

Appendix B. Heterogeneous Effects on Hours of Meeting and Unfruitful Meetings

Figure B1. Hours of Meeting

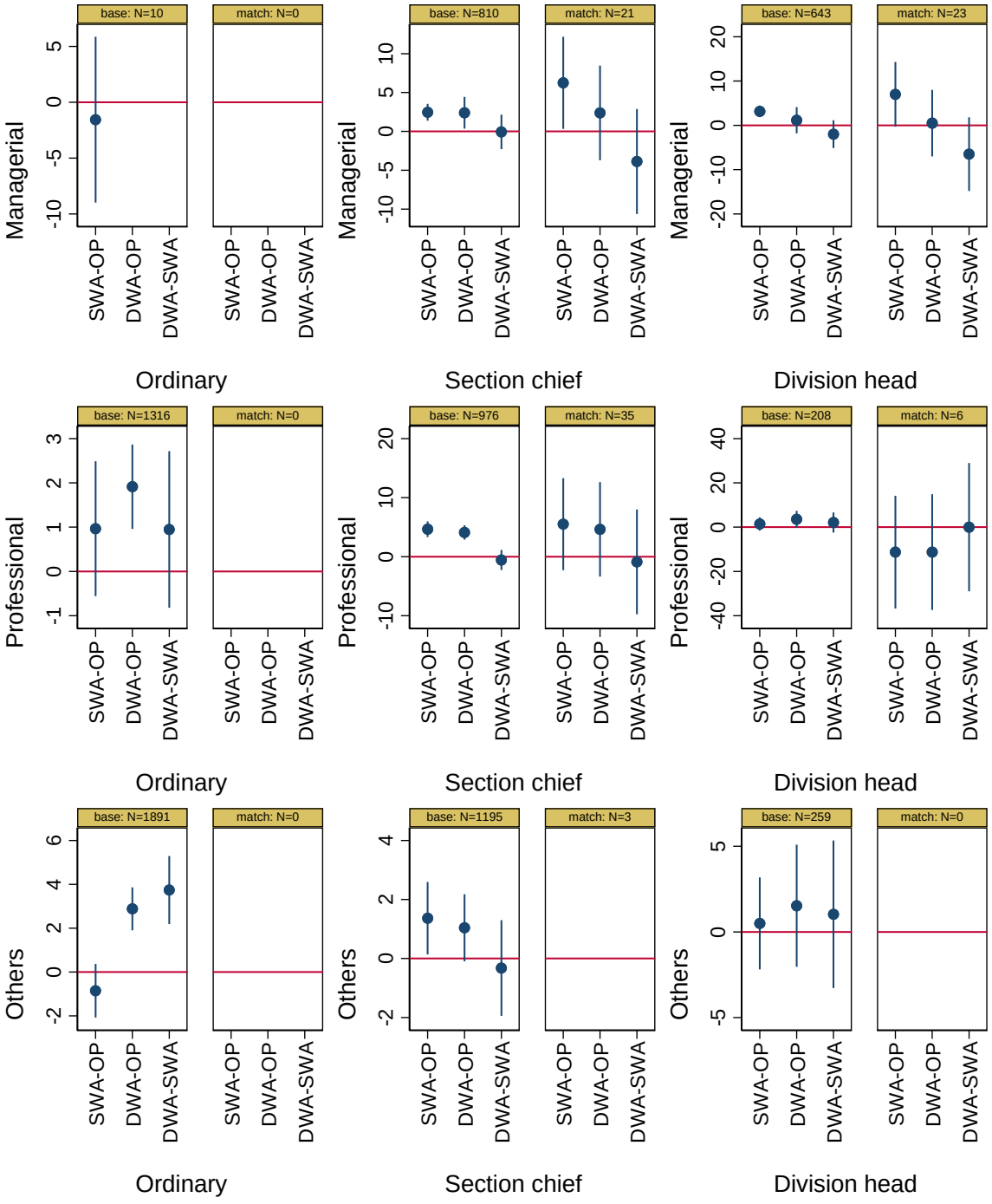
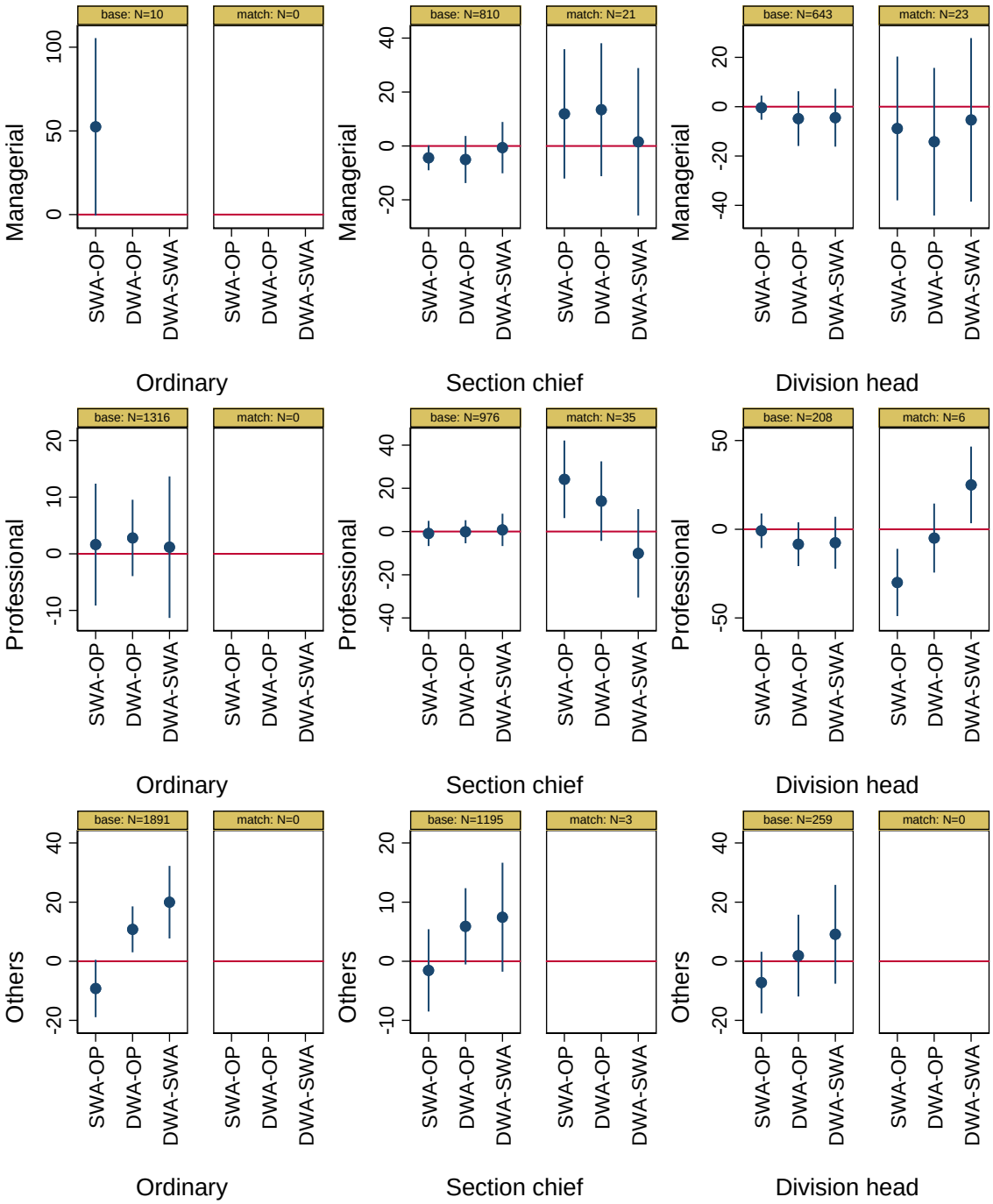


Figure B2. Percent of Unfruitful Meetings



Appendix C. Results of L1-minimizing Algorithm

Figure C1. Trade-Off between Sample Size and Covariate Balance

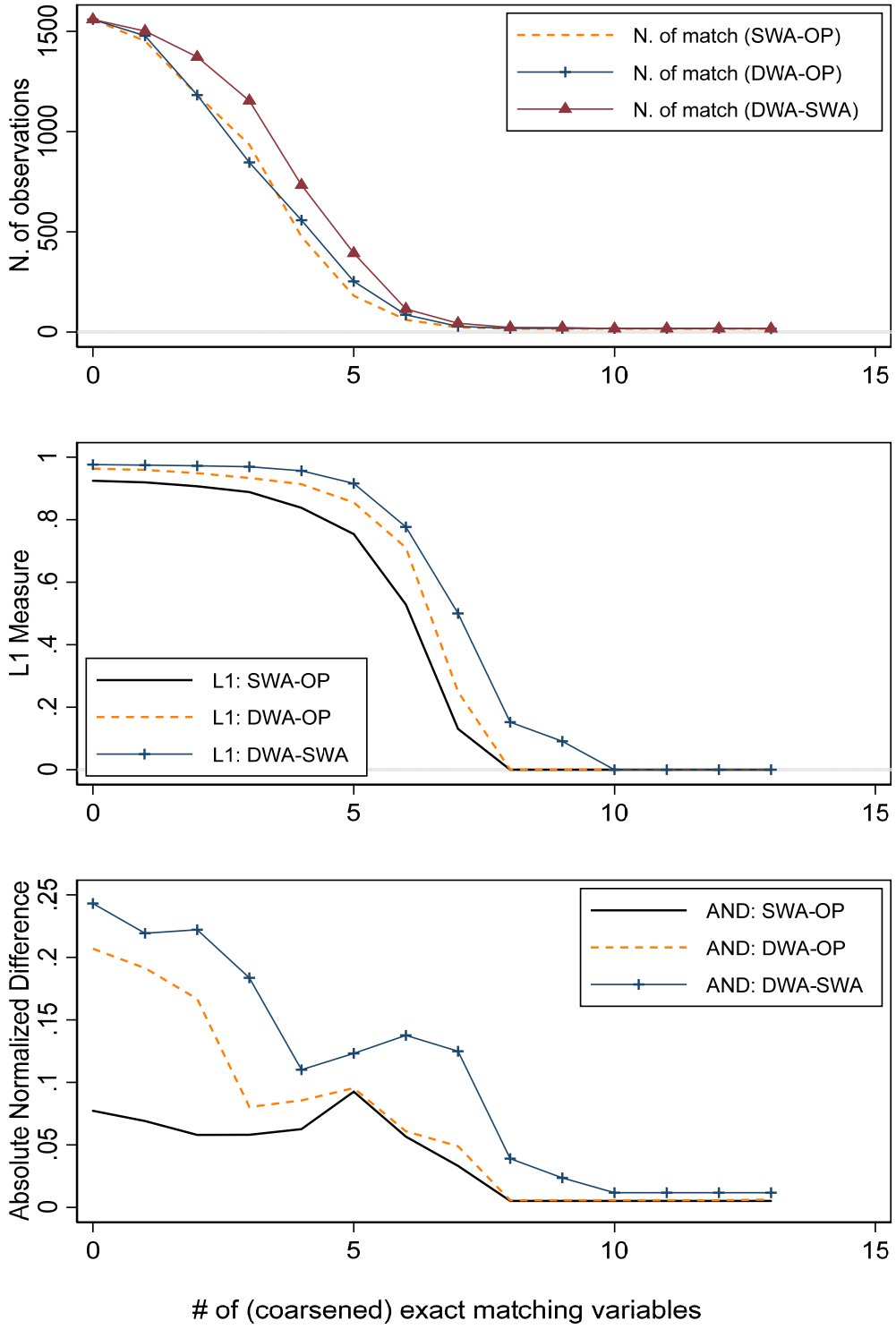
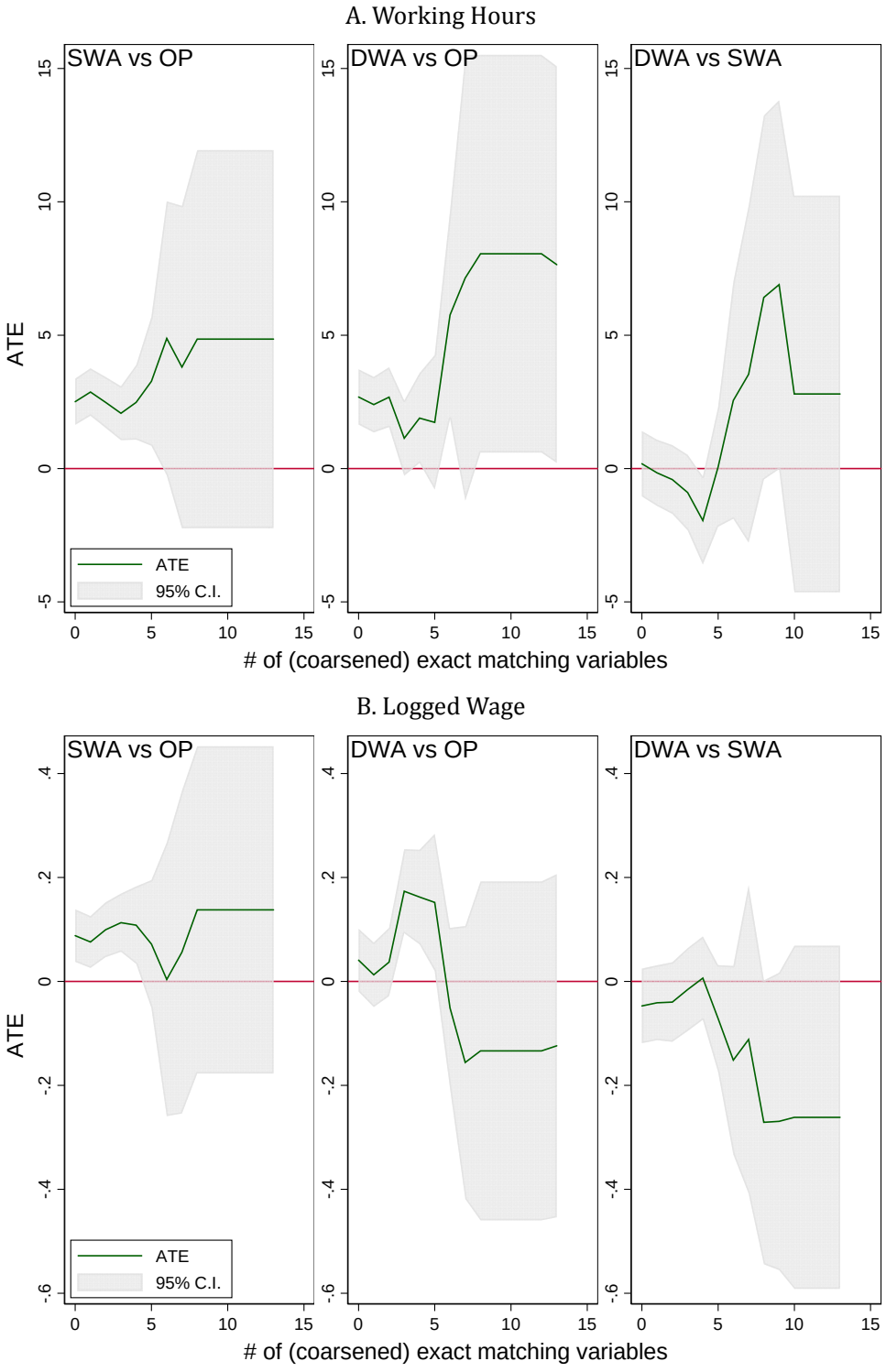
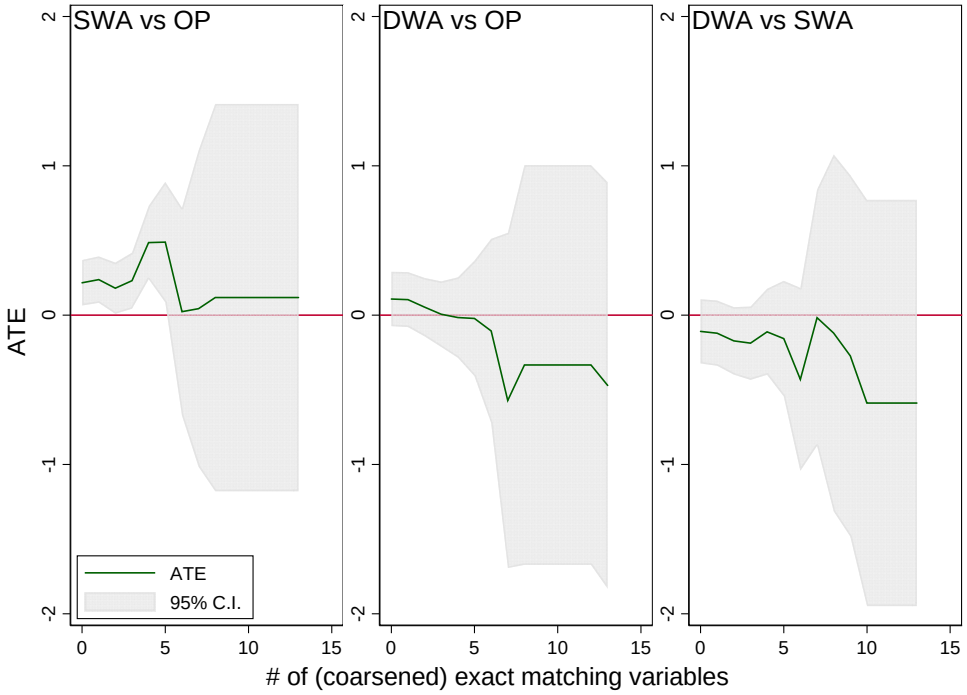


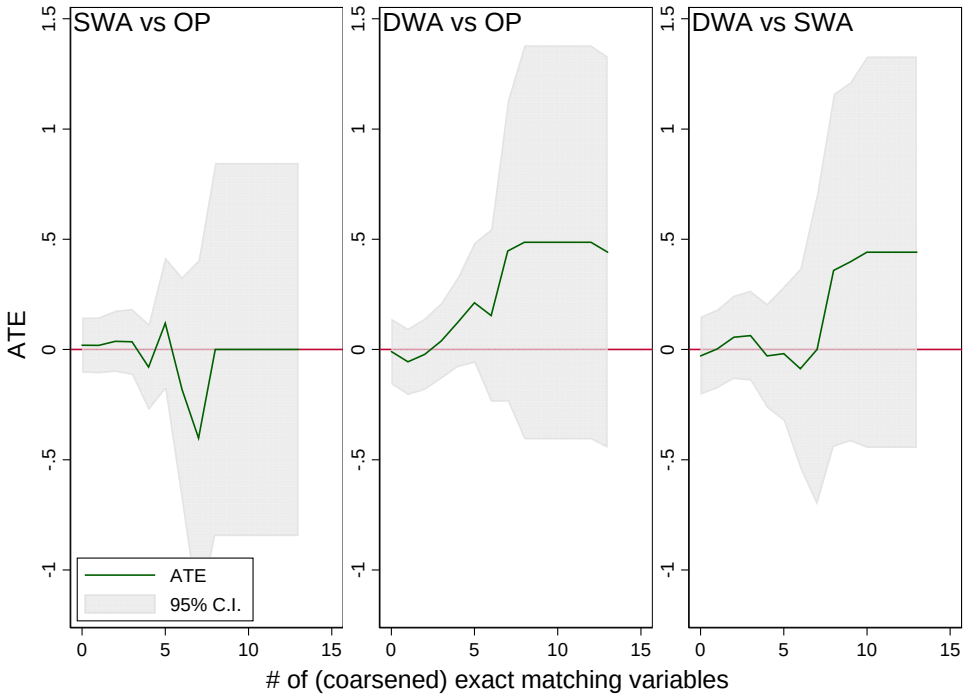
Figure C2. Effect of CEM on ATEs



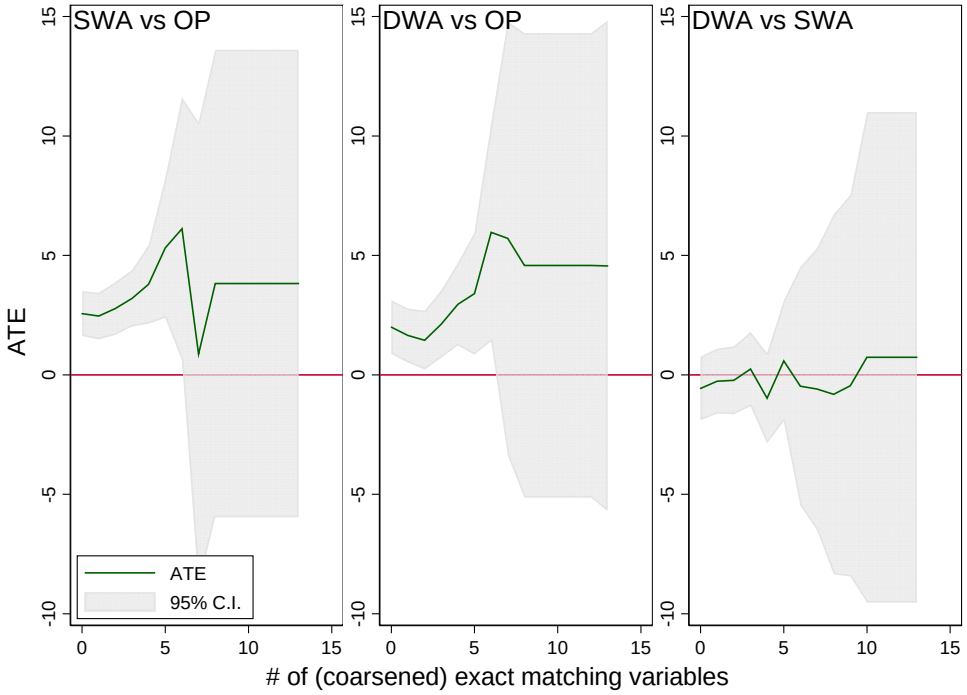
C. Satisfaction from Work



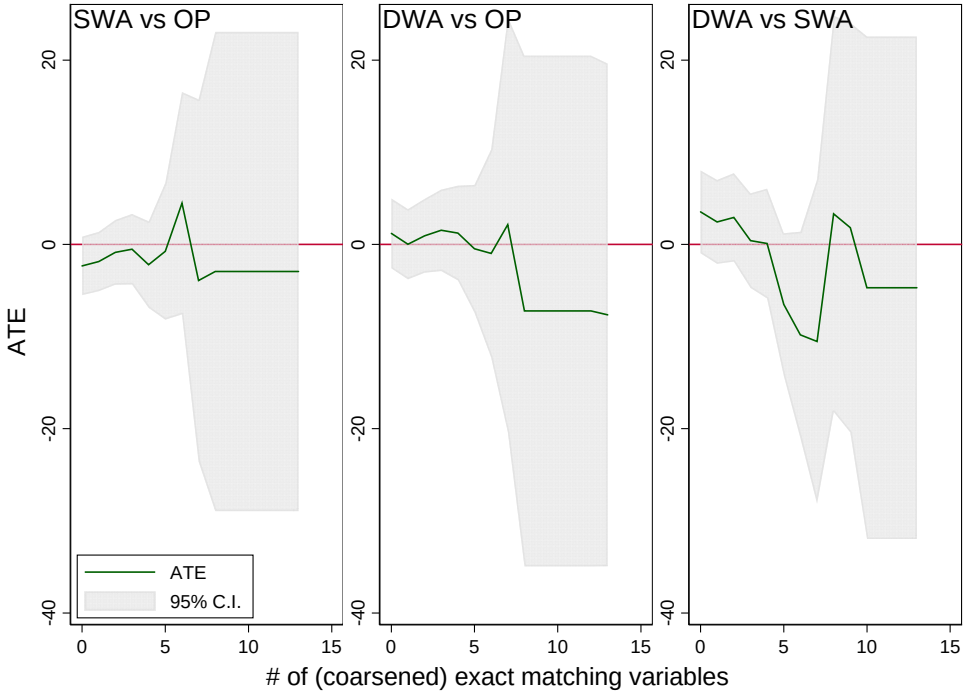
D. Frequency of Stress



E. Hours of Meeting



F. Percent of Fruitless Meeting



Appendix D. Results on the Sample with Propensity Score = 0.25-0.45

Figure D1. Trade-Off between Sample Size and Covariate Balance

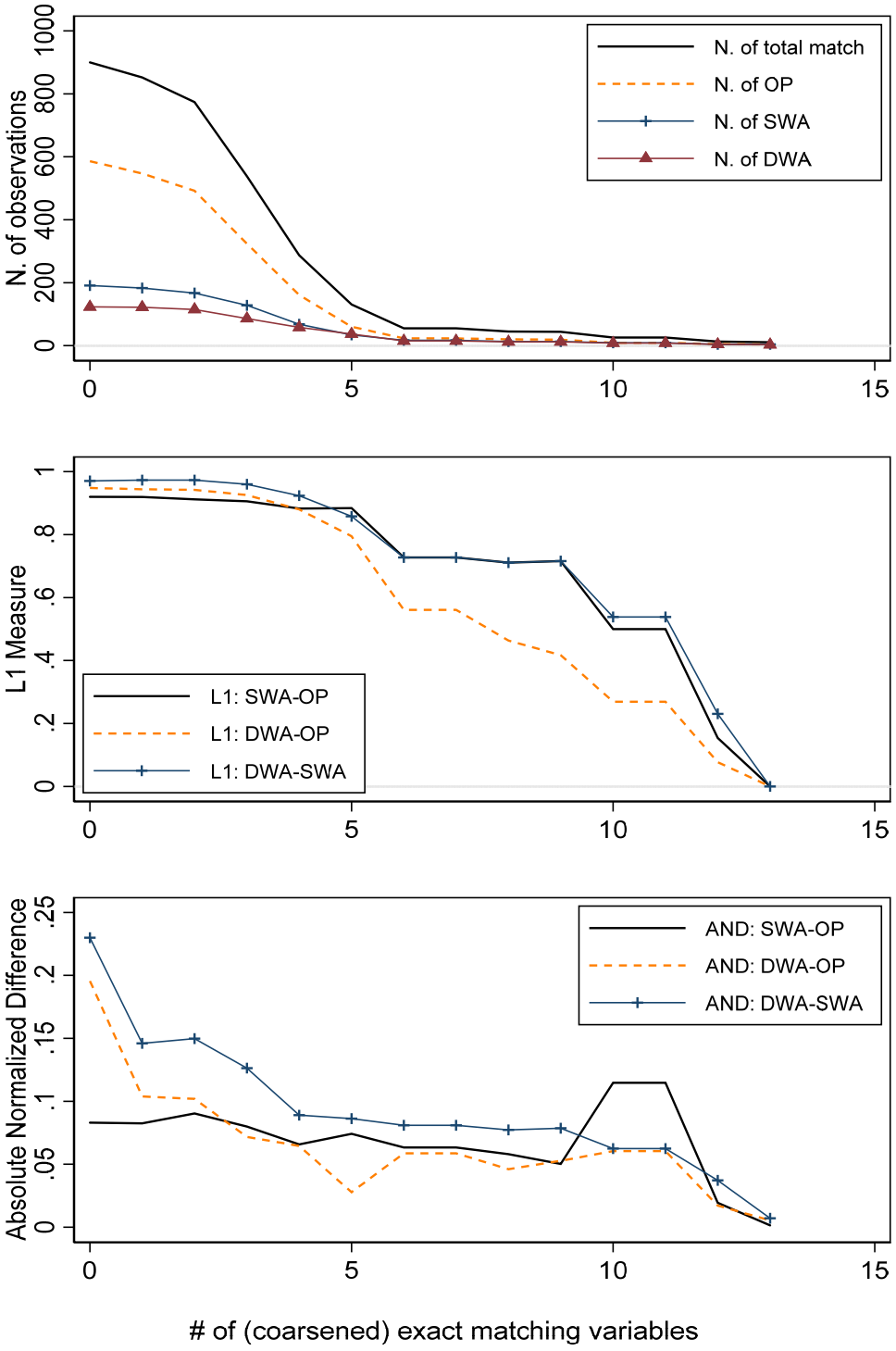
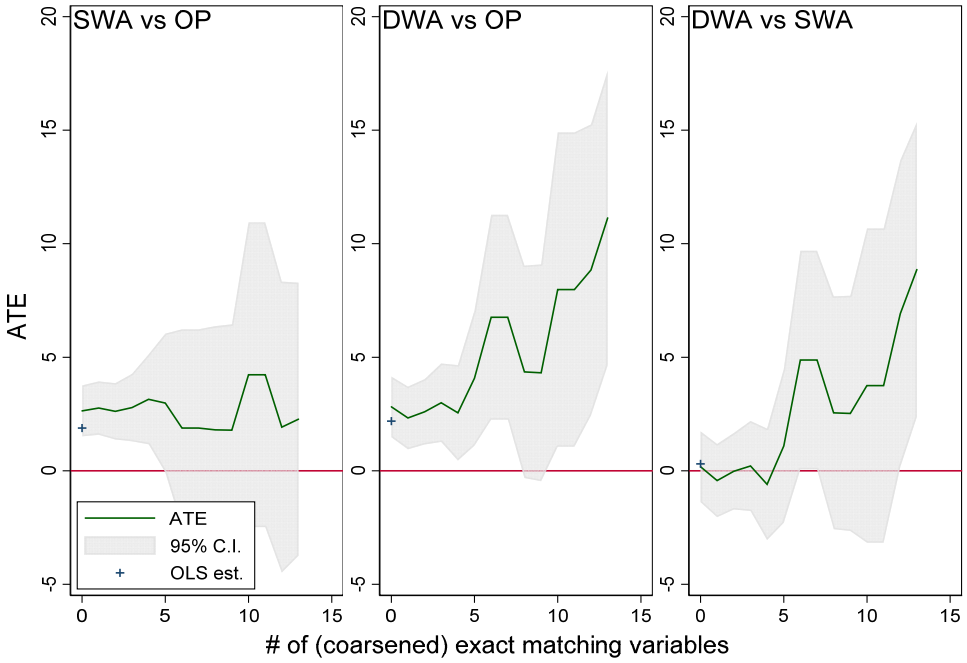
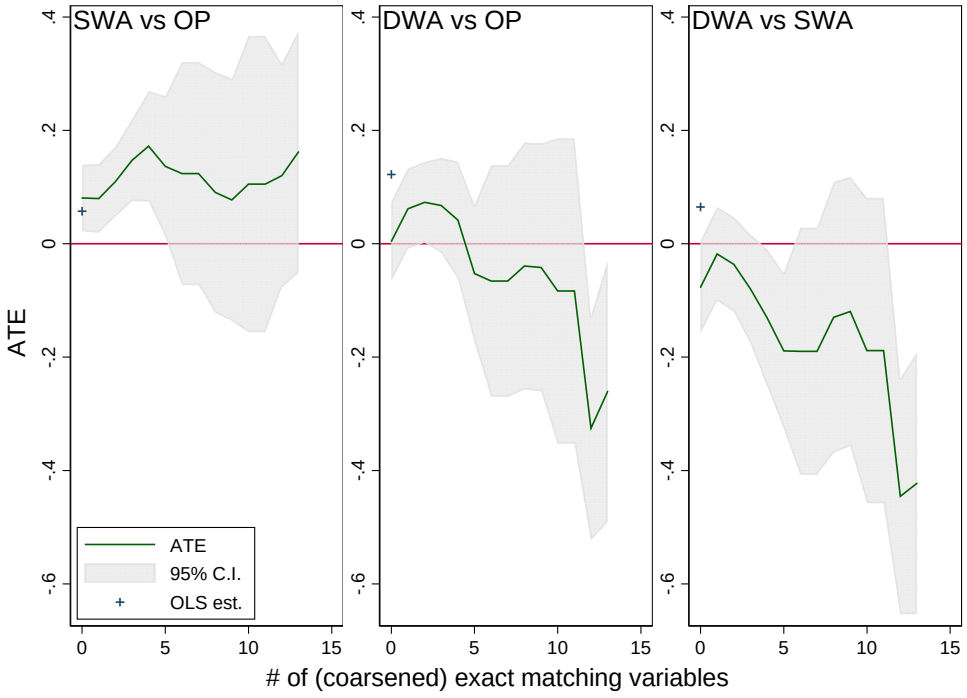


Figure D2. Effect of CEM on ATEs

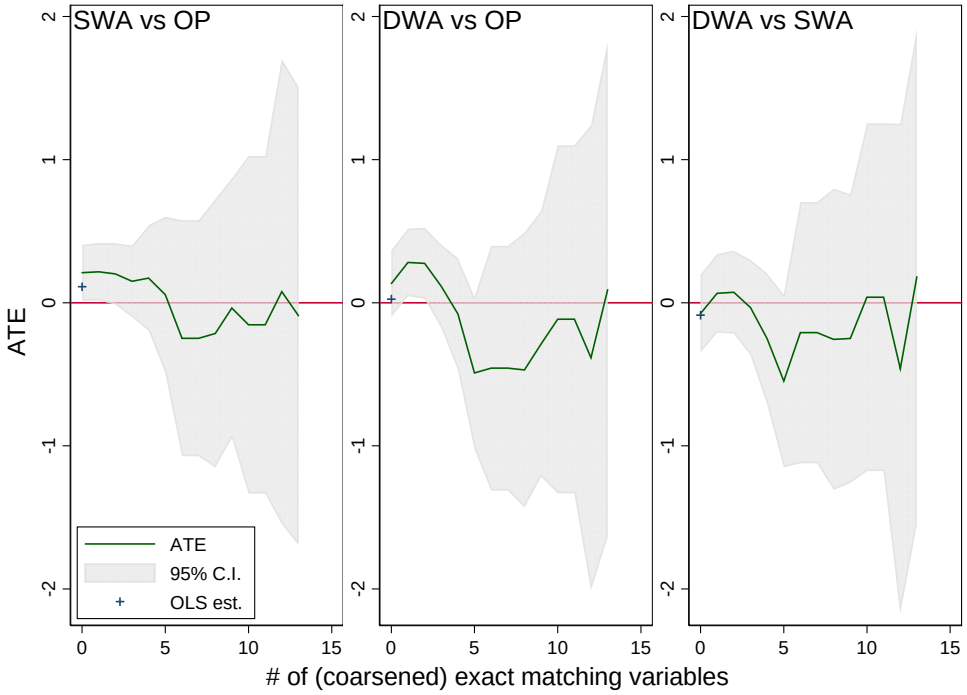
A. Working Hours



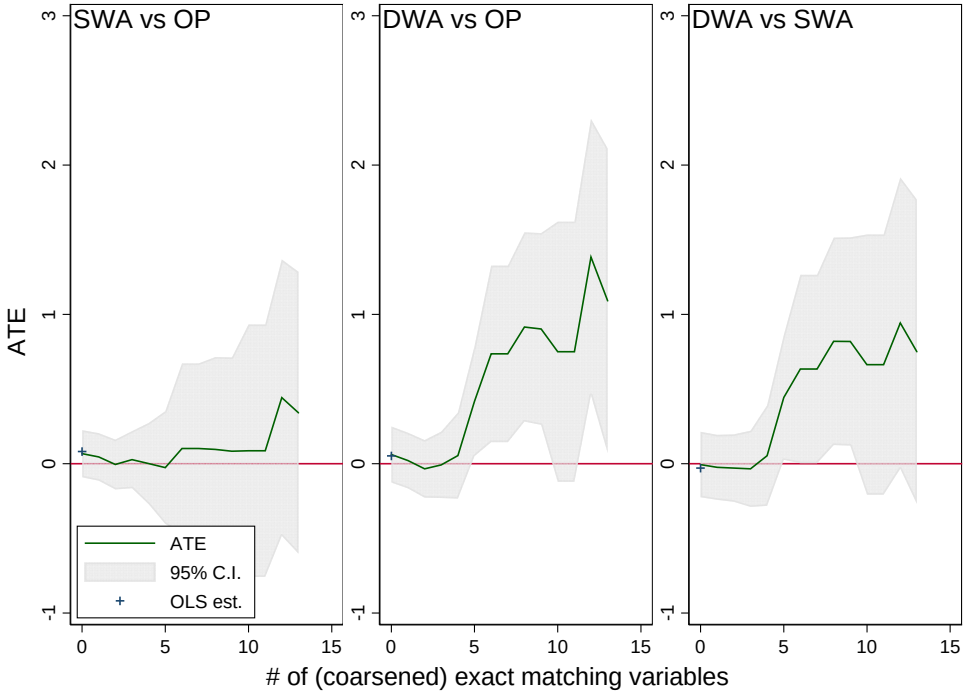
B. Logged Wage



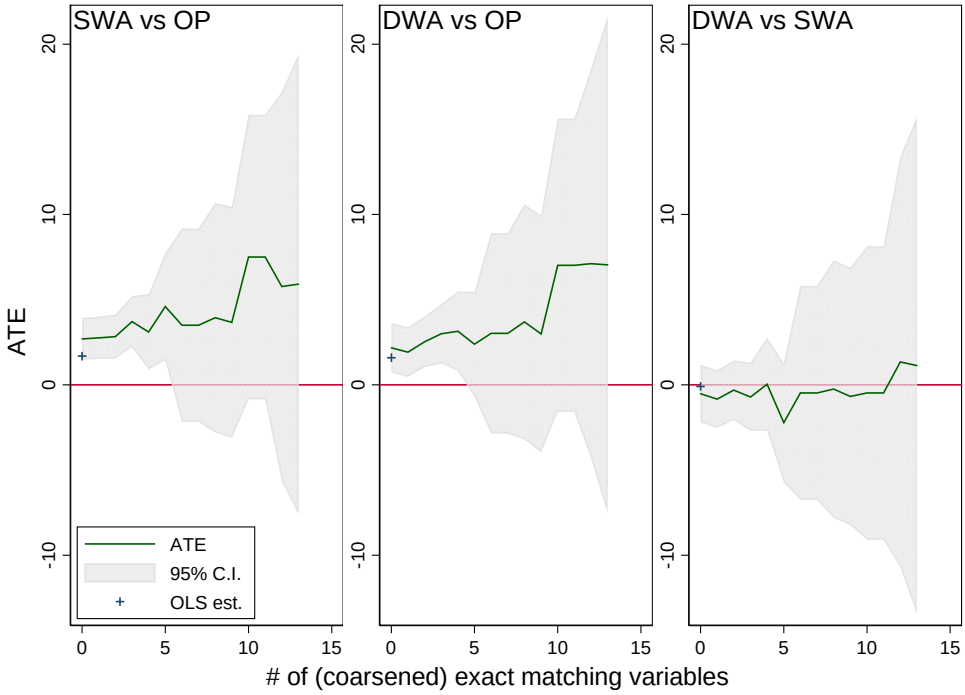
C. Satisfaction from Work



D. Frequency of Stress



E. Hours of Meeting



F. Percent of Fruitless Meeting