Measuring ACME Manufacturing's Career 2030 Training Program

Group Project 1: A/B Testing + Observational Study

Team Member

Ella Lee
Valerie Chan
Irene Wang
Bradley Ge
Jim Tiao

Executive Summary

This project aims to utilize various causal inference techniques to mitigate biases in an observational study and investigate the causal relationship between the implementation of the ACME's Career Training Program and employees' promotion rates.

To adjust for the confounding effect and the imbalanced group distribution in the dataset, the team applied three matching methods and compared the performance of balance based on the standardized mean differences (SMD) metric. Both Propensity Score Matching and Inverse Probability of Treatment Weighting achieved satisfactory balance across most covariates and maintained a significantly large sample size. However, the greater practicality and better weight performance of the former one ultimately persuaded our choice.

Using the dataset balanced by the Propensity Score Matching, the team further employed the Instrumental Variable Analysis technique to address the endogeneity and unobserved confounding effects. According to the results of the two-stage regression analysis, it is evident that the instrumental variable identified, or the distance from employees' homes to the training facility, has a negative correlation with the participation rate of the training program. Additionally, the analysis confirmed that there is a statistically significant positive causal relationship between employees attending the training program and getting promoted.

To maximize the impact of the career training program, we suggest prioritizing initiatives aimed at enhancing access and refining content. Implementing online training solutions or providing transportation support can mitigate geographical barriers to participation. Additionally, focusing on soft skills development, particularly in networking and communication, can significantly improve employees' promotion prospects. Tailoring content to engage employees with higher test scores is also crucial. These recommendations can offer greater career advancement opportunities to ACME employees.

Introduction

ACME initiated the Career 2030 training initiative, focused on boosting employee promotions and retention. A total of 6,000 ACME employees participated in this program, with a split between those who received training and those who did not, representing 10% of the company's total employee base. The company seeks to analyze the impact of the program on its workforce.

Our objectives are twofold: (1) to assess the effectiveness of the training sessions, and (2) to formulate strategic recommendations. First, we want to determine the efficiency of the training sessions, eliminating any confounding variables to ensure that observed outcomes are truly attributable to the training. If we establish that the training is effective, we will then explore strategic approaches to potentially expand its implementation.

Preliminary Analysis

Sample

The data involved a sample of 6,000 employees from ACME, with a subset of approximately 38.2% (n = 2,291) participating in the Career 2030 training program and the remaining 61.8% (n = 3,709) not participating. The participants' average age was 42.99 years (SD = 11.53), encompassing 3,338 males and 2,662 females. Within this workforce, managers made up 14.4% (n = 864), while non-managers accounted for 85.6% (n = 5,136).

Descriptive Statistics

The project examines the effect of the Career 2030 training program on employee promotions, with "training" serving as the independent variable and "promoted" as the dependent variable. Critical control variables such as "manager" to identify managerial roles, "salary" to assess compensation tiers, "age" and "sex" for demographic insights, and "edu" indicating levels of education, are included to evaluate the program's efficacy.

The summary statistics indicate an imbalance between the two groups regarding some covariates, which undermines the ignorability assumption where potential outcomes are independent of the treatment assignment (Table 1). Due to this imbalance, we cannot conclude that any differences in outcome between the treated and untreated groups can be attributed to the treatment itself, rather than differences in the covariates. Take, for example, the notable difference in merit increases during the last review cycle. Overlooking such a discrepancy precludes a definitive conclusion on whether the training effect is genuine or merely a reflection of inherent performance differences—those who received merit increases might inherently perform better and thus are more likely to be promoted. To ensure a more rigorous analysis, these imbalances must be addressed.

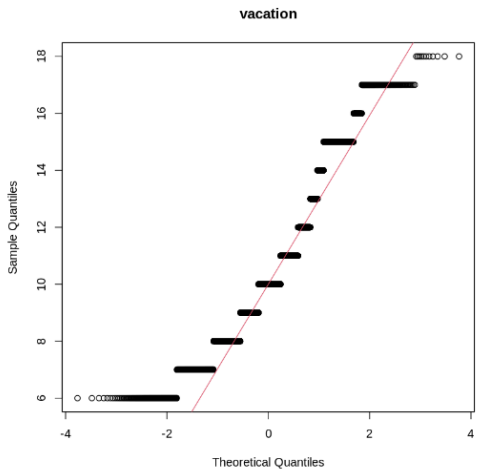
	Training Group	**Control Group**
Age	Median: 44 Mean: 42	Median: 45 Mean: 43.26
Sex	58.49% Male	53.86% Male
Vacation	Median: 10 days Mean: 10.5 days	Median: 10 days Mean: 10.31 days
Manager Proportion	16.54% Manager	13.07% Manager
Education Level	Median: 12 Mean: 11.8	Median: 12 Mean: 11.51
Raise	28.98% had merit increase in the last review cycle	41.53% had merit increase in the last review cycle
Distance From Home	Median: 16 miles Mean: 15.69 miles	Median: 26 miles Mean: 25.61 miles
Test Score	Median: 58 Mean: 56.67	Median: 67.5 Mean: 66.22
Race	Black: 15.36% White: 78.18% Other: 6.46%	Black: 16.75% White: 77.29% Other: 5.96%

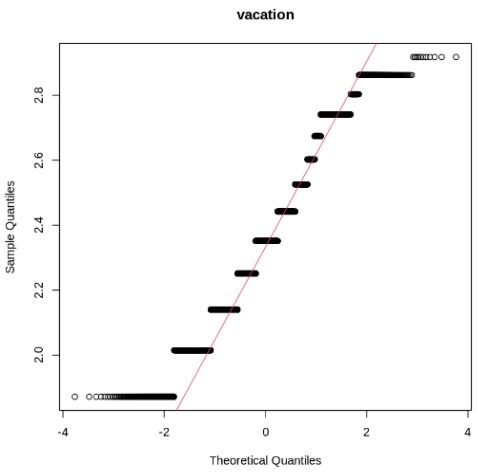
Table 1: Descriptive Statistics of Training and Control Group

Data Preprocessing

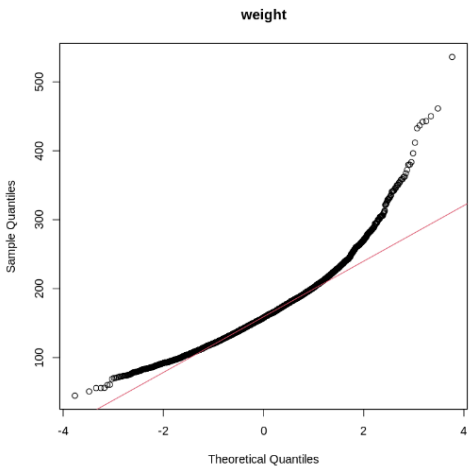
There is one negative value in the ‘edu’ column, which does not provide a meaningful interpretation in the context, as the number of educational years cannot logically be negative. In order to maintain the integrity of our dataset, we removed the corresponding rows.

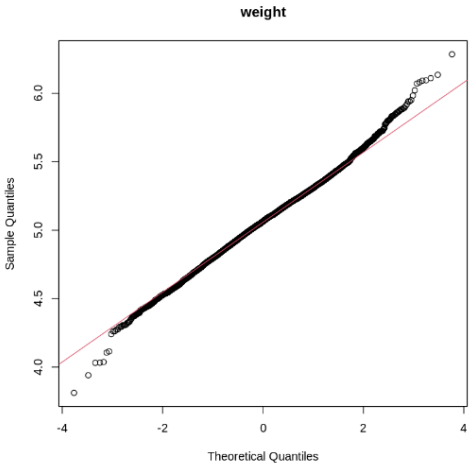
In the data preprocessing phase, we identified skewness in the ‘weight’, ‘vacation’, and ‘edu’ variables using Q-Q plot and boxplot visualizations. By applying log transformations to skewed variables, we mitigate the potential for skewed distributions to influence causal inference results, enabling a more robust and reliable analysis of causal relationships.





Graph 1 & 2: Q-Q Plots of 'vacation' before and after transformation





Graph 3 & 4: Q-Q Plots of 'weight' before and after transformation

Method

Before analyzing the impact of training sessions on promotion, we first examined the assumption of ignorability, considering both observed and unobserved confounders. Standardized mean differences (SMDs) in the non-matched dataset indicated a violation of this assumption.

To address the issue of observed confounders, we explored three distinct matching techniques to create comparable groups of employees who had, and had not, participated in the training: 1:1 Matching, Propensity Score Matching (PSM), and Inverse Probability of Treatment Weighting (IPTW). Upon comparison, PSM emerged as the superior method in terms of ensuring positivity, achieving better balance as indicated by SMDs, and maintaining an adequate sample size.

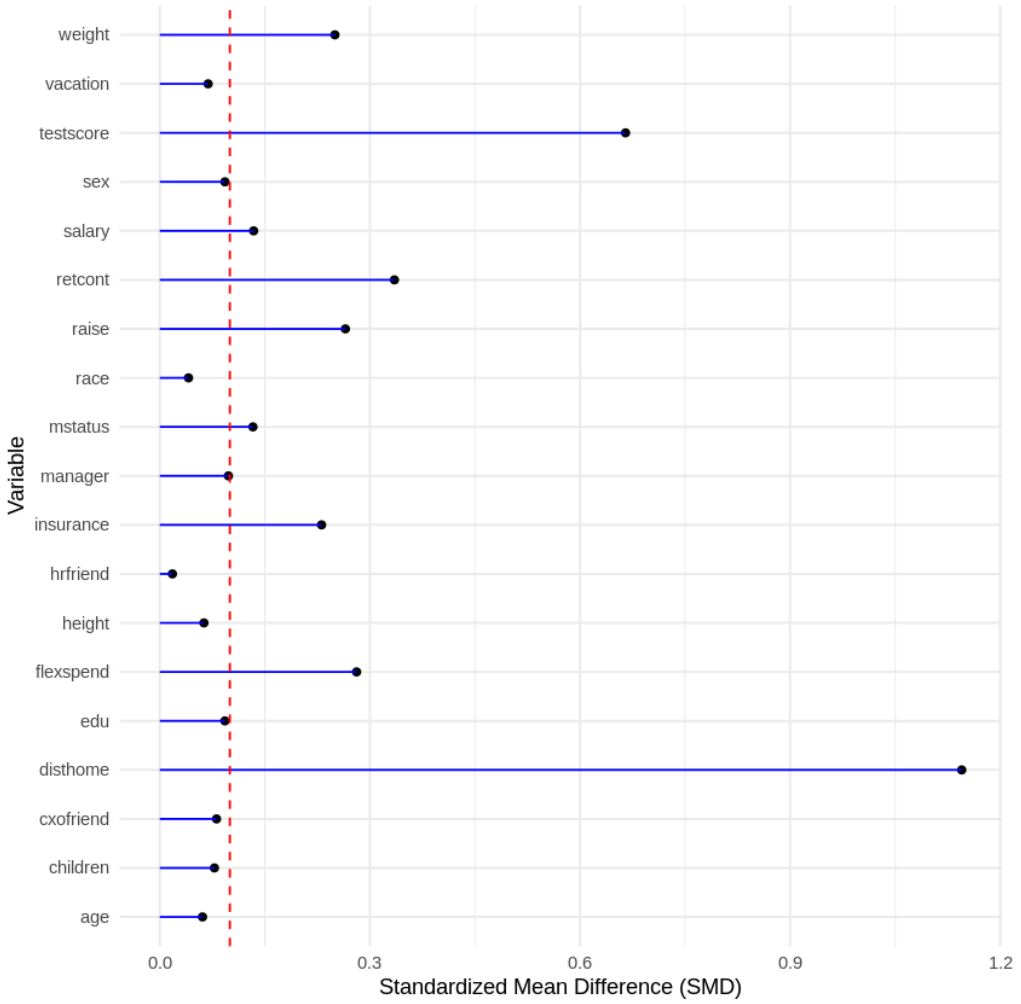
Next, to deal with unobserved confounders, we employed an instrumental variable (IV) strategy to mitigate potential endogeneity, thereby aiming to provide a more precise estimation of the training's causal effect on the likelihood of employee promotions.

Assumption Testing

In our preliminary evaluation of potential confounders, the data revealed discrepancies in covariate balances between the treated and control groups prior to any matching interventions (See *Graph 5*). Specifically, the standardized mean differences (SMDs) highlight imbalances: with 9 out of 19 covariates exhibiting SMDs above 0.1, and 7 out of these 19 showing SMDs exceeding 0.2. This imbalance indicates significant differences in key covariates between the two groups, suggesting that without corrective measures, any observed effects might not accurately reflect the true impact of training but rather the disparities in other covariates (Detailed Table in *Appendix Table 1*).

Particularly, imbalances were evident in variables such as test scores, raise and salary. These factors are pivotal to further examine since individuals with higher test scores and salaries may be more inclined to engage in training and also possess a greater probability of promotion. As such, disentangling the genuine effect of the training from these motivational elements becomes challenging.

To address this issue and enhance the comparability of the treated and control groups, we employed three distinct matching techniques: 1:1 Matching, Propensity Score Matching (PSM), and Inverse Probability of Treatment Weighting (IPTW). The purpose of these matching methods is to pair individuals from the treated and control groups who have similar covariate profiles, thereby correcting the observed imbalances.



Graph 5: Initial assessment of covariate balance before matching

1:1 Matching

We conducted various matching techniques and evaluated their standardized mean differences SMDs to identify the most effective method. The outcomes are outlined below: The 1:1 matching method is used with the optimal caliper set as 1.7 (See *Table 2*). When the caliper is set to 1.7, we have the most number of pairs and maximum number of covariables with SMDs less than 0.1 while keeping the most number of pairs. We now have 636 number of pairs, with weight slightly higher than 0.1 and only two covariates test score and distance from home having high SMDs, 0,401 and 0.583 representatively, and no assumptions violated.

Caliper	Num of Pairs	Positivity Assumption Violation	Covariates SMDs over 0.1 Count
0.8	13	No	5
1	87	No	1
1.3	274	No	2
1.5	445	No	4
1.7	**636**	**No**	**3**

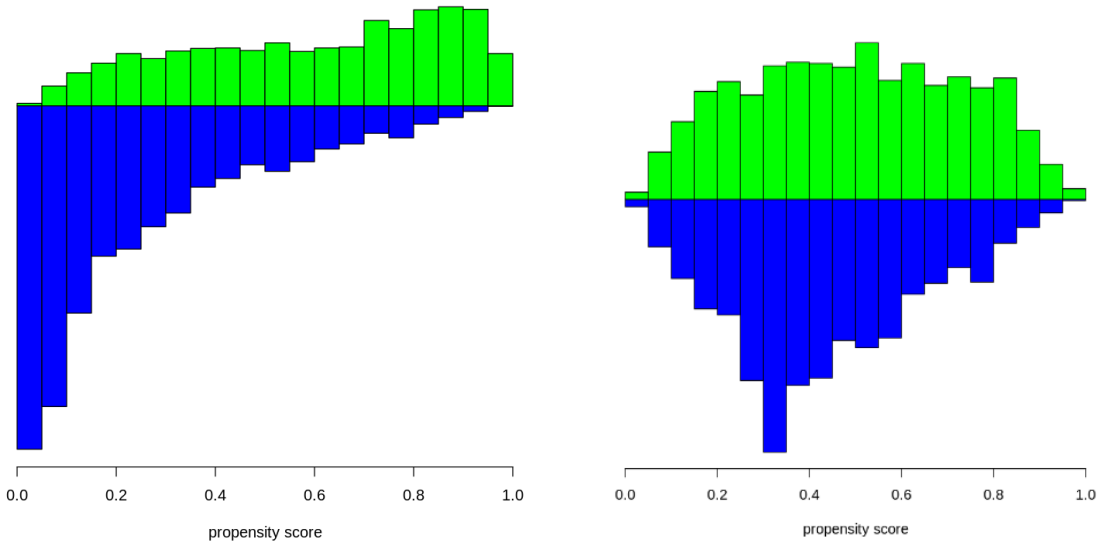
Table 2: Caliper Trials and Selection for 1:1 Matching

Propensity Score Matching

We also conducted propensity score matching aiming to find more pairs with lower SMDs. When the caliper is set to 0.80, we have the optimal result. It shows that, with no violation of positivity assumption, we have 1646 pairs and only two covariates, distance from home and test score, that have a SMD slightly over 0.1. Other caliper values may lead to the violation of positivity assumption, and also have much fewer pairs and more covariates that have more than 0.1 or even 0.2 SMDs . Given that our goal is to balance the treated and control group with no violation of assumptions, 0.8 caliper is considered the optimal solution (See *Appendix Table 3*). Utilizing the optimal caliper value of 0.80, we effectively paired the data, achieving a notable balance as visually depicted in *Graphs 6* and *7.* The table of covariate balance is in *Appendix Table 2*.

Caliper	Num of Pairs	Positivity Assumption Violation	Covariates SMDs over 0.1 Count
0.6	1566	Weak	1
0.8	**1646**	**No**	**2**
1.0	1760	Weak	2 (1 over 0.2)
1.2	1856	Weak	2 (1 over 0.2)
1.5	1991	No	4 (2 over 0.2)

Table 3: Caliper Trials and Selection for 1:1 Matching



Graph 6 & 7: Propensity score before and after matching

Inverse Probability of Treatment Weighting

Lastly, we conducted Inverse Probability of Treatment Weighting (IPTW) to reduce the confounding effect. This method successfully achieved balance between treated and control groups with most variables showing SMDs below 0.1, except for 'disthome'. Despite achieving substantial sample sizes of 5868 and 5568 for the control and treated groups respectively, we observed high maximum weights up to 30.4, deviating significantly from the median weight of 1.3 (see *Appendix Table 3*). Given the potential noise and bias associated with large weights and technical challenges in using IPTW for instrumental variable analysis, we opted for Propensity Score Matching as the preferred method for sample balancing, considering its effectiveness and practicality.

Min	1st Quartile	Median	3rd Quartile	Max
1.002	1.106	1.312	1.919	30.402

Table 4: Summary Statistics of Inverse Probability Treatment Weights

Final Matching Method Selection

Upon comparison, PSM emerged as the superior method in terms of ensuring positivity, achieving better balance as indicated by SMDs, and maintaining an adequate sample size (see *Table 5*).

	1:1 Matching	PSM Matching	IPTW
Positivity Assumption Violation	No Violation	No Violation	High Weights
Num of Covariates SMDs over 0.1	3 covariates (2 over 0.2)	2 covariates close to 0.1	
Number of employees/Pairs	636 pairs	1646 pairs	

Table 5: Propensity score before and after matching

Instrumental Variable

The experiment suggests that external factors such as manager intervention and individual motivation may influence the training attendance rate. Anecdotal evidence suggests that managers might intervene on behalf of their direct reports to ensure their inclusion in the program. Additionally, motivated individuals are more inclined to attend training sessions due to perceived career advancement opportunities.

However, as these factors were not directly observed in the experiment, it is crucial for the team to employ Instrumental Variable Analysis to address potential confounding effects. This analysis can assess the causal relationship between training program and promotion while also controlling for unobserved variables that may bias the results.

After examination, our team has identified the variable 'disthome,' representing the distance from the training facility to employees' homes, as an instrumental variable that satisfies the three assumptions:

a. ***Relevance:*** The traveling distance and time is potentially associated with employees' willingness to attend the training program.
b. ***Exclusion:*** There is no obvious relationship between the distance from employees' home to facility and their promotional opportunities.
c. ***Exogeneity:*** The distance variable is independent of the two unmeasured covariates, individual's motivation and manager intervention.

Results

Two-step Regression

In the first stage of the logistic regression, it is shown that the 'training' variable is indeed associated with the 'disthome' variable with a correlation coefficient of -0.16. This confirms our hypothesis that a longer distance from an employee's home to facility might decrease the likelihood of participation in the training program.

As seen in the table of the second stage of the regression, seven covariates show a strong association with the outcome variable ‘promoted’: training, marriage status, age, vacation days taken, friendship with HR, and friendship with the C-level executives. This suggests that older employees, those who took fewer vacation days the previous year, and those who have connections with HR or C-level executives are more strongly associated with being promoted within a year after undergoing the training.

Additionally, the treatment variable 'training' demonstrates a statistically significant causal relationship with promotion, with a p-value below 0.05. Upon exponentiating the coefficient of 'trainingest,' it is evident that attending the training program leads to a 5.15 times higher odds ratio of promotion.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.739141	0.958589	-7.030	2.06e-12	***
trainingest	**1.949231**	**0.811069**	**2.403**	**0.016248**	*****
raiseYes	-0.174053	0.105277	-1.653	0.098273	.
salary$20-$40k	-0.194120	0.204469	-0.949	0.342424	
salary$40-$80k	-0.042269	0.197503	-0.214	0.830534	
salaryUnder $20k	0.218092	0.202820	1.075	0.282241	
children	-0.039909	0.047328	-0.843	0.399099	
mstatusmarried	**-0.606534**	**0.161558**	**-3.754**	**0.000174**	*******
mstatussingle	**-1.035163**	**0.131383**	**-7.879**	**3.30e-15**	*******
age	**0.038747**	**0.004901**	**7.906**	**2.66e-15**	*******
sexMale	0.174766	0.096672	1.808	0.070635	.
edu	0.314977	0.193514	1.628	0.103596	
vacation	**-1.173692**	**0.197130**	**-5.954**	**2.62e-09**	*******
hrfriendYes	**0.393562**	**0.093784**	**4.196**	**2.71e-05**	*******
cxofriendYes	**0.800247**	**0.094488**	**8.469**	**< 2e-16**	*******
insuranceCovered & Medicaid	-0.062569	0.170431	-0.367	0.713529	
insuranceCovered & Medicare	0.125764	0.679947	0.185	0.853259	
insuranceMedicaid	-0.038297	0.138680	-0.276	0.782431	
insuranceMedicare	0.226346	0.493656	0.459	0.646586	
insuranceMedicare & Medicaid	10.015786	313.087677	0.032	0.974480	
insuranceOther	-0.104319	0.221347	-0.471	0.637433	
flexspendYes	-0.189339	0.104069	-1.819	0.068857	.
retcontYes	0.287684	0.178470	1.612	0.106974	
raceother	-0.105511	0.223024	-0.473	0.636149	
racewhite	-0.064518	0.133482	-0.483	0.628849	
testscore	**0.117709**	**0.005012**	**23.487**	**< 2e-16**	*******

Table 6: Results of Two-step Regression on Promotion

Recommendations

Through our analysis, we concluded that ACME's Career Training Program has proven effective. As such it holds the potential for broader impact if extended to more employees. To do this effectively, two main areas require attention: enhancing access to the program and refining its content to better prepare employees for career advancement.

Addressing the issue of access to training is fundamental. For instance, employees living far from the workplace are less likely to participate in the training program as highlighted in *Table 7* below. The introduction of online training solutions or providing transportation support are practical steps that can eliminate the barriers posed by physical distance.

In terms of training session content, prioritizing the development of soft skills is recommended given their established connection with promotion success, especially in the context of networking with HR and C-level executives as detailed in *Table 6.* Training focused on improving communication and networking abilities could dramatically improve the visibility and promotion prospects of employees within ACME. Moreover, considering that individuals with higher test scores may be less inclined to participate in the training (see *Table 7*), the program content should be designed to also engage and challenge these employees, thereby aligning with their advancement needs and encouraging their participation.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.108067	0.629084	1.761	0.07817	.
disthome	**-0.033402**	**0.005063**	**-6.598**	**4.17e-11**	***
raiseYes	**-0.154134**	**0.078091**	**-1.974**	**0.04841**	*
salary$20-$40k	0.002141	0.159074	0.013	0.98926	
salary$40-$80k	-0.011087	0.151715	-0.073	0.94174	
salaryUnder $20k	-0.026689	0.157079	-0.170	0.86508	
children	0.005584	0.035646	0.157	0.87553	
mstatusmarried	0.126248	0.116330	1.085	0.27781	
mstatussingle	0.021508	0.093267	0.231	0.81762	
age	0.001656	0.003716	0.446	0.65583	
sexMale	0.057788	0.072321	0.799	0.42426	
edu	0.080552	0.144205	0.559	0.57644	
vacation	0.006465	0.148391	0.044	0.96525	
hrfriendYes	0.002353	0.070944	0.033	0.97355	
cxofriendYes	0.076121	0.072471	1.050	0.29355	
insuranceCovered & Medicaid	-0.030412	0.126185	-0.241	0.80955	
insuranceCovered & Medicare	-0.463123	0.512819	-0.903	0.36648	
insuranceMedicaid	-0.096382	0.105089	-0.917	0.35907	
insuranceMedicare	-0.559180	0.363345	-1.539	0.12381	
insuranceMedicare & Medicaid	-0.013363	1.425980	-0.009	0.99252	

insuranceOther	0.007210	0.171998	0.042	0.96656	
flexspendYes	0.095499	0.077922	1.226	0.22036	
retcontYes	**-0.363604**	**0.133996**	**-2.714**	**0.00666**	**
raceother	0.125812	0.170002	0.740	0.45926	
racewhite	0.054964	0.100643	0.546	0.58498	
testscore	**-0.013625**	**0.002711**	**-5.027**	**4.99e-07**	***

Table 7: Results of Regression on Training Session

Appendix

Appendix Table 1: SMD table of unmatched dataset

Stratified by training

	0		1		SMD
n	3708		2291		
manager = Yes (%)	485	(13.1)	379	(16.5)	0.098
raise = Yes (%)	1540	(41.5)	664	(29.0)	0.265
salary (%)					0.134
> $80k	269	(7.3)	204	(8.9)	
$20-$40k	743	(20.0)	471	(20.6)	
$40-$80k	531	(14.3)	412	(18.0)	
Under $20k	2165	(58.4)	1204	(52.6)	
children (mean (SD))	1.35	(1.16)	1.44	(1.21)	0.078
mstatus (%)					0.133
divorced	910	(24.5)	467	(20.4)	
married	724	(19.5)	554	(24.2)	
single	2074	(55.9)	1270	(55.4)	
age (mean (SD))	43.26	(11.95)	42.56	(10.80)	0.062
sex = Male (%)	1997	(53.9)	1340	(58.5)	0.093
edu (mean (SD))	2.45	(0.29)	2.48	(0.29)	0.080
vacation (mean (SD))	2.35	(0.26)	2.37	(0.24)	0.091
weight (mean (SD))	5.03	(0.27)	5.10	(0.25)	0.283
height (mean (SD))	64.48	(6.65)	64.91	(6.85)	0.064
hrfriend = Yes (%)	2000	(53.9)	1215	(53.0)	0.018
cxofriend = Yes (%)	1984	(53.5)	1318	(57.5)	0.081
insurance (%)					0.231
Covered	1236	(33.3)	940	(41.0)	
Covered & Medicaid	640	(17.3)	393	(17.2)	
Covered & Medicare	27	(0.7)	10	(0.4)	
Medicaid	1516	(40.9)	789	(34.4)	
Medicare	85	(2.3)	16	(0.7)	
Medicare & Medicaid	14	(0.4)	1	(0.0)	
Other	190	(5.1)	142	(6.2)	
flexspend = Yes (%)	1057	(28.5)	957	(41.8)	0.281
retcont = Yes (%)	594	(16.0)	131	(5.7)	0.336
race (%)					0.041
black	621	(16.7)	352	(15.4)	
other	221	(6.0)	148	(6.5)	
white	2866	(77.3)	1791	(78.2)	
testscore (mean (SD))	66.22	(15.57)	56.57	(13.36)	0.665
disthome (mean (SD))	25.62	(8.67)	15.69	(8.68)	1.145

Appendix Table 2: SMD table of 1:1 matching dataset

Stratified by training

	0		1		SMD
n	636		636		
manager = Yes (%)	85	(13.4)	85	(13.4)	<0.001
raise = Yes (%)	182	(28.6)	182	(28.6)	<0.001

salary (%)					<0.001
> $80k	43	(6.8)	43	(6.8)	
$20-$40k	91	(14.3)	91	(14.3)	
$40-$80k	86	(13.5)	86	(13.5)	
Under $20k	416	(65.4)	416	(65.4)	
children (mean (SD))	1.18	(1.01)	1.21	(1.02)	0.032
mstatus (%)					<0.001
divorced	129	(20.3)	129	(20.3)	
married	87	(13.7)	87	(13.7)	
single	420	(66.0)	420	(66.0)	
age (mean (SD))	44.70	(10.10)	44.25	(10.02)	0.045
sex = Male (%)	380	(59.7)	380	(59.7)	<0.001
edu (mean (SD))	2.48	(0.21)	2.48	(0.22)	0.029
vacation (mean (SD))	2.35	(0.22)	2.36	(0.22)	0.051
weight (mean (SD))	5.06	(0.23)	5.09	(0.22)	0.132
height (mean (SD))	64.58	(6.07)	64.45	(6.38)	0.021
hrfriend = Yes (%)	330	(51.9)	330	(51.9)	<0.001
cxofriend = Yes (%)	380	(59.7)	380	(59.7)	<0.001
insurance (%)					<0.001
Covered	206	(32.4)	206	(32.4)	
Covered & Medicaid	126	(19.8)	126	(19.8)	
Medicaid	291	(45.8)	291	(45.8)	
Other	13	(2.0)	13	(2.0)	
flexspend = Yes (%)	222	(34.9)	222	(34.9)	<0.001
retcont = Yes (%)	23	(3.6)	23	(3.6)	<0.001
race (%)					<0.001
black	56	(8.8)	56	(8.8)	
other	6	(0.9)	6	(0.9)	
white	574	(90.3)	574	(90.3)	
testscore (mean (SD))	64.41	(13.63)	59.34	(11.54)	0.401
disthome (mean (SD))	22.73	(7.99)	17.97	(8.34)	0.583

Appendix Table 3: SMD table of PSM matching dataset

Stratified by training

	0		1		SMD
n	1646		1646		
manager = Yes (%)	239	(14.5)	248	(15.1)	0.015
raise = Yes (%)	562	(34.1)	525	(31.9)	0.048
salary (%)					0.031
> $80k	135	(8.2)	140	(8.5)	
$20-$40k	332	(20.2)	342	(20.8)	
$40-$80k	257	(15.6)	267	(16.2)	
Under $20k	922	(56.0)	897	(54.5)	
children (mean (SD))	1.40	(1.19)	1.42	(1.20)	0.024
mstatus (%)					0.049
divorced	372	(22.6)	355	(21.6)	
married	353	(21.4)	386	(23.5)	
single	921	(56.0)	905	(55.0)	
age (mean (SD))	42.81	(11.63)	42.63	(10.93)	0.015
sex = Male (%)	917	(55.7)	939	(57.0)	0.027

edu (mean (SD))	2.47	(0.27)	2.48	(0.28)	0.029
vacation (mean (SD))	2.36	(0.26)	2.36	(0.23)	0.001
weight (mean (SD))	5.07	(0.27)	5.08	(0.25)	0.048
height (mean (SD))	64.69	(6.74)	64.95	(6.90)	0.038
hrfriend = Yes (%)	865	(52.6)	862	(52.4)	0.004
cxofriend = Yes (%)	940	(57.1)	946	(57.5)	0.007
insurance (%)					0.058
Covered	625	(38.0)	648	(39.4)	
Covered & Medicaid	285	(17.3)	296	(18.0)	
Covered & Medicare	10	(0.6)	7	(0.4)	
Medicaid	617	(37.5)	590	(35.8)	
Medicare	21	(1.3)	15	(0.9)	
Medicare & Medicaid	1	(0.1)	1	(0.1)	
Other	87	(5.3)	89	(5.4)	
flexspend = Yes (%)	593	(36.0)	633	(38.5)	0.050
retcont = Yes (%)	150	(9.1)	117	(7.1)	0.073
race (%)					0.041
black	278	(16.9)	255	(15.5)	
other	95	(5.8)	103	(6.3)	
white	1273	(77.3)	1288	(78.3)	
testscore (mean (SD))	60.62	(15.43)	59.05	(12.46)	0.112
disthome (mean (SD))	19.93	(6.83)	18.64	(7.84)	0.175