Assessing Selection Bias of Non-experimental Estimates in Returns to Workplace Training[*]

Jan Sauermann[†] Anders Stenberg[‡]

November 11, 2019

Abstract

We assess selection bias in estimated returns to formal workplace training. Using a field experiment with random assignment to a training program, the estimated causal effect is compared to non-experimental estimates based on a sample selected by management not to participate in the experiment. Our results show that non-experimental estimates are biased, yielding returns about twice as large as the causal effect. We find that this bias is eliminated almost entirely when controlling for worker-specific characteristics, notably through the inclusion of individual fixed effects or by controlling for pre-treatment performance.

Keywords: returns to training, selection bias, field experiment

JEL Classification Numbers: J24, C93, M53.

[*] We thank Sebastian Butschek, Matthias Heinz, José Montalbán Castilla, Dirk Sliwka, Frederik Thenee, and Petra Todd. Jan Sauermann thanks the Swedish Research Council (grant number 2017-03432) and the Jan Wallanders och Tom Hedelius Stiftelse for financial support (Grant number P2014-0236:1). The field experiment used in this study is registered as AEARCTR-0004967 in the AEA RCT registry.

[†] Swedish Institute for Social Research (SOFI), Stockholm University, Institute for the Study of Labor (IZA), Research Centre for Education and the Labour Market (ROA), *and* Center for Corporate Performance (CCP). *E-mail*: jan.sauermann@sofi.su.se

[‡] Swedish Institute for Social Research (SOFI), Stockholm University, *and* IZA. *E-mail*: anders.stenberg@sofi.su.se

1 Introduction

It is widely acknowledged that an important part of human capital accumulation takes place in the labor market, for example, through informal on-the-job learning or formal workplace training programs. For the US, employees are estimated to receive an average of 47 hours of training per year, with total investment in formal workplace training estimated at 87.6 billion USD in 2018 (Freifeld, 2018). Consequently, assessing returns to workplace training is important to better understand the development of human capital, productivity and wages throughout the worker life-cycle. From a policy perspective, many governments subsidize investment in formal workplace training, e.g., by providing tax incentives to stimulate the adaptation of new skills (Cedefop, 2009). The existence of public subsidies suggests a belief that there are market imperfections (Bassanini et al., 2007), a view that may find empirical support from high estimated returns to workplace training.

Studies evaluating returns to workplace training have produced mixed results. The literature is in general difficult to assess because most studies are based on non-experimental data, leaving unknown the extent of bias remaining in these estimates (see, e.g., Pischke, 2007). Therefore, addressing selection bias is crucial. A potential way to do that is to carry out a large number of randomized control trials to provide a set of reliable estimates of the returns to workplace training. However, this would be both expensive and time-consuming. An alternative is to examine to what degree standard econometric methods are able to reduce selection bias by comparing non-experimental estimates against the benchmark of existing experimental estimates. Such estimates have been reported for evaluations of e.g. active labor market programs, antipoverty programs, and early education programs (see, e.g., LaLonde, 1986; Heckman et al., 1998; Diaz and Handa, 2006; Griffen and Todd, 2017). If non-experimental estimates of the returns to workplace training are only modestly biased, the findings could provide guidance on how to assess non-experimental estimators and ultimately to improve our understanding of the impact of workplace training.

The aim of this study is to explore to what degree endogenous training participation causes selection bias in estimated returns to workplace training. We exploit a field experiment with random participation of workers to a one-week training course, conducted in a call center of a multinational telephone company in the Netherlands (De Grip and Sauermann, 2012). This experiment provides an unbiased estimate of the returns to workplace training. We compare this estimate to non-experimental estimates based on a sample of agents that were not selected to be part of the experiment. These individuals, who worked in the same firm during the same time period, therefore constitute an endogenously selected sample. This allows us to use the experimental estimate as a benchmark, against which the potentially biased non-experimental estimates can be compared. This provides novel results by assessing the bias remaining in estimators based on standard regression techniques. We first replicate the main finding of De Grip and Sauermann (2012), which shows that the workplace training program led to an increase in productivity, as measured by average handling time, of about 11%. Using the non-experimental control group, the same specification leads to an estimated return to training of 21%, i.e., almost twice the experimental estimate. By exploiting the panel structure of the data, we show that when including pre-treatment performance as a proxy of worker ability, or when including worker fixed effects to capture unobserved individual-specific factors, the non-experimental estimates are very similar to those reported from the experimental sample. This suggests that in this setting, non-experimental approaches can yield estimates with relatively modest bias compared with experimental evaluations.

This study relates both to the long-standing literature quantifying the returns to workplace training on workers' wages or worker productivity, as well as to earlier assessments of selection bias in non-experimental settings. Returns to workplace training on wages are typically estimated using survey data, while the few studies estimating returns on worker productivity use personnel data from individual firms that contain direct measures of worker productivity, such as the number of pieces produced by garment workers (Adhvaryu et al., 2018), amount of sales for sales clerks (Prada et al., 2019), and average handling time for call agents (Liu and Batt, 2007; De Grip and Sauermann, 2012).[1] Table A.1 in the Online Appendix provides an overview of estimated returns to both wages and

[1] While the potential impact on wages is interesting in itself, reflecting workers' payoff to training, productivity is arguably more interesting from a societal point of view. For instance, social cost-benefit calculations of public policies typically seek to determine whether productivity increases are sufficient to cover the associated social costs.

worker productivity, displaying a high degree of heterogeneity among studies.[2] This is explained partly by differences in study setup. For example, in studies based on (representative) survey data, differences in estimates may arise because the individuals compared are in different occupations and in different labor markets, and because training is often broadly defined to include programs of different lengths and different contents (Bartel, 1995; Pischke, 2007). Further, rent-sharing between workers and firms implies that wage returns should be generally smaller than the returns on underlying worker productivity. Irrespective of the nature of the data, however, estimates are likely upward biased if participants are positively selected into training programs. Based on the results in Heckman et al. (1998), this bias may be more modest in studies using personnel data that analyze participants in the same labor market, employed by the same firm, and subject to the same training program.

To the best of our knowledge, only three studies use randomized assignment to training to evaluate effects on worker-level outcomes. These studies use personnel data from individual firms and report fairly large effects from workplace training on measures of worker productivity of between 10% and 20% (De Grip and Sauermann, 2012; Adhvaryu et al., 2018; Prada et al., 2019).[3]

Conventional non-experimental ordinary least squares (OLS) estimates on the returns to wages, shown in Table A.1 in the Online Appendix, range from -3% to 21.8% (average of 5.7%).[4] Overall, the small or even zero wage effects reported (e.g., Krueger and Rouse, 1998; Goux and Maurin, 2000; Leuven and Oosterbeek, 2004, 2008) contrast with those reported on worker productivity from randomized assignment to training but can be reconciled if firms pass on only a small share of profits to employee wages, as shown by Adhvaryu et al. (2018). A few studies seek to elicit exogenous variation when estimating returns (on wages), e.g., by exploiting age-cutoffs in tax deductions of training investments

[2] In addition to studies analyzing the returns to worker outcomes, other studies estimate the effect of firm training investments on the firm or establishment-level outcomes. See, e.g., Bartel (2000), Dearden et al. (2006), and Konings and Vanormelingen (2015) and the literature cited therein.

[3] Furthermore, Dimitriadis and Koning (2019) find that randomly assigned participation in a two-hour communications training session significantly increases self-reported profits for entrepreneurs. Alfonsi et al. (2019) report that randomly offering firms wage subsidies to train workers on-the-job shows no significant earnings returns.

[4] For a systematic meta-study on the returns to workplace training, see Haelermans and Borghans (2012).

(Leuven and Oosterbeek, 2004), by exploiting unanticipated withdrawals from training (Leuven and Oosterbeek, 2008), or by using worker rank measures to instrument for training investments (Bartel, 1995). However, the majority of studies take selection into account by controlling for observable characteristics or worker fixed effects. We aim to provide guidance on the performance of different estimators when assessing returns to workplace training using non-experimental data.

This study also relates to earlier assessments of selection bias in non-experimental estimates that focus on active labor market programs, antipoverty programs, and early education programs (e.g., LaLonde, 1986; Heckman et al., 1998; Diaz and Handa, 2006; Griffen and Todd, 2017). Exploiting randomized participation in an employment program in the US, LaLonde (1986) defines non-experimental control groups using data from the Panel Study of Income Dynamics (PSID) and from the Current Population Survey (CPS). Relative to the experimental benchmark, difference-in-differences and control function approaches result in substantial deviations from experimental estimates, suggesting that it is difficult to uncover a causal effect from non-experimental data.

Analyzing an active labor market training program, Heckman et al. (1998) decompose selection bias into distinct parts, and emphasize the importance of common support, i.e., that there are no differences in conditional treatment probability between treated and untreated individuals. If there are probabilities for which only treated individuals exist, even methods that could solve the selection problem can compare only non-comparable individuals. Heckman et al. (1998) also highlight the importance of data quality, notably that treated and non-treated individuals are subject to the same regional labor markets, collected from the same data sources, and that treated individuals are subject to the same training programs. Following these adjustments, the findings indicate that the remaining bias is only a small fraction (7%) of the conventional measure of selection bias and not statistically significantly different from zero (Heckman et al., 1998, Table V). These results suggest that the poor performance of non-experimental estimators in LaLonde (1986) are, to a large extent, driven by these issues of data quality.

Other assessments of non-experimental estimates have since involved large-scale policy programs such as PROGRESA, an antipoverty program in rural areas of Mexico (Diaz and Handa, 2006; Bifulco, 2012), and Head Start, an early childhood intervention program in the US (Griffen and Todd, 2017). These studies conclude that when limiting analysis to the region of common support and after careful consideration of data quality, remaining bias is modest compared with experimental estimates. Overall, these results suggest that mean differences in outcomes between treated and untreated groups can be explained largely by selection on observed characteristics. Studies concerning other treatments, however, find that non-experimental estimates do not recover the experimental estimate. This applies to programs to prevent dropping out of high-school (Agodini and Dynarski, 2004) and measuring income gains from migration (McKenzie et al., 2010).[5]

The contribution of the present study is to assess to what extent conventional OLS estimates of returns to workplace training are able to correct for selection bias using non-experimental data. To the best of our knowledge, this study is the first to present such an evaluation for workplace training. We find that controlling for measures of pre-treatment performance and controlling for observed and unobserved ability through worker fixed effects removes most of the bias in OLS estimates using non-experimental data. Our estimates are obtained using personnel data where agents are employed in the same firm, participate in the same training program and are subject to the same labor market. Given that evaluating workplace training with random assignment is costly and rarely done, our findings suggest that these approaches to correct for selection of OLS estimates may offer an interesting alternative to IV estimates, since the latter is often much less precisely estimated.[6]

While our data is specific to a call center, we highlight two aspects regarding the generalizability of this study. First, as opposed to other industries, there is usually no vocational education for call agents, despite extensive use of information technology in the call center sector (Sieben et al., 2009). This implies that call centers themselves need to undertake the bulk of human capital investment themselves. Second, the size of the workforce in typical call center occupations in the US was estimated to be 3.8 million in

[5] For advertisements on Facebook, Gordon et al. (2019) find that observational estimates often fail to recover the experimental estimates.

[6] Given that some evidence indicates that carefully crafted OLS estimates are relatively close to experimental estimates, one may argue that there is a trade-off between modestly biased OLS estimates and less precise but unbiased IV estimates. As pointed out by Black et al. (2017, page 6), this may be viewed as similar to how researchers trade off bias and variance via choice of a bandwidth or other tuning parameter.

2018, or 2.7% of the total workforce (Batt et al., 2005; Bureau of Labor Statistics, 2018). Although modest in size, call center employees represent a non-negligible proportion of lower skilled employees in service sector occupations that gained in hours worked and real hourly wages in the context of polarization of the labor market (Autor and Dorn, 2013).

2 Setting

2.1 Institutional setting

Our field experiment was conducted in the call center of a multinational telecommunications company located in the Netherlands.[7] The call center consists of several departments, the largest of which is for customers with fixed cellphone contracts. Customers phone the call center, e.g., with technical problems, billing problems, or complaints, and are routed to available agents to take their calls. The agents' sole task is to answer these customer calls, and to enter notes into the customer database about the call during or after the call. Agents are not involved in other tasks, such as back-office work or sales.

Our estimation sample, which lasts from week 45/2008 to week 24/2009, includes a total of 157 agents organized in 13 teams.[8] Each team is led by a team leader responsible for one team and its agents only. The primary purpose of team leaders is to monitor and evaluate agents efficiently. There are no team-related incentives or specialized tasks.

Despite recording several key performance indicators of agent performance, agents are formally evaluated only once a year. Appraisal interviews between team leaders and their subordinates result in a grade that determines both an annual bonus and a wage increase.[9] There is no piece rate pay or other performance incentives for agents in this call center. Before entering the call center, agents participate in an intensive four-week training course. Throughout their career in the call center, agents receive additional,

[7] For a more in-depth description of the firm and the field experiment, the reader is referred to De Grip and Sauermann (2012).

[8] In week 50/2008, management decided which agents to include in the field experiment. Our analysis is restricted to agents employed in the department at the time of the announcement of the training program. Also including agents who started after the announcement yields very similar results.

[9] The annual wage increase typically ranges between 0 and 8%.

shorter training courses, e.g., to acquire or improve technical skills or communication skills.

2.2 Field experiment and sample definition

Firm management introduced a new training program that aimed to improve the department's main key performance indicator, which is average handling time of customer calls. The program was designed as a week-long group training program running from Monday to Friday and was held in the in-house training center, physically separated from the work spaces. Participating agents were paid in full for the training week irrespective of their contractual hours. The training consisted of two parts: roughly half the sessions consisted of discussing which skills agents were lacking to efficiently do their job, and, for example, how agents could help each other on the work floor. In the remaining sessions, agents worked on selected customer calls with direct support from the team coach. Due to capacity constraints, a maximum of 10 agents could be trained at once. All training sessions were held and led by a team coach.

Timing of experiment Of all agents selected to participate in the field experiment, the randomly selected treatment group was trained in weeks 10/2009 to 14/2009. Agents from the randomly selected control group were trained after week 24/2009. While the weeks prior to the first training of the treatment group (in week 10/2009) serve as a pre-treatment period, the weeks between the last training of the treatment group and the first training of the control group serve as a post-treatment period, during which only agents of the treatment group had been trained (weeks 15/2009-24/2009).

In January 2009, the training program was announced in a general message from management. The actual weeks in which agents were trained was communicated about four weeks prior to the training week, when agents were typically informed about their schedule.

Assignment to treatment and control groups, and the non-experimental control group. Out of the 157 agents working in the call center during the sample period, management selected a total of 74 agents to be part of the field experiment.[10] The main criterion for including agents in the field experiment was tenure: management selected agents with longer tenure to avoid losing training investment due to high turnover rates, which are common to call centers. While management did not apply a strict tenure threshold to be assigned to the field experiment, the data show that 71% of those *not* selected for the field experiment have a tenure of one year or less, whereas the corresponding figure for agents selected for the field experiment is only 19%.

Agents assigned to participate in the field experiment were randomly assigned to be treated during the treatment period ($N = 34$), or to be treated after the end of the experiment ($N = 40$). Due to the restriction that agents should be trained with other agents from the same team, half the teams were randomly assigned to the treatment group, whereas the other half were assigned to the control group. Each team was then randomly split into different training groups, due to size constraints of the training center. Descriptive statistics for treatment and control groups, as well as t-statistics for differences, are shown in Table 1. Column (5) replicates the finding of De Grip and Sauermann (2012) showing that agents in the experimental treatment and control groups do not differ significantly in terms of the observable characteristics. The F-test on joint significance is 0.74 with a p-value of 0.64.[11]

We additionally make use of the 73 agents who were *not* selected to be part of the field experiment and who constitute the *non-experimental control group*.[12] Column (6) of Table 1 shows that these agents had, relative to agents assigned to the experimental treatment group, on average 2.9 fewer years of tenure and an average performance that was 0.8 of a standard deviation lower. The relatively low performance of agents not selected into

[10] The field experiment also included 10 agents re-assigned from the treatment to the control group, and vice versa, for example, in case of illness or scheduled vacations (see De Grip and Sauermann, 2012). These agents are similar on observables, and are excluded from the main results in this study. Including these 10 in analyses presented below only has a marginal impact on the estimates.

[11] Table A.2 in the Online Appendix provides detailed information on how groups are defined. Note that the total number of agents in this study ($N = 157$) differs from the number reported in De Grip and Sauermann (2012, $N = 179$). A re-evaluation of the data used in De Grip and Sauermann (2012) shows that Column (1) and (3) of their Table 1 also includes individuals who did not work during the observation period, and thus cannot be used to identify the treatment effect of training participation. These observations therefore do not affect their estimates. In the present paper, these individuals are excluded from the sample.

[12] Agents in the non-experimental control group have not been previously analyzed.

the field experiment likely reflects both their lower tenure, and factors unobservable to the researcher. Agents in the non-experimental sample are also more likely to leave the department. If sample attrition is correlated with either unobserved ability or treatment status, it could bias our estimation results. We further explore this in Section 4.1.

Table 1: Descriptive statistics by group and *t*-tests of differences between samples

	(1)	(2)	(3)	(4)	(5)	(6)
		Exp. sample	Exp. sample	Non-exp.		
	All agents	Treat. group	Control group	sample	*t*-test (2)-(3)	*t*-test (2)-(4)
Gender (1=male)	0.293	0.382	0.275	0.301	0.107	0.081
	(0.457)	(0.493)	(0.452)	(0.462)	(0.111)	(0.100)
Age (in years)	33.009	34.070	36.146	29.866	-2.076	4.204*
	(11.304)	(10.095)	(11.640)	(10.847)	(2.527)	(2.186)
Tenure (in months)	36.975	53.353	49.450	16.110	3.903	37.243***
	(47.377)	(49.578)	(51.857)	(30.053)	(11.812)	(9.201)
Working hours	20.154	19.053	20.182	20.864	-1.129	-1.812
	(6.410)	(6.348)	(5.682)	(7.041)	(1.411)	(1.365)
Share peak hours	0.547	0.554	0.523	0.565	0.031	-0.011
	(0.098)	(0.106)	(0.129)	(0.063)	(0.027)	(0.020)
Pre-treatment performance	0.336	0.364	0.374	0.285	-0.010	0.079***
	(0.094)	(0.072)	(0.071)	(0.087)	(0.017)	(0.016)
Turnover	0.338	0.206	0.250	0.493	-0.044	-0.287***
	(0.474)	(0.410)	(0.439)	(0.503)	(0.099)	(0.092)
Number of agents	157	34	40	73	74	107

Notes: *** p<0.01, ** p<0.05, * p<0.1. Columns (1) to (4) show means and standard deviations in parentheses; Columns (5) and (6) show differences between experimental treatment group and control group (Column 5), and experimental treatment group and non-experimental control group (Column 6), respectively, and standard errors are in parentheses. Performance is defined as the inverse of *average handling time*, share of peak hours is defined as number of hours worked during high customer demand hours of the day, pre-treatment performance is defined as average performance before management's assignment to the field experiment, and turnover is defined as whether an agent left the department before the end of the observation period. A more detailed description of all variables is given in Table A.3 in the Online Appendix.

Out of the 73 agents in the non-experimental control group, a total of 7 agents were later selected to be trained along with agents in the treatment group. This group, denoted G1, was placed by management in the training program to fill vacant slots during the training period in weeks 10/2009 to 14/2009. In the regressions estimating the treatment effect, G1 agents will therefore also contribute to identifying the treatment effect based on non-experimental data. The remaining agents in the non-experimental control group ($N = 66$), denoted G2, were not assigned to the field experiment, and were also not added to the training program throughout the observation period. Both groups, G1 and G2, therefore did not have a randomized treatment status but were *endogenously* assigned by management.[13]

2.3 Measuring performance

To measure performance of individual call agents, it is important to have a measure that is comparable *between* agents, but also *within* agents over time. For purposes of this study, we use the main key performance indicator used by call center management, average handling time, variants of which have been used in several studies (e.g., Liu and Batt, 2007; De Grip and Sauermann, 2012; Breuer et al., 2013). In this call center, customer calls are randomly assigned to individual agents. Agents are not able to pick out particular calls, e.g., to get calls with shorter expected length. For this reason, all agents have a priori the same probability of receiving calls from customers with very short or very long length.

Average handling time is defined as the average length of all calls an agent handled in a week. Short calls are interpreted as good calls, not least because they are less expensive for the firm. For this reason, we use the inverse of average handling time, multiplied by 100 so that high levels of our measure can be interpreted as high performance. Our measure of performance is available for any week an agent is working.[14]

Average handling time is driven both by individual-specific characteristics as well as by period-specific effects. The latter can occur, for example, if the department has problems with the IT-infrastructure of the firm, systematic errors in invoices, or deviations in the number of predicted incoming customer calls. In our observation period, individual (worker) fixed effects alone explain 58% of total variation in handling time, whereas additionally controlling for week fixed effects adds only 7 percentage points in the variation explained.

[13] Descriptive statistics for the two groups, G1 and G2, and corresponding *t*-tests are shown in Table A.4 in the Online Appendix.

[14] Quality of agent calls is assessed in two ways. First, team leaders regularly listen in to calls. Second, key performance indicators, such as the share of customers calling back after talking to a call agent, are used to assess quality continuously. Previous studies using data from this firm show limited trade-off between average handling time and quality of calls (De Grip and Sauermann, 2012; De Grip et al., 2016).

3 Estimation framework

To estimate the returns to training for the experimental treatment group against the experimental and non-experimental control groups, respectively, we regress the logarithm of worker productivity on a dummy, being 1 if an individual is treated, and zero otherwise. The estimation equation can then be written as

$$log(y_{it}) = \alpha + \tau d_{it} + \beta_1 X_{it} + \beta_2 t_t + u_{jt} \quad (1)$$

where y_{it} denotes our measure of productivity of worker i in week t, which is based on average handling time and for which high levels of y_{it} are interpreted as high performance. The treatment dummy d_{it} is defined as being 1 in each week after an agent has participated in the training, and 0 otherwise. Equation (1) also contains time varying controls X_{it} and a common time trend t. The idiosyncratic error term u_{it} is clustered at the team level to account for team level randomization (Abadie et al., 2017).[15]

In the absence of exogenous variation in the treatment indicator d_{it}, we use different methods to account for selection into treatment. To get a first assessment of bias, we estimate Equation (1) with no controls, comparing the results obtained with the experimental sample and the non-experimental sample. We then apply specifications based on explanatory variables used in De Grip and Sauermann (2012), which includes an agent's working hours, share of peak hours, total number of incoming calls in week t, and time trend t_t.[16] To mimic management's decisions on selection into the training program, we then include two variables that were explicitly mentioned by management as reasons for selecting agents: pre-treatment performance and agent tenure.[17]

We also explore to what degree worker fixed effects contribute to reducing bias. The idea here is that unobserved individual-specific components μ_i are correlated with the decision to participate in the training program $Cov(\mu_i, d_{it}) \neq 0$. Augmenting Equation

[15] Results do not change qualitatively when clustering at the individual (agent) level.

[16] See Table A.3 in the Online Appendix for detailed variable definitions.

[17] Pre-training wages were shown by Blau and Robins (1987) to strongly reduce conventional OLS estimates (see Table A.1 in the Online Appendix).

(1) with an individual-specific term μ_i and demeaning then eliminates any time-constant, individual-specific characteristics:

$$\begin{aligned} log(y_{it}) - \overline{log(y_i)} &= \tau(d_{it} - \overline{d}_i) + \gamma(\mu_i - \overline{\mu}_i) + (\epsilon_i - \overline{\epsilon}_i) \\ &= \tau(d_{it} - \overline{d}_i) + \epsilon'_i \end{aligned} \tag{2}$$

Even if our models capture important differences between treated and untreated, Heckman et al. (1998) characterize three sources of bias that may remain. First, selection bias may originate from differences in common support, i.e., potential differences in background variables such that there are only (non-)treated individuals for certain values of X_{it}. For instance, if everyone with a certain tenure is (un)treated, the lack of common support between treated and untreated individuals prevents a comparison of comparable individuals. To test whether differences in the common support of observable characteristics influence the estimated treatment effect of experimental and non-experimental samples, we replicate our main analysis with a sample restricted to all agents within common support, i.e., to the propensity score distribution between the 5th and 95th percentiles, and the 20th to 80th percentiles, respectively. A second source of bias stems from differences in the *distributions* of the observable characteristics within the area of common support. For example, treated agents may be over-represented among those with long tenure, but under-represented for short periods of tenure. Our simple common support restriction does not address this bias. Third, there may be systematic differences in unobservable characteristics between treated and untreated individuals. This is perhaps the most often discussed problem and arises if variables unobserved to the researcher, e.g., motivation or ability, influence the estimates. To the extent that unobservable characteristics are individual-specific and time-constant, the specification including worker fixed effects accounts for this source of bias.

4 Results

4.1 Baseline results

Table 2 shows our main results. Each estimate in the table is a treatment effect stemming from a separate regression in which the experimental treatment group is either compared to agents in the experimental control group (Panel A) or to agents in the non-experimental control group (Panel B). Panel C shows measures of selection bias in the returns to training. It reports both the absolute difference between the estimates shown in Panel B and Panel A, and the relative difference between the two estimates. Each column shows results from a different specification.

Column (1) shows that, for the 74 agents who were part of the field experiment, participants in the training program display increased post-treatment performance of 10.9%. This result replicates De Grip and Sauermann (2012) and may be given a causal interpretation since treatment was randomly assigned to agents. The estimate in Panel B shows the corresponding estimate for the non-experimental sample. This regression is based on the experimental treatment group and agents who were selected by management *not* to participate in the field experiment (see Section 2.2). Using this non-experimental control group, the estimated treatment effect is 21.8%. The biased non-experimental estimate in this Column is 99% larger than the unbiased, causal estimate. This supports the view that selection bias in returns to workplace training can be substantial.[18]

The key question here is whether it is possible to explain this strong difference in treatment effect estimates between experimental and non-experimental samples shown in Column (1) of Table 2. To explore this, Columns (2) to (6) of Table 2 provide analogous estimation results using different sets of control variables.[19] Column (2) adds an agent's weekly working hours, her share of hours worked during hours of the day with high customer load (share peak hours), and number of customer calls divided by number of full-time equivalent agents in a week (calls per FTE). These are the same control variables

[18] Table A.5 in the Online Appendix shows the equivalent regressions for subgroups G1 and G2 of the non-experimental control group. Agents in group G1, i.e., agents who were initially not selected for the field experiment but later placed in the training, are more similar to agents in the field experiment than to those in group G2 (see Table A.4).

[19] Corresponding estimation results for subgroups G1 and G2 show that selection bias and its reduction are driven mainly by agents in G2 (cf. Table A.5 in the Online Appendix).

Table 2: Treatment effects of workplace training

Dependent variable: logarithm of worker productivity						
	(1)	(2)	(3)	(4)	(5)	(6)
A: Experimental sample						
Treatment dummy ($\widehat{\tau}_E$)	0.1092***	0.1127***	0.1195***	0.0835***	0.1252***	0.1167***
	(0.0179)	(0.0182)	(0.0201)	(0.0217)	(0.0100)	(0.0105)
Adjusted R-squared	0.0315	0.0668	0.0809	0.0730	0.0662	0.4124
Number of agents	74	74	74	74	74	73
Number of observations	1,859	1,859	1,859	1,859	1,859	1,850
B: Non-experimental sample						
Treatment dummy ($\widehat{\tau}_N$)	0.2175***	0.2160***	0.1868***	0.1242***	0.1346***	0.1321***
	(0.0369)	(0.0387)	(0.0233)	(0.0200)	(0.0116)	(0.0245)
Adjusted R-squared	0.0929	0.1089	0.1365	0.1501	0.0628	0.4692
Number of agents	107	107	107	107	107	104
Number of observations	2,383	2,383	2,383	2,383	2,383	2,336
C:Selection bias						
$\widehat{\tau}_N - \widehat{\tau}_E$	0.1083	0.1033	0.0673	0.0407	0.0094	0.0154
p-value ($\widehat{\tau}_N - \widehat{\tau}_E$)	0.0078	0.0108	0.0060	0.1620	0.3053	0.5049
Bias in % (($\widehat{\tau}_N - \widehat{\tau}_E$)/$\widehat{\tau}_E$*100)	99.1	91.6	56.3	48.8	7.5	13.2
Control variables	No	Yes	Yes	Yes	No	No
Tenure (linear + squared)	No	No	Yes	No	No	No
Common trend	No	No	No	Yes	No	No
Worker FE	No	No	No	No	Yes	No
Pre-treatment performance	No	No	No	No	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: $log(y_{it})$. Standard errors are clustered at the team level. Control variables include working hours, share peak hours, and calls per full-time equivalents (FTE). All regressions include a constant. Pre-treatment performance is defined as average performance over weeks 45/2008 to week 49/2008, i.e., before management decided on assignment to the training program.

used in De Grip and Sauermann (2012). The non-experimental estimate remains almost double the size of the experimental estimate.

In Column (3), we control for agent tenure by including a linear term capturing tenure measured in months, and its squared term. Agent tenure may be important to include for two reasons. First, tenure was mentioned as the main argument for including agents in the field experiment. Second, the outcome variable used, average handling time, exhibits a strong tenure pattern, with a non-linear increase in performance over the first year of tenure (De Grip et al., 2016). In the specification controlling for tenure, the experimental estimate is 11.95%, whereas the corresponding non-experimental estimate is 18.7%, which corresponds to reducing selection bias by almost half from 91.6% to 56.3%.

In Column (4), the specification includes control variables and a linear time trend. The experimental estimate is then reduced to 8.4%, with the non-experimental estimate still almost 50% larger (12.4%). The difference between the estimates, however, is not statistically significant.

When including individual fixed effects (Column 5) or the measure of pre-treatment performance (Column 6), the non-experimental estimates are only slightly larger than estimates based on the experimental sample, and the bias corresponds to 7.5% and 13.2% respectively. Taken together, the results found in Table 2 show that conditioning on individual fixed effects or pre-treatment performance reduces selection bias to the point where it is relatively small and not significantly different from zero.

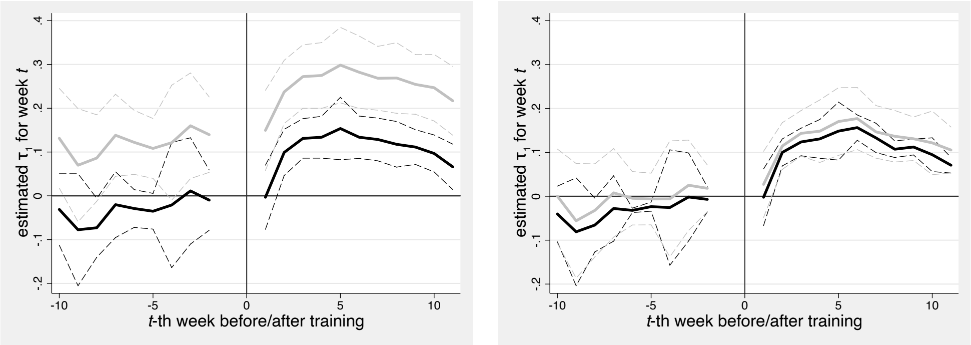
To check if attrition influences our findings, we provide estimates for the sample of agents who do not exit the department. For the baseline specification without controls, the estimated bias is 72% (Table A.6 in the Online Appendix). For the specifications including worker fixed effects, and including pre-treatment performance, the estimated bias is almost the same as without this sample restriction (7% and 12%, respectively). This suggests that the main results are not driven primarily by differential attrition between the experimental and non-experimental samples.

4.2 Dynamics

The weekly frequency of the outcome variable y_{it} allows us to compare treated and untreated agents by week before and after training participation. Panels (a) and (b) of Figure 1 show the treatment dummy estimates for different specifications and samples by weeks relative to the training. Solid black lines show the treatment effect for the experimental sample when not including any control variables (Panel (a)), and when including worker fixed effects (Panel (b)). For both specifications, the results show that the training program already caused performance to increase substantially in the second week following training, but it decreased a few weeks later. As one would expect in a randomized experiment, the estimates prior to training participation are close to zero and insignificant.

Gray lines show corresponding estimates for the non-experimental control group. The estimates in Panel (a), i.e., those taken from a regression with no controls, show a similar dynamic pattern, but are clearly larger than the experimental estimates both before and after treatment. Estimates for the non-experimental group *including* worker fixed effects (Panel (b)), however, are remarkably close to the experimental estimate, which are shown as a solid black line. Thus, despite using data from the non-experimental control group, including worker fixed effects yields trajectories that closely follow those of the experimental estimates.

Figure 1: Dynamic treatment effects for experimental, non-experimental samples



(a) No controls

(b) Worker fixed effects

Notes: This figure shows the treatment dummy estimates by weeks before and after training participation for the experimental sample (black lines) and the non-experimental sample (gray lines). The figure in Panel (a) shows estimates for the specification with no controls; the figure in Panel (b) shows the corresponding results when including worker fixed effects. Solid lines show point estimates from regressions with treatment dummies only; dashed lines show the corresponding 95% confidence intervals. Period $t-1$ serves as the reference period. There is no performance information in the training week itself.

4.3 Imposing common support restrictions

The descriptive statistics in Table 1 show that agents in the non-experimental control group have, on average, much shorter tenure and, as a result, substantially lower pre-treatment performance. These differences in observable characteristics may yield bias due to a lack of common support (see Section 3). Below, we combine regressions with common support restrictions to make the sample of treated and non-treated individuals more comparable (Heckman et al., 1998; Imbens and Wooldridge, 2009).

The common support restriction is applied by estimating a propensity score, defined as the probability of assignment to either the experimental group or the non-experimental group. This probability is predicted via a probit model including all pre-treatment variables, i.e., an agent's gender and age, tenure and tenure squared, working hours, share of peak hours as well as pre-treatment performance (see Table A.7 in the Online Appendix). The distributions of the estimated propensity scores for the experimental sample and the non-experimental control group are shown in Figure A.1. The figure shows that both groups are represented in the propensity score span of 0.1 to 0.9. The non-experimental control group shows no observations in the rightmost parts of the distribution. Conversely, on the leftmost end of the distribution, there is a large proportion from the non-experimental group, but no individuals from the experimental group.

Table 3: Treatment effects under common support

Dependent variable: logarithm of worker performance						
	(1)	(2)	(3)	(4)	(5)	(6)
Common support restriction	5th-95th	5th-95th	5th-95th	20th-80th	20th-80th	20th-80th
A: Experimental sample						
Treatment dummy ($\widehat{\tau}_E$)	0.1079***	0.1295***	0.1158***	0.1163***	0.1262***	0.1146***
	(0.0217)	(0.0070)	(0.0097)	(0.0330)	(0.0051)	(0.0155)
Adjusted R-squared	0.0315	0.0788	0.4405	0.0437	0.0804	0.3294
Number of agents	67	67	67	47	47	47
Number of observations	1,684	1,684	1,684	1,173	1,173	1,173
B: Non-experimental sample						
Treatment dummy ($\widehat{\tau}_N$)	0.1924***	0.1348***	0.1272***	0.1367***	0.1245***	0.1105***
	(0.0299)	(0.0071)	(0.0161)	(0.0342)	(0.0048)	(0.0165)
Adjusted R-squared	0.0888	0.0682	0.4159	0.0591	0.0676	0.3408
Number of agents	88	88	88	57	57	57
Number of observations	2,040	2,040	2,040	1,397	1,397	1,397
C:Selection bias						
$\widehat{\tau}_N - \widehat{\tau}_E$	0.0845	0.0053	0.0114	0.0204	-0.0016	-0.0041
p-value ($\widehat{\tau}_N - \widehat{\tau}_E$)	0.0212	0.4328	0.5445	0.3056	0.3160	0.7727
Bias in % (($\widehat{\tau}_N - \widehat{\tau}_E$)/$\widehat{\tau}_E$*100)	78.3	4.1	9.8	17.6	-1.3	-3.5
Control variables	No	No	No	No	No	No
Tenure (linear + squared)	No	No	No	No	No	No
Common trend	No	No	No	No	No	No
Worker FE	No	Yes	No	No	Yes	No
Pre-treatment performance	No	No	Yes	No	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: $log(y_{it})$. Standard errors clustered at the team level. Control variables include working hours, share peak hours, and calls per FTE. All regressions include a constant. Pre-treatment performance is defined as average performance over weeks 45/2008 to 49/2008, i.e., before management decided on assignment to the training program. Regressions in Columns (1) to (3) include agents with an estimated propensity score within the 5th to 95th percentiles of the propensity score distribution (cf. Table A.7 in the Online Appendix). Columns (4) to (6) are further restricted to the 20th to 80th percentiles.

In Table 3, Columns (1) to (3) show results where the estimation sample is restricted to the 5th to 95th percentiles of the propensity score distribution (indicated by solid lines in Figure A.1). The specifications are without controls (Column 1), with controls for worker fixed effects (Column 2), or with controls for pre-treatment performance (Column 3). The bias in the specification without controls remains large (78.3% in Column 1), whereas bias in the specifications including worker fixed effects (Column 2) and pre-treatment performance (Column 3) are slightly smaller than for the unrestricted sample (4.1% and

9.8%, respectively). Compared to the unrestricted sample, shown in Table 2, the estimates for the non-experimental sample in Table 3 are relatively similar, with the difference between 0 and 2.5 percentage points. While this seems at odds with the idea that a common support restriction should decrease selection bias, the tails of the propensity score distribution remain unbalanced even when applying the 5th/95th percentile restriction.

Columns (4) to (6) of Table 3 show results when further restricting the sample to between the 20th and 80th percentiles (indicated by dashed lines in Figure A.1). With this admittedly strong common support restriction, both groups are represented over the remaining distribution. Even when not conditioning on control variables or worker fixed effects, bias decreases substantially to 17.6% (Column 4). When further including worker fixed effects or pre-treatment performance, the estimated bias is relatively close to zero (-1.3% in Column 5 and -3.5% in Column 6, respectively). These results suggest that, for the training program examined here, the common support restriction achieves a sizeable reduction in selection bias, especially when requiring the common support restriction to be fairly large.

5 Conclusion

The scarcity of experimental evidence on returns to workplace training has forced academics and policymakers to rely on estimates based primarily on selection on observables or fixed effects approaches. This paper assesses bias in OLS estimators of returns to workplace training using non-experimental data. We apply data from a field experiment with random assignment to training (De Grip and Sauermann, 2012), and compare the causal estimates with a non-experimental sample endogenously chosen not to be part of the field experiment. To the best of our knowledge, this is the first study of its kind concerning returns to workplace training. In fact, the overall scarcity of experimental studies leaves few opportunities to compare different OLS estimates against the benchmark of an experimental estimate.

Our results indicate that the biased (non-experimental) estimate is up to twice the size of the causal estimate. When controlling for individual fixed effects or including pre-treatment performance, the remaining bias is modest, in several specifications below 10 percent. Importantly, the results of this paper are obtained using personnel data from an individual firm where agents are subject to the same labor market, employed by the same firm, and the treated individuals participate in the same training program. The quality of the data may explain why most of the bias is removed even without conditioning on common support. In other settings, e.g., if data is collected from population surveys, lack of common support has been emphasized as a major source of bias (Heckman et al., 1998).

How can we relate our findings to previous studies estimating the returns to formal workplace training? Although it is not possible to assess the degree of selection bias in existing non-experimental studies on returns to formal workplace training, it is possible to assess the change in estimated returns when including worker fixed effects to account for endogenous training participation. Some of the studies shown in the Appendix (Table A.1) provide estimates both with and without worker fixed effects. When *including* worker fixed effects, estimated returns decrease by 61% from an average of 8.9% (OLS) to 3.2% (OLS with worker fixed effects). In comparison, our results in Table 2 show that, for the non-experimental sample, including worker fixed effects reduces the estimate by 38%, from 21.8% (OLS) to 13.5% (OLS with fixed effects). The more modest reduction in our estimates could be explained by the similarity between our sample of treated and untreated individuals along several important dimensions.

An additional factor when interpreting findings based on personnel data is peer effects from treated to untreated workers. For the setting of this field experiment, De Grip and Sauermann (2012) show that untrained workers in the experimental sample increased performance even though they had not received training themselves. The presence of peer effects implies that our estimates of the non-experimental treatment effects are actually underestimated: peer effects increase the non-experimental control group's performance, resulting in lower treatment effects. While these general equilibrium effects are especially important for studies using personnel data that capture social interaction, they can affect any evaluation of workplace training.

References

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge (2017): “When Should You Adjust Standard Errors for Clustering?” NBER Working Paper 24003.

Adhvaryu, Achyuta, Namrata Kala, and Anant Nyshadham (2018): “The Skills to Pay the Bills: Returns to On-the-job Soft Skills Training,” NBER Working Paper 24313.

Agodini, Roberto and Mark Dynarski (2004): “Are experiments the only option? A look at dropout prevention programs,” *Review of Economics and Statistics*, 86(1): 180–194.

Alfonsi, Livia, Oriana Bandiera, Vittorio Bassi, Robin Burgess, Imran Rasul, Munshi Sulaiman, and Anna Vitali (2019): “Tackling Youth Unemployment: Evidence from a Labour Market Experiment in Uganda,” unpublished manuscript.

Arulampalam, Wiji and Alison L. Booth (2001): “Learning and Earning: Do Multiple Training Events Pay? A Decade of Evidence from a Cohort of Young British Men,” *Economica*, 68(271): 379–400.

Autor, David H. and David Dorn (2013): “The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market,” *American Economic Review*, 103(5): 1553–97.

Barron, John M, Mark C Berger, and Dan A Black (1997): “How Well Do We Measure Training?” *Journal of Labor Economics*, 15(3): 507–28.

Bartel, Ann P (1995): “Training, Wage Growth, and Job Performance: Evidence from a Company Database,” *Journal of Labor Economics*, 13(3): 401–25.

Bartel, Ann P. (2000): “Measuring the Employer’s Return on Investments in Training: Evidence from the Literature,” *Industrial Relations*, 39(3): 502–524.

Bassanini, Andrea, Alison Booth, Giorgio Brunello, Maria De Paola, and Edwin Leuven (2007): “Workplace Training in Europe,” in *Education and Training in Europe*, ed. by Giorgio Brunello, Pietro Garibaldi, and Etienne Wasmer, Oxford: Oxford University Press, chap. 8-13.

Batt, Rosemary, Virginia Doellgast, and Hyunji Kwon (2005): "Service Management and Employment Systems in U.S. and Indian Call Centers," in *Brookings Trade Forum 2005: Offshoring White-Collar Work—The Issues and Implications*, ed. by Susan M. Collins and Lael Brainard, Washington, D.C.: Brookings Institution Press, 335–372.

Beyer, Joy De (1990): "The incidence and impact on earnings of formal training provided by enterprises in Kenya and Tanzania," *Economics of Education Review*, 9(4): 321–330.

Bifulco, Robert (2012): "Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison," *Journal of Policy Analysis and Management*, 31(3): 729–751.

Black, Dan A., Joonhwi Joo, Robert LaLonde, Jeffrey Andrew Smith, and Evan J. Taylor (2017): "Simple Tests for Selection: Learning More from Instrumental Variables," CESifo Working Paper Series 6392.

Blau, David M. and Philip K. Robins (1987): "Training Programs and Wages: A General Equilibrium Analysis of the Effects of Program Size," *Journal of Human Resources*, 22(1): 113–125.

Blundell, Richard, Lorraine Dearden, and Costas Meghir (1996): *The determinants and effects of work-related training in Britain*, R50, IFS Reports, Institute for Fiscal Studies.

Booth, Alison L. (1991): "Job-Related Formal Training: Who Receives It And What Is It Worth?" *Oxford Bulletin of Economics and Statistics*, 53(3): 281–294.

——— (1993): "Private Sector Training and Graduate Earnings," *Review of Economics and Statistics*, 75(1): 164–170.

Booth, Alison L. and Mark L. Bryan (2005): "Testing Some Predictions of Human Capital Theory: New Training Evidence from Britain," *Review of Economics and Statistics*, 87(2): 391–394.

Booth, Alison. L., Marco Francesconi, and Gylfi Zoega (2003): "Unions, Work-Related Training, and Wages: Evidence for British Men," *Industrial and Labor Relations Review*, 57(1): 68–91.

Breuer, Kathrin, Petra Nieken, and Dirk Sliwka (2013): "Social ties and subjective performance evaluations: An empirical investigation," *Review of Managerial Science*, 7(2): 141–157.

Brown, James N. (1989): "Why Do Wages Increase with Tenure? On-the-Job Training and Life-Cycle Wage Growth Observed within Firms," *American Economic Review*, 79(5): 971–991.

Budria, Santiago and Pedro Telhado Pereira (2007): "The Wage Effects of Training in Portugal: Differences across skill groups, genders, sectors, and training types." *Applied Economics*, 39: 787–807.

Bureau of Labor Statistics (2018): "Occupational Employment Statistics," May 2018, Bureau of Labor Statistics (BLS).

Cedefop (2009): *Using tax incentives to promote education and training*, Luxembourg: Office for Official Publications of the European Communities.

De Grip, Andries and Jan Sauermann (2012): "The Effects of Training on Own and Co-Worker Productivity: Evidence from a Field Experiment," *Economic Journal*, 122(560): 376–399.

De Grip, Andries, Jan Sauermann, and Inge Sieben (2016): "Tenure-Performance Profiles and the Role of Peers: Evidence from Personnel Data," *Journal of Economic Behavior & Organization*, 126: 39–54.

Dearden, Lorraine, Howard Reed, and John Van Reenen (2006): "The Impact of Training on Productivity and Wages: Evidence from British Panel Data," *Oxford Bulletin of Economics and Statistics*, 68(4): 397–421.

Diaz, Juan Jose and Sudhanshu Handa (2006): "An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico's PROGRESA Program," *Journal of Human Resources*, 41(2): 319–345.

Dimitriadis, Stefan and Rembrand Koning (2019): “The value of communication:Evidence from a field experiment with entrepreneurs in Togo,” Tech. rep., SSRN Working Paper Series No. 3459643.

Evertsson, Marie (2004): “Formal On-the-Job Training: A Gender-Typed Experience and Wage-Related Advantage?” *European Sociological Review*, 20(1): 79–94.

Freifeld, Lorri (2018): *2018 Training Industry Report*, Training Magazine.

Gordon, Brett R., Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky (2019): “A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook,” *Marketing Science*, 38(2): 193–225.

Goux, Dominique and Eric Maurin (2000): “Returns to firm-provided training: evidence from French worker-firm matched data,” *Labour Economics*, 7(1): 1–19.

Griffen, Andrew S. and Petra E. Todd (2017): “Assessing the Performance of Nonexperimental Estimators for Evaluating Head Start,” *Journal of Labor Economics*, 35(S1): S7–S63.

Haelermans, Carla and Lex Borghans (2012): “Wage Effects of On-the-Job Training: A Meta-Analysis,” *British Journal of Industrial Relations*, 50(3): 502–528.

Heckman, James J. (1979): “Sample Selection Bias as a Specification Error,” *Econometrica*, 47(1): 153–161.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra Todd (1998): “Characterizing Selection Bias Using Experimental Data,” *Econometrica*, 66(5): 1017–98.

Hill, Elizabeth T. (1995): “Labor Market Effects Of Women’s Post-school-age Training.” *Industrial and Labor Relations Review*, 49(1): 138–149.

Imbens, Guido W and Jeffrey M Wooldridge (2009): “Recent developments in the econometrics of program evaluation,” *Journal of Economic Literature*, 47(1): 5–86.

Kawaguchi, Daiji (2006): “The Incidence and Effect of Job Training among Japanese Women,” *Industrial Relations: A Journal of Economy and Society*, 45(3): 469–477.

Konings, Jozef and Stijn Vanormelingen (2015): "The Impact of Training on Productivity and Wages: Firm-Level Evidence," *Review of Economics and Statistics*, 97(2): 485–497.

Krueger, Alan and Cecilia Rouse (1998): "The Effect of Workplace Education on Earnings, Turnover, and Job Performance," *Journal of Labor Economics*, 16(1): 61–94.

LaLonde, Robert J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76(4): 604–620.

Leuven, Edwin and Hessel Oosterbeek (2004): "Evaluating the effects of a tax deduction on training," *Journal of Labor Economics*, 22(2): 461–488.

——— (2008): "An Alternative Approach to Estimate the Wage Returns to Private-Sector Training," *Journal of Applied Econometrics*, 23(4): 423–434.

Lillard, Lee and Hong Tan (1992): "Private Sector Training: Who Get's it and What Are Its Effects?" *Research in Labor Economics*, 13(1).

Liu, Xiangmin and Rosemary Batt (2007): "The Economic Pay-Offs to Informal Training: Evidence from Routine Service Work," *Industrial and Labor Relations Review*, 61(1): 75–89.

Loewenstein, Mark A. and James R. Spletzer (1998): "Dividing the Costs and Returns to General Training," *Journal of Labor Economics*, 16(1): 142–171.

Lynch, Lisa M. (1992): "Private-Sector Training and the Earnings of Young Workers," *American Economic Review*, 82(1): 299–312.

Marcotte, Dave E. (2000): "Continuing Education, Job Training, and the Growth of Earnings Inequality," *Industrial and Labor Relations Review*, 53(4): 602–623.

McKenzie, David, Steven Stillman, and John Gibson (2010): "How Important is Selection? Experimental vs. Non-Experimental Measures of the Income Gains from Migration," *Journal of the European Economic Association*, 8(4): 913–945.

O'Connell, Philip J. and Delma Byrne (2010): "The Determinants and Effects of Training at Work: Bringing the Workplace Back in," *European Sociological Review*, 28(3): 283–300.

Parent, Daniel (1999): “Wages and Mobility: The Impact of Employer-Provided Training,” *Journal of Labor Economics*, 17(2): 298–317.

——— (2003): “Employer-supported training in Canada and its impact on mobility and wages,” *Empirical Economics*, 28(3): 431–459.

Pischke, Jörn-Steffen (2001): “Continuous training in Germany,” *Journal of Population Economics*, 14(3): 523–548.

Pischke, Jörn-Steffen (2007): “Comment on ‘Workplace training in Europe’, by Andrea Bassanini et al.” in *Education and Training in Europe*, ed. by Giorgio Brunello, Pietro Garibaldi, and Etienne Wasmer, Oxford: Oxford University Press, 330–342.

Prada, Maria, Graciana Rucci, and Sergio Urzua (2019): “Training, Soft Skills and Productivity: Evidence from a Field Experiment,” IZA Discussion Paper 12447.

Salas-Velasco, Manuel (2009): “Beyond lectures and tutorials: Formal on-the-job training received by young European university graduates,” *Research in Economics*, 63(3): 200–211.

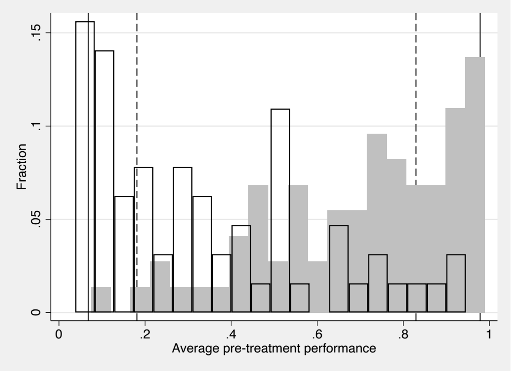
Schone, Pal (2004): “Firm-financed training: Firm-specific or general skills?” *Empirical Economics*, 29(4): 885–900.

Sieben, Inge, Andries De Grip, Danielle Van Jaarsveld, and Ole Sørensen (2009): “Technology, Selection, and Training in Call Centers,” *Industrial and Labor Relations Review*, 62(4): 553–572.

Vignoles, Anna, Fernando Galindo-Rueda, and Leon Feinstein (2004): “The Labour Market Impact of Adult Education and Training: A Cohort Analysis,” *Scottish Journal of Political Economy*, 51(2): 266.

Online appendix

Figure A.1: Propensity scores for experimental sample and non-experimental control group



Notes: This figure shows the estimated propensity score for agents in the field experiment (experimental treatment group and experimental control group: gray bars) and for agents in the non-experimental control group (white framed bars). The vertical solid (dashed) lines indicate the 5th and 95th (20th and 80th) percentiles of the propensity score distribution.

Table A.1: Estimates of returns to participating in formal workplace training programs on worker outcomes

Study	*Outcome variable*	*Estimate*	*Method*	*Details*
Adhvaryu et al. (2018)	Worker productivity Wages	20% 0.5%	DiD (field experiment)	Female garment workers in India (2013-15); modular training with 80 training hours in total. Up to 18 month follow-up
Arulampalam and Booth (2001)	Wage growth	2.3% (n.s.) 34.2%	OLS IV	Survey data from UK (National Child Development Survey, 1981 and 1991). Cohort of men aged 23 in 1981. Training defined as at least one training course between 1981 and 1991.
Barron et al. (1997)	Wage growth	0.2% (0.3%)	OLS	Data from SBA 1992 (EOPP 1982). Effect on having participated in training between hiring and two years later. Estimate is percentage increase in wage growth for a 10% increase in training participation.
Bartel (1995)	Wage growth	1% 10.6%	OLS IV	Personnel data from large manufacturing firm (1986-1990). Various training programs implemented in the firm
Bassanini et al. (2007)	Hourly wages	3.7-21.6% -3%-10.5% (partly n.s.)	OLS OLS (fixed effects)	Data from the European Community Household Panel (1995, -97, -99, 2001). Estimates by country. Training defined as accumulated training over the survey years.
Blau and Robins (1987)	Wage growth	14.8% (13.9%) 4.8% (3.7%)	OLS OLS (w/ pre-training wages)	Data from EOPP 1980. Effect on having participated in training during employment spell. Results for men (women)
Blundell et al. (1996)	Wage growth	3.6% (4.8%) 4.1% (0.3%-n.s.)	OLS Heckman (1979) correction	Survey data from UK (National Child Development Survey, 1981 and 1991). Training in current job. Results for men (women)
Booth (1991)	Wages	10.6% (16.6%)	Censored regression	Cross-sectional data for UK (1987). Effects for men (women). Censored regression estimated with maximum likelihood due to banded earnings data.
Booth (1993)	Wages	-1.5% (-2.7%) – n.s. -1.7% (.2%) – n.s.	OLS (on-the-job) OLS (inhouse)	National Survey of 1980 Graduates and Diplomates (1986-87). Results for in-house training courses / on-the-job training. Effects for men (women)
Booth and Bryan (2005)	Wages	2.4%	OLS (fixed effects)	BHPS (1998-2000). Results for training in current job
Booth et al. (2003)	Hourly wages	3.3% 1.0% (n.s.)	OLS OLS (fixed effects)	BHPS (1991-96). Results for training in current job
Brown (1989)	Hourly wages	21.8%	OLS (fixed effects)	PSID 1976-1986. Training is defined as cumulative training.
Budria and Pereira (2007)	Wages	12.7% (8.4%) 30.3% (37.5%)	OLS 2SLS	Portuguese Labor Force Survey (LFS, 1998-2000). Effects for men (women)
Dearden et al. (2006)	Wages	0.3%	GMM estimation	UK Labor Force Survey (LFS, 1983-96). Authors also estimate establishment returns. Estimate refers to a 1% increase in share of workers trained.
Beyer (1990)	Earnings	3.9% (15.5%)	OLS	Survey among firms in Kenya (Tanzania) in 1980.
De Grip and Sauermann (2012)	Worker productivity	10%	OLS (field experiment)	Call agents in the Netherlands (2008-2009); one-week training course; short follow-up ($\leq$ 12 weeks)
Evertsson (2004)	Earnings	6% (4%)	OLS	Level of Living Conditions Survey in Sweden (1994-98). Results are for men (women).
Goux and Maurin (2000)	Wages	6.6% -5.7% (n.s.)	OLS Bivariate probit	Household data from France (1988-93). Estimation model accounts for both decision to receive training and job-to-job mobility
Hill (1995)	Wage growth	4-6%	OLS	NLS Mature Women's Cohort (1984)
Kawaguchi (2006)	Wages	3% 1.8% (n.s.)	OLS OLS (fixed effects)	Japanese panel data for women (1994 and 1998)

...continues on next page

Table A.1: (*continued*)

Study	*Outcome variable*	*Estimate*	*Method*	*Details*
Konings and Vanormelingen (2015)	Wages	20% 17%	OLS ACF	Belgian firm database 1997 to 2006. Authors also estimate effects on firm productivity. ACF refers to estimation of the production function in which all inputs are specified.
Krueger and Rouse (1998)	Wages	1.9% (0.5%)	OLS (fixed effects)	Service sector (manufacturing) company in US. Performance measures (nominations, awards) are not defined as logarithms
Leuven and Oosterbeek (2004)	Wages	3% (n.s.) -6.3% (n.s.)	OLS 2SLS	Dutch survey data aged 16-64 from 1994 and 1999. 2SLS identification based on age-discontinuity
Leuven and Oosterbeek (2008)	Wages	10.6% 0.9% (n.s.)	"Endog." comparison "Exog." comparison	Dutch population aged 16-64. Approach asks survey respondents for reasons for (non) participation. Endogenous comparison includes all participants and non-participants; exogenous comparison includes non-participants only with "random" reasons for non-participation.
Lillard and Tan (1992)	Wages	5.6% (11.9%)	OLS	Current population survey (CPS) for men, 1982 (National longitudinal survey (NLS) young men 1966-69)
Liu and Batt (2007)	Worker productivity	0.06%	OLS (fixed effects)	Call agents in a large US telecommunications company. Estimate shows productivity increase for 10% increase in training hours.
Loewenstein and Spletzer (1998)	Wages	2.7% (n.s.) 3.5%	OLS OLS (fixed effects)	Data from NLSY 1988-91. Individuals aged 23-34. Training program or on-the-job training to improve job skills or learn a new job.
Lynch (1992)	Wages	0.2% 0.2%	OLS Heckman (1979) correction	NLSY data (1980-83). Individuals aged 16-26. Training in the current job
Marcotte (2000)	Wages	10.5-14%	OLS	NLS (1981) and NLSY (1993)), restricted to white males.
O'Connell and Byrne (2010)	Wages	3.6%	OLS	Irish Survey of EmployeesÕ Attitudes and Experiences of the Workplace (2003)
Parent (1999)	Hourly wages	16.9% 11.5%	OLS IV	NLSY (1979-91)
Parent (2003)	Hourly earnings	11.9% (8.3%) 10.3% (1.7%–n.s.)	OLS OLS (fixed effects)	Data for Canada. Non-college graduates aged 18-20 in 1991 and re-interviewed in 1995. Any career- or work-related training in the current job. Results for men (women)
Pischke (2001)	Wages	1.2% (n.s.)	OLS (fixed effects)	Data for Germany (German Socio-Economic Panel), 1986-1989. Any type of work-related training in the three years prior to interview
Prada et al. (2019)	Sales	10-12.1%	DiD (field experiment)	Store managers and sales associates of retail chain in Chile (2014-16). Follow-up between 3.5 and 6 months.
Salas-Velasco (2009)	Wages	12.4% 52.4% (n.s.)	OLS Heckman (1979) correction	Survey on European graduates who graduated in 1994-95
Schone (2004)	Hourly wages	3.6% 3.7% (n.s.)	OLS IV	Norwegian Survey of Organizations and Employees (NSOE) in 1993. Training defined as training participation in 1993
Vignoles et al. (2004)	Wage growth	4.8% 5.0% (n.s.)	OLS IV	Survey data from UK (National Child Development Survey, 1981 and 1991). Results for men

Notes: n.s.–not significant. Unless otherwise mentioned, OLS results include larger sets of control variables. All estimates for training incidence (unless mentioned otherwise).

Table A.2: Group definitions of all 157 agents

Name	Description	Number of agents
Field experiment		
Treatment group	Agents randomly assigned to training in weeks 10/2009 to 14/2009.	$N = 34$
Control group	Agents randomly assigned to training after week 24/2009.	$N = 40$
Re-assigned agents	Agents initially assigned to treatment or control group but re-assigned, e.g., due to illness or vacation plans (see Footnote 10). These agents are not included in our analysis.	$N = 10$
Non-experimental control group		
G1	Agents initially not selected to be part of the field experiment, but were eventually trained during the sample period.	$N = 7$
G2	Agents initially not selected to be part of the field experiment who were not trained during the sample period.	$N = 66$

Notes: This table summarizes the groups portion of this study. All agents in all groups are observable over the full sample period from week 45/2008 to week 24/2009.

Table A.3: Variable definitions

Variable name	Definition
Gender	1 if male agent, zero otherwise
Age	Age measured in years at start of observation period
Tenure	Tenure measured in months at start of observation period in week 45/2008
Working hours	Number of actual working hours in week t
Share peak hours	Share of agent i's hours worked during peak hours (defined as hours between 12:00 and 18:00)
Pre-treatment performance	Average performance over weeks 45/2008 to 49/2008, i.e., before management decided on assignment to the training program
Turnover	Dummy equaling 1 if agent exits department before end of observation period, and zero otherwise. Exiting agents could either move to other departments or leave the firm entirely
Calls per FTE	Total number of incoming calls normalized by number of full-time equivalents working in the same week
Common trend	Linear time trend

Table A.4: Descriptive statistics for groups G1 and G2

	(1)	(2)	(3)	(4)	(5)
			t-test	*t*-test	*t*-test
	G1	G2	G1-G2	G1-Exp. TG	G2-Exp. TG
Gender (1=male)	0.143	0.318	0.175	0.239	0.064
	(0.378)	(0.469)	(0.154)	(0.166)	(0.102)
Age (in years)	23.045	30.548	7.503***	11.025***	3.521
	(4.180)	(11.087)	(2.227)	(2.431)	(2.246)
Tenure (in months)	22.000	15.485	-6.515	31.353	37.868***
	(43.780)	(28.626)	(16.918)	(18.604)	(9.204)
Working hours	24.091	20.522	-3.569	-5.038*	-1.470
	(5.579)	(7.128)	(2.284)	(2.373)	(1.398)
Share peak hours	0.596	0.561	-0.034*	-0.042	-0.007
	(0.043)	(0.064)	(0.018)	(0.024)	(0.020)
Pre-treatment performance	0.305	0.283	-0.022	0.059	0.081***
	(0.080)	(0.088)	(0.034)	(0.035)	(0.017)
Turnover	0.143	0.530	0.387**	0.063	-0.324***
	(0.378)	(0.503)	(0.156)	(0.159)	(0.094)
Number of agents	7	66	73	41	100

Notes: *** p<0.01, ** p<0.05, * p<0.1. The sample used in G1 is defined as all agents from the non-experimental sample who were assigned by management to be treated with the treatment group during the training period. G2 includes all agents who were *not* trained during the observation period used in this sample. Columns (1) and (2) show means and standard deviations in parentheses for Groups G1 and G2; Columns (3) to (5) show differences between G1 and G2, differences between G1 and the experimental treatment group, and differences between G2 and the experimental treatment group, respectively. Columns (3) to (5) show standard errors in parentheses.

Table A.5: Treatment effects for varying covariates with varying control group definitions

Dependent variable: logarithm of worker performance						
	(1)	(2)	(3)	(4)	(5)	(6)
*A: TG-G1 (*N=41, n=1,049*)*						
Treatment dummy	0.1472***	0.1420***	0.1415***	0.1647**	0.1346***	0.1386***
	(0.0200)	(0.0207)	(0.0216)	(0.0427)	(0.0124)	(0.0174)
Adjusted R-squared	0.1067	0.1270	0.1357	0.1276	0.1780	0.4897
*B: TG-G2 (*N=100, n=2,190*)*						
Treatment dummy	0.2205***	0.2179***	0.1769***	0.1261***	0.1252***	0.1143***
	(0.0402)	(0.0417)	(0.0193)	(0.0211)	(0.0098)	(0.0138)
Adjusted R-squared	0.0832	0.0996	0.1305	0.1421	0.0472	0.4760
Control variables	No	Yes	Yes	Yes	No	No
Tenure (linear + squared)	No	No	Yes	No	No	No
Common trend	No	No	No	Yes	No	No
Worker FE	No	No	No	No	Yes	No
Pre-treatment performance	No	No	No	No	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: $log(y_{it})$. Standard errors clustered at the team level. Control variables include working hours, share peak hours, and calls per FTE. All regressions include a constant. Pre-treatment performance is defined as average performance during weeks 45/2008 to 52/2008. TG-G1 contains all agents from the non-experimental sample assigned by management to be treated with the treatment group during the training period. G2 includes all agents who were *not* trained during the observation period used in this sample.

Table A.6: Treatment effects of workplace training for non-leavers

Dependent variable: logarithm of worker performance	(1)	(2)	(3)
A: Experimental sample			
Treatment dummy ($\widehat{\tau}_E$)	0.0989***	0.1267***	0.1041***
	(0.0177)	(0.0095)	(0.0138)
Adjusted R-squared	0.0276	0.0732	0.4240
Number of agents	57	57	57
Number of observations	1,558	1,558	1,558
B: Non-experimental sample			
Treatment dummy ($\widehat{\tau}_N$)	0.1704***	0.1357***	0.1169***
	(0.0272)	(0.0114)	(0.0211)
Adjusted R-squared	0.0865	0.0756	0.3804
Number of agents	64	64	62
Number of observations	1,722	1,722	1,684
C:Selection bias			
$\widehat{\tau}_N - \widehat{\tau}_E$	0.0715	0.0090	0.0129
p-value ($\widehat{\tau}_N - \widehat{\tau}_E$)	0.0263	0.3527	0.5370
Bias in % (($\widehat{\tau}_N - \widehat{\tau}_E$)/$\widehat{\tau}_E$*100)	72.3	7.1	12.4
Control variables	No	No	No
Tenure (linear + squared)	No	No	No
Common trend	No	No	No
Worker FE	No	Yes	No
Pre-treatment performance	No	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: $log(y_{it})$. Standard errors clustered at the team level. The sample is restricted to agents who remained in the department until the end of the observation period in week 24/2009.

Table A.7: Propensity score estimation

	(1)
Gender (1=male)	0.0668
	(0.2704)
Age (in years)	-0.0005
	(0.0147)
Tenure (in months)	0.0515***
	(0.0166)
Tenure (sq.)	-0.0003**
	(0.0001)
Working hours	0.0451**
	(0.0194)
Share peak hours	-1.1862
	(1.2690)
Pre-treatment performance	7.3280***
	(1.7331)
Constant	-3.4666***
	(1.0538)
Number of agents	137
Pseudo R-squared	0.3030

Notes: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: dummy whether agent i is part of the field experiment or not. The regression includes all agents used for estimating Table 2). For 10 of the 147 agents, there is either no information on age or no information on pre-treatment performance available.