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

Jan Sauermann† Anders Stenberg‡

June 2021

Abstract

We assess selection bias in estimated returns to workplace training by exploiting a field experiment with random assignment of workers to a one-week training program. We compare the causal estimate, which is estimated using randomized treatment and control groups, with a non-experimental estimate, which uses an endogenously selected control group. The results show that non-experimental estimates are biased, yielding returns about twice as large as the causal effect. When controlling for pre-treatment performance or individual fixed effects, only about one tenth of this bias remains. Similarly, the bias is almost entirely removed when applying common support restrictions.

Keywords: returns to training, selection bias, field experiment

JEL Classification Numbers: J24, C93, M53

*We thank Michael Burda, Sebastian Butschek, Andries De Grip, Matthias Heinz, Edwin Leuven, José Montalbán Castilla, Jan Nimczik, Dirk Sliwka, Frederik Thenee, and Petra Todd, as well as seminar and conference participants at the University of Cologne, EALE 2019, Humboldt University Berlin, the IZA workshop Workplace and management practices in Dublin, and the University of Potsdam. Jan Sauermann thanks the Swedish Research Council (2017-03432) and the Jan Wallanders och Tom Hedelius Stiftelse for financial support (P2014-0236:1). The field experiment used in this study is registered as AEARCTR-0004967 in the AEA RCT registry.

†Institute for Evaluation of Labour Market and Education Policy (IFAU), Box 513, 75120 Uppsala, +46 737 169 102; CCP, Copenhagen Business School; Institute of Labor Economics (IZA); ROA, Maastricht University; UCLS, Uppsala University. *E-mail*: jan.sauermann@ifau.uu.se

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

1 Introduction

In labor markets faced by changing tasks and new skill demands, investing in skills is important for both individuals and firms (Acemoglu and Pischke, 1998; Acemoglu and Autor, 2010). Estimates for the US show that employees receive an average of 47 hours of training per year, with firms' total investments in formal workplace training estimated at 87.6 billion USD in 2018 (Freifeld, 2018). Despite a large number of influential papers, several of which were cited more than 500 times, there is no consensus whether and how strongly workplace training affects worker productivity. A major reason behind this lack of consenus is the difficulty to account for selection bias due to endogenous training participation (see, e.g., Bassanini et al., 2007; Pischke, 2007). Reliable estimates on the causal returns to training, however, are important for managers making decisions about whether or not to invest in workplace training, for decisions on whether governments should allocate resources to stimulate workplace training (Cedefop, 2009), and more generally to better understand the development of human capital, productivity and wages throughout the worker life-cycle. While experimental methods are increasingly used in the estimation of the returns to training (Murthy et al., 2008; De Grip and Sauermann, 2012; Adhvaryu et al., 2018; Prada et al., 2019), experiments are not always feasible or even possible. It is therefore important to learn to what extent non-experimental evaluations using standard econometric methods reduce selection bias.

The aim of this study is to explore whether endogenous training participation causes selection bias in estimated returns to workplace training, and whether standard regression techniques based on non-experimental data are capable of reducing bias in the estimated

treatment effects. We exploit data from 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. This experiment provides an unbiased estimate of the returns to workplace training (De Grip and Sauermann, 2012). We compare this estimate with a non-experimental estimate based on a sample of workers that were selected to *not* be included in the experiment. These individuals, who worked in the same firm during the same time period, therefore constitute an endogenously selected sample of non-participants (henceforth referred to as the non-experimental control group). This set-up allows us to use the experimental estimate as a benchmark, against which the potentially biased non-experimental estimates can be compared, and it enables us to assess the bias remaining in estimators when using standard regression techniques. We first replicate the main finding of the field experiment from De Grip and Sauermann (2012), showing that the workplace training program led to an increase in productivity, as measured by average handling time, of 10.9%. We then use the non-experimental control group, and find that the same specification leads to an estimated return to training of 21.8%, i.e., almost twice the experimental estimate and indicating a bias of 99.1%. By expanding the model to control for pre-treatment performance as a proxy of worker ability, or to include worker fixed effects to capture unobserved individual-specific factors, the bias remaining in the non-experimental estimates are reduced from 99.1% to 13.2% and 7.5% respectively, i.e. similar to those reported from the experimental sample. This suggests that relatively parsimonious non-experimental approaches can yield estimates with modest bias. 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.

Our empirical strategy essentially follows the practice of the small but important literature assessing selection bias in non-experimental estimates, originating from studies analyzing the effectiveness of training programs targeting disadvantaged groups. These studies exploit random variation in treatment status to analyze to what extent different econometric methods account for selection, e.g. difference-in-differences, (propensity score) matching or control function approaches. Seminal studies in this literature such as LaLonde (1986) and Fraker and Maynard (1987) assessed the bias in non-experimental evaluations of the National Supported Work Demonstration (NSW) program finding that non-experimental data deviated substantially from the experimental estimates. This indicates that it would be difficult for a researcher to uncover the causal effect without a randomized control group. Heckman et al. (1998), however, argue that the poor performance of non-experimental estimators was in part due to issues of data quality and that greater homogeneity among treated and non-treated individuals reduces bias. This can be achieved, for example, by focusing on data for treated and non-treated individuals that are collected from the same data sources, if treated and non-treated individuals are part of the same regional labor markets, and if treated individuals participate in similar training programs. Their results indicate that the remaining bias is only a fraction (7%) of the conventional measure of selection bias and not statistically significantly different from zero (Heckman et al., 1998, Table V).[1]

[1]Other assessments of selection bias have concerned large-scale policy programs such as PROGRESA, an antipoverty program in rural areas of Mexico (Diaz and Handa, 2006; Bifulco, 2012), (early) childhood interventions such as Head Start (Griffen and Todd, 2017) and Project STAR (Wilde and Hollister, 2007),

The present study is foremost related to the long-standing literature quantifying the returns to workplace training on workers' wages or worker productivity and is the first to evaluate selection bias for workplace training programs provided by firms. In the literature on the returns to workplace training, the returns to training on wages are typically estimated using cross-sectional or yearly survey data with self-reported information on training participation (see, e.g., Bassanini et al., 2007). While the impact on wages is interesting in itself, it only reflects the workers' payoff to training. From a societal point of view, the effect of training on productivity is arguably more interesting.[2] 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; Murthy et al., 2008; De Grip and Sauermann, 2012). Besides having direct measures of worker productivity, focusing on individual firms also allows researchers to hold constant data sources, labor market conditions, and the type of training program (cf. Heckman et al., 1998).

Table A.1 in the Online Appendix provides an overview of estimated effects of training on wages and worker productivity, displaying a large degree of heterogeneity among estimates.[3] Non-experimental ordinary least squares (OLS) estimates of the wage returns to training show an average return of 8.7% with a standard deviation of 5.7%.[4] To account for

programs to prevent dropping out of high-school (Agodini and Dynarski, 2004) and income gains from migration (McKenzie et al., 2010).

[2]For instance, social cost-benefit calculations of public policies typically seek to determine whether productivity increases are sufficient to cover the associated social costs.

[3]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.

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

worker selection into training, some studies seek to elicit exogenous variation by exploiting age-cutoffs in tax deductions of training investments (Leuven and Oosterbeek, 2004), unanticipated withdrawals from training (Leuven and Oosterbeek, 2008; Görlitz, 2011), or by using worker rank measures to instrument for training investments (Bartel, 1995). The majority of studies, however, take selection into account by controlling for observable characteristics or worker fixed effects. The general conclusion of these studies is that accounting for selection reduces the estimated effects of training on wages. To which degree these estimates recover the causal estimate, however, is difficult to assess in the absence of experimental variation of the treatment.

To the best of our knowledge, only four studies use randomized assignment to training to evaluate effects of workplace training. These studies use personnel data of individual firms with direct measures of worker productivity and report returns to training between 10% and 21% (Murthy et al., 2008; De Grip and Sauermann, 2012; Adhvaryu et al., 2018; Prada et al., 2019). Randomized participation in the evaluated training programs allow to make causal interpretations regarding returns to workplace training programs.[5]

The contribution of our study is to provide a first assessment of the extent to which it is possible to recover the causal estimate of participation in workplace training programs using non-experimental data. Due to the scarcity of experimental approaches in the literature on the returns to training, there are no studies explicitly assessing selection bias in workplace training programs by comparing non-experimental estimates to an experi-

[5]There is also a related strand of literature that evaluates training programs focused on soft-skill training programs targeted at adolescents and entrepreneurs in developing countries (see, e.g., Groh et al., 2012; Acevedo et al., 2017; Campos et al., 2017; Dimitriadis and Koning, 2019; Ashraf et al., 2020). Also related to this study, Alfonsi et al. (2020) report that randomly offering firms wage subsidies to train workers on-the-job shows no significant earnings returns.

mental benchmark. We find that controlling for measures of pre-treatment performance, or controlling for worker fixed effects, removes most of the bias in OLS estimates using non-experimental data. We also find that applying common support restrictions further reduces selection bias in the returns to training. These results provide guidance for evaluations of workplace training when, as is most often the case, randomized control trials are not feasible or even not possible.

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 2018, or 2.7% of the total workforce (Batt et al., 2005; Bureau of Labor Statistics, 2018). Call center employees also represent a non-negligible share of workers in the service sector occupations that, as a consequence labor market polarization, gained both in overall hours worked and real hourly wages (Autor and Dorn, 2013).

2 Setting

2.1 Institutional setting

The field experiment was conducted in the call center of a multinational telecommunications company located in the Netherlands.[6] 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.[7] 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.[8] 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, shorter train-

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

[7]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.

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

ing 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 treated (weeks 15/2009-24/2009).

In January 2009, the training program was announced in a general message from man-

agement. 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 experimental 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.[9] 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%.

The 74 agents who were assigned to participate in the field experiment were randomly assigned to be treated during the treatment period (experimental treatment group: $N = 34$), or to be treated after the end of the experiment (experimental control group: $N = 40$). 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.

[9]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. Results are available upon request from the authors.

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.[10]

We additionally make use of the 73 agents who were *not* included in the field experiment. Because these agents were *endogenously* selected by management not to participate in the field experiment and the training program, they do not have a randomized treatment status. These agents constitute the *non-experimental control group*.[11] 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 the field experiment likely reflects both their lower tenure, and unobservable factors that management used for assigning agents to the field experiment.[12] Agents in the non-experimental sample are also more likely to leave the department.[13]

[10]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.

[11]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. The remaining agents in the non-experimental control group ($N = 66$), denoted G2, did not participate in the training program. Descriptive statistics for the two groups, G1 and G2, and corresponding t-tests are shown in Table A.4 in the Online Appendix. Agents in the non-experimental control group have not previously been analyzed.

[12]When analyzing the determinants of the probability of participating in the field experiment, tenure and pre-treatment performance are the two most important factors and explain about 14% of the overall variation.

[13]If sample attrition is correlated with either unobserved ability or treatment status, it could bias our estimation results. We further explore this in Section 5.1.

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

	(1) All agents	(2) Experimental treat. group	(3) Experimental control group	(4) Non-experimental control group	(5) Difference (2)-(3)	(6) Difference (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: Columns (1) to (4) show means and standard deviations in parentheses; Columns (5) and (6) show differences between the experimental treatment group and the experimental control group (Column 5), and the experimental treatment group and the non-experimental control group (Column 6), respectively. Parentheses and asterisks in Columns (5) and (6) are from a two-sided t-test on the respective differences (*** p<0.01, ** p<0.05, * p<0.1). 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.

2.3 Measuring performance

To measure performance of individual call agents, it is important to have a measure that is comparable between 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; Murthy et al., 2008; 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.

3 Estimation framework

The main challenge when estimating treatment effects is that counterfactual outcomes are not observed. The post-treatment outcome of a treated individual, denoted Y_{1i}, cannot be compared to the unobserved outcome of the same individual who was not treated Y_{0i}. In a setting with randomized assignment to a treatment D_i, i.e. where $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i$, the average outcome of the untreated $E[Y_{0i}]$ is an unbiased estimate of the counterfactual outcome. The average treatment effect (ATE) is then $\tau_{\text{ATE}} = E[Y_{1i}] - E[Y_{0i}]$. When estimating treatment effects using non-experimental data, however, one must consider systematic differences in individual characteristics X_i between treated and untreated individuals. Assuming homogenous treatment effects, the estimated difference between treated

[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).

and untreated agents can then be written as:

$$E[Y_{1i}|D_i = 1, X_i] - E[Y_{0i}|D_i = 0, X_i] = \tag{1}$$

$$\tau_{\text{ATE}} + E[Y_{0i}|D = 1, X_i] - E[Y_{0i}|D = 0, X_i]$$

where the latter terms represent the selection bias ($E[Y_{0i}|D = 1, X_i] - E[Y_{0i}|D = 0, X_i]$).

In our regression framework, we follow the workplace training literature in that we present estimates of the returns to training by regressing the logarithm of worker productivity on a dummy, being 1 if an individual is treated, and 0 otherwise. The estimation equation can be written as

$$log(y_{it}) = \alpha + \tau d_{it} + \beta_1 X_{it} + \beta_2 t_t + u_{it} \tag{2}$$

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 (2) 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]

Equation (2) is estimated for two samples, the experimental sample which comprises all agents that were included in the field experiment and which are randomly assigned to treatment and control groups. Given the random variation and irrespective the specifi-

[15]When clustering at the individual agent level, the implications of all our estimates remain the same as shown in the results section.

cation, this sample should result in an unbiased estimate of the ATE of the returns to training. We refer to this estimate as the causal return to training. We contrast this ATE with an estimate of the returns to training that is likely affected by selection into training, and where we expect selection bias $E[Y_{0i}|D=1] - E[Y_{0i}|D=0]$ to be positive.[16]

To get a first assessment of bias, we estimate Equation (2) 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.[17] 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: agent tenure and pre-treatment performance.[18]

We also explore to what degree worker fixed effects contribute to reducing bias. The idea is that the error term u_{it} consists of an unobserved individual-specific component μ_i and an idiosyncratic component ϵ_i, i.e. $u_{it} = \mu_i + \epsilon_{it}$, and that the unobserved individual-specific component μ_i is correlated with the decision to participate in the training program $Cov(\mu_i, d_{it}) \neq 0$. Augmenting Equation (2) with an individual-specific term μ_i and de-

[16]Strictly speaking, compared with Equation (1), we make the additional assumption for our experimental estimate $\widehat{\tau}_E$ that the individual was chosen by management to be included in the experimental sample. All agents in the experimental sample were treated but the timing was randomized. Agents that were not included in the field experiment comprise our non-experimental control group. Thus, similarly, the non-experimental estimate should be interpreted with the added assumption that the non-treated were actively chosen by the management *not* to participate in the field experiment.

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

[18]Blau and Robins (1987) found pre-training wages to strongly reduce conventional OLS estimates (see Table A.1 in the Online Appendix).

meaning then eliminates any time-constant, individual-specific characteristics:

$$\begin{aligned} log(y_{it}) - \overline{log(y_i)} &= \tau(d_{it} - \overline{d}_i) + \gamma(\mu_i - \overline{\mu}_i) + (\epsilon_{it} - \overline{\epsilon}_i) \\ &= \tau(d_{it} - \overline{d}_i) + \epsilon'_i \end{aligned} \tag{3}$$

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. This is addressed by nearest neighbor matching techniques, but due to our small sample size, we only apply a common support restriction which does not explicitly 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). The comparison of experimental treatment and control groups results in the causal estimate of the returns to the workplace training program, whereas the comparison of the experimental treatment group to the non-experimental control group results in the biased, or endogenous, treatment effect. 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.[19,20]

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***	0.1167***	0.1252***
	(0.0179)	(0.0182)	(0.0201)	(0.0217)	(0.0105)	(0.0100)
Adjusted R-squared	0.0315	0.0668	0.0809	0.0730	0.4124	0.0662
Number of agents	74	74	74	74	73	74
Number of observations	1,859	1,859	1,859	1,859	1,850	1,859
B: Non-experimental sample						
Treatment dummy ($\widehat{\tau}_N$)	0.2175***	0.2160***	0.1868***	0.1242***	0.1321***	0.1346***
	(0.0369)	(0.0387)	(0.0233)	(0.0200)	(0.0245)	(0.0116)
Adjusted R-squared	0.0929	0.1089	0.1365	0.1501	0.4692	0.0628
Number of agents	107	107	107	107	104	107
Number of observations	2,383	2,383	2,383	2,383	2,336	2,383
C: Selection bias						
$\widehat{\tau}_N - \widehat{\tau}_E$	0.1083	0.1033	0.0673	0.0407	0.0154	0.0094
S.E. ($\widehat{\tau}_N - \widehat{\tau}_E$)	(0.0407)	(0.0405)	(0.0245)	(0.0291)	(0.0231)	(0.0092)
Bias in % ($(\widehat{\tau}_N - \widehat{\tau}_E)/\widehat{\tau}_E$*100)	99.1	91.6	56.3	48.8	13.2	7.5
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
Pre-treatment performance	No	No	No	No	Yes	No
Worker FE	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.

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

[19]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). For this reason, it is not surprising that the selection bias is much smaller when using G1 agents as the control group.

[20]Standard errors of tests are calculated using the suest command in Stata which accounts for overlapping samples.

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. 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 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). Selection bias is reduced by almost half from 91.6% to 56.3%, as the experimental estimate is 12.0%, whereas the corresponding non-experimental estimate is 18.7%.

When including a linear time trend (Column (4)), pre-treatment performance (Column (5)), or individual fixed effects (Column (6)), the difference between the estimated treatment effects for treatment and control groups gradually decrease towards zero and is not significantly different from zero. The bias is smallest when controlling for pre-treatment performance (Column 5), or individual fixed effects (Column (6)), with 13.2% and 7.5%, respectively. Taken together, the results in Table 2 show that conditioning on pre-treatment performance or individual fixed effects reduces selection bias to the point where it is rela-

tively small and not significantly different from zero.

How can we relate our findings to previous studies estimating the returns to formal workplace training? Due to the different sources of selection bias, it is not possible for us to assess the degree of selection bias in existing non-experimental studies on returns to formal workplace training. It is possible, however, to assess the *change* in estimated returns when including worker fixed effects to account for endogenous training participation. For the studies shown in Table A.1 that provide estimates both with and without worker fixed effects, introducing fixed effects reduces returns to training by 65% (8.1% vs. 2.9%; N=16). In comparison, our results in Table 2 show that, for the non-experimental sample, including worker fixed effects reduces the estimate by up to 38%, from 21.8% (Column (1)) to 13.5% (Column (6)). One explanation for why introducing fixed effects has a smaller effect on the estimated returns to training is that our sample of treated and untreated individuals is more homogenous along several important dimensions. Agents are subject to the same labor market, employed by the same firm, and the treated individuals participate in the same training program.

4.2 Treatment effects across weeks

The weekly frequency of the outcome variable y_{it} allows us to compare treated and untreated agents by week before and after training participation. Figure 1 shows 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.

The solid gray lines in Figure 1 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. Thus, despite using data from the non-experimental control group, including worker fixed effects yields trajectories that closely follow those of the experimental estimates.

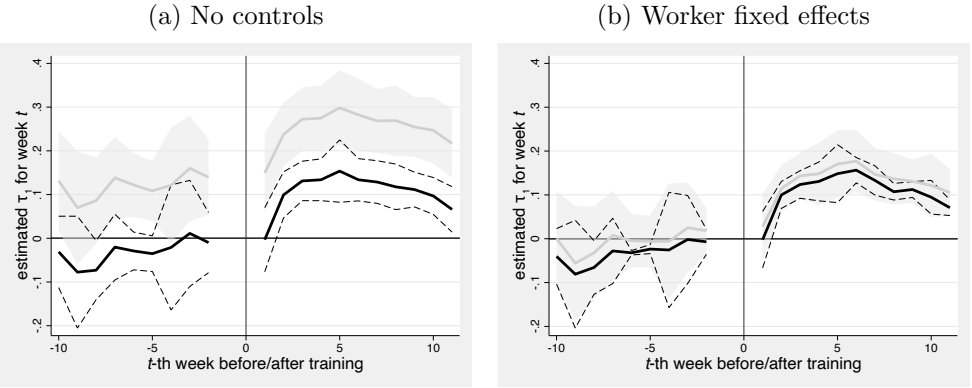
The results for both the experimental and the non-experimental sample in Figure 1 also show that the treatment effect decreases over the course of the post-treatment period. While this could be driven by different mechanisms, De Grip and Sauermann (2012) show this is partially driven by externalities from treated agents to untreated agents such as knowledge spillovers. Panel (b) suggests that this is equally the case for untreated agents in the experimental control group as in the non-experimental control group.[21]

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

[21] Alternatively, the decline could also be driven by behavioral effects, such as motivation (cf. De Grip and Sauermann, 2012).

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



Notes: This figure shows the treatment dummy estimates by weeks before and after the training week for the experimental sample (black line) and the non-experimental sample (gray line). 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. The black dashed line (gray shaded area) indicates the corresponding 95% confidence intervals for the experimental (non-experimental) sample. Period $t-1$ serves as the reference period. There is no performance information in the training week itself.

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 the pre-treatment variables 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 in the Online Appendix. The figure shows that both groups are represented in the propensity score span of about 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***	0.1158***	0.1295***	0.1163***	0.1146***	0.1262***
	(0.0217)	(0.0097)	(0.0070)	(0.0330)	(0.0155)	(0.0051)
Adjusted R-squared	0.0315	0.4405	0.0788	0.0437	0.3294	0.0804
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***	0.1272***	0.1348***	0.1367***	0.1105***	0.1245***
	(0.0299)	(0.0161)	(0.0071)	(0.0342)	(0.0165)	(0.0048)
Adjusted R-squared	0.0888	0.4159	0.0682	0.0591	0.3408	0.0676
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	0.0114	0.0053	0.0204	-0.0041	-0.0016
S.E. ($\widehat{\tau}_N - \widehat{\tau}_E$)	(0.0367)	(0.0188)	(0.0067)	(0.0199)	(0.0140)	(0.0016)
Bias in % ($(\widehat{\tau}_N - \widehat{\tau}_E)/\widehat{\tau}_E$*100)	78.3	9.8	4.1	17.6	-3.5	-1.3
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
Pre-treatment performance	No	No	No	No	Yes	No
Worker FE	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 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 pre-treatment performance (Column 2), or with controls for worker fixed effects (Column 2). The bias in the specification without controls remains large (78.3% in Column 1), whereas bias in the specifications including pre-treatment performance (Column 2) and worker fixed effects (Column 3) are 9.8% and 4.1%, respectively, which is slightly smaller than for the unrestricted sample.

When applying the 5th/95th percentile restriction, the tails of the propensity score distribution remain unbalanced. Columns (4) to (6) of Table 3 therefore also 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 pre-treatment performance or worker fixed effects, the estimated bias is close to zero (-3.5% in Column 5 and -1.3% in Column 6, respectively). These results suggest that, for the training program examined here, applying the common support restriction is linked with a further reduction in selection bias.

5 Robustness checks

5.1 Worker turnover

As shown in Table 1, agents in the non-experimental control group are much more likely to exit the call center. To check whether this explains the selection bias found for the non-experimental sample, Table A.6 in the Online Appendix shows analogous results for the sample of agents who do not exit the department. For the baseline specification without controls, the estimated bias is 72% (Column (1)). For the specifications including pre-treatment performance and including worker fixed effects, the estimated bias is almost the same as without this sample restriction (12% and 7%, respectively). This suggests that the main results are not driven primarily by differential attrition between the experimental

and non-experimental samples.

5.2 Spillover effects

Another source of bias in the estimated returns to training are spillover effects from treated workers to untreated co-workers. These spillovers can arise due to knowledge spillovers or peer pressure and are typically positive (Falk and Ichino, 2006; Mas and Moretti, 2009; De Grip and Sauermann, 2012; Cornelissen et al., 2017). Spillover effects from treatment to control group thereby create an increased productivity in the otherwise untreated control group, creating a downward bias in the estimated treatment effect. It would mean that the estimated effect can be interpreted as a lower bound of the causal effect. In the analysis so far, we have implicitly assumed that spillovers do not exist, i.e. that the stable unit treatment value assumption (SUTVA) is not violated. In the context of this study, it may also be relevant to consider the relative importance of spillover effects for the experimental control group and for the non-experimental control group. If non-experimental control agents benefit more strongly from spillover effects, the selection bias in the non-experimental estimates shown in Table 2 are a lower bound of the true bias. If, however, experimental control group agents benefit more from spillover effects, the downward bias in the experimental estimate implies that we *overstate* the selection bias in the returns to training estimates.

There are two plausible hypotheses in the context regarding the strength of spillover effects of this setting. First, agents from the experimental control group benefit more from spillovers from treated agents. Treatment group agents and agents from the experimental

control group are similar in terms of tenure and are arguably more likely to have formed friendships and co-worker networks.[22] Second, agents from the non-experimental control group benefit from strongly from working in the same team as treated agents. Agents from the non-experimental control group have a relatively steep learning curve and are therefore likely to benefit most from knowledge spillovers (De Grip et al., 2016).

For the field experiment analyzed in this study, De Grip and Sauermann (2012) show that untrained workers in the experimental sample increased their own performance when working with trained peers: a 10 percentage point increase in the share of treated peers improves untreated co-workers' performance by 0.5% (De Grip and Sauermann, 2012).[23] When estimating the same effect for non-experimental control group agents, we find that a 10 percentage point increase in the share of treated peers results in 0.7% larger performance of agents in the non-experimental control group.[24] While these effects are not significantly different from eachother, they suggest that non-experimental control group agents benefit more strongly from working with treated peers and that, if at all, our measure of selection bias *understates* the true selection bias.

6 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

[22] This would be consistent with the finding by Cornelissen et al. (2017) who show stronger peer effects among workers with higher tenure.

[23] In De Grip and Sauermann (2012), peer effects are only estimated for the experimental control group and not the non-experimental control group.

[24] Table available upon request.

or fixed effects approaches. We provide novel evidence that, based on non-experimental data, regression estimates of the impact of workplace training may only be modestly biased. We assess bias in OLS estimators by comparing estimates from a field experiment, using random assignment to training (De Grip and Sauermann, 2012), with estimates obtained using a control group from a non-experimental sample endogenously chosen not to be included in the field experiment. We show that the biased (non-experimental) estimate is up to twice the size of the causal estimate. When including pre-treatment performance or controlling for individual fixed effects, the remaining bias is modest, in several specifications below 10 percent. This suggests that non-experimental regression estimates of the returns to workplace training may be an interesting alternative to experimental estimates.

To put these results into perspective, it is important to note that this paper makes use of data obtained 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. This might, in comparison to data collected from population surveys, explain why we find relatively little remaining bias even without conditioning on common support (cf. Heckman et al., 1998). While these type of single-firm studies are not representative, an increasing number of experimental studies on the returns to training allows us to learn more about the distribution of the returns to training across different settings and labor markets (De Grip and Sauermann, 2012; Adhvaryu et al., 2018; Prada et al., 2019).

While experimental variation in training participation allows to credibly estimate average treatment effects, non-experimental studies have the advantage that they are typically less costly since one does not need to set up a field experiment, conduct lotteries, or worry

about non-compliers. Especially in similar settings to ours, e.g. when managers are interested in evaluating the effectiveness of workplace training programs, our findings suggest that it is possible to elicit modestly biased average treatment effects without random assignment. This knowledge may also be useful for the overall process of providing evidence of the returns to workplace training, and may ultimately lead to an improved understanding of how human capital evolves over the working life.

References

Abadie, A., Athey, S., Imbens, G. and Wooldridge, J. (2017), When should you adjust standard errors for clustering?, NBER Working Paper 24003, National Bureau of Economic Research.

Acemoglu, D. and Autor, D. (2010), Skills, tasks and technologies: Implications for employment and earnings, *in* O. Ashenfelter and D. E.Card, eds, 'Handbook of Labor Economics', Vol. 4 of *Handbook of Labor Economics*, Elsevier, Amsterdam.

Acemoglu, D. and Pischke, J.-S. (1998), 'Why do firms train? theory and evidence', *Quarterly Journal of Economics* **113**(1), 79–119.

Acevedo, P., Cruces, G., Gertler, P. and Martinez, S. (2017), Living up to expectations: How job training made women better off and men worse off, NBER Working Paper 23264, National Bureau of Economic Research.

Adhvaryu, A., Kala, N. and Nyshadham, A. (2018), The skills to pay the bills: Returns to on-the-job soft skills training, NBER Working Paper 24313, National Bureau of Economic Research.

Agodini, R. and Dynarski, M. (2004), 'Are experiments the only option? a look at dropout prevention programs', *Review of Economics and Statistics* **86**(1), 180–194.

Alfonsi, L., Bandiera, O., Bassi, V., Burgess, R., Rasul, I., Sulaiman, M. and Vitali, A. (2020), 'Tackling Youth Unemployment: Evidence from a Labour Market Experiment in Uganda', *Econometrica* **forthcoming**.

Arulampalam, W. and Booth, A. L. (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.

Ashraf, N., Low, C. and McGinn, K. (2020), 'Negotiating a better future: How interpersonal skills facilitate intergenerational investment', *Quarterly Journal of Economics* **135**(2), 1095–1151.

Autor, D. H. and Dorn, D. (2013), 'The growth of low-skill service jobs and the polarization of the us labor market', *American Economic Review* **103**(5), 1553–97.

Barron, J. M., Berger, M. C. and Black, D. A. (1997), 'How well do we measure training?', *Journal of Labor Economics* **15**(3), 507–28.

Bartel, A. P. (1995), 'Training, wage growth, and job performance: Evidence from a company database', *Journal of Labor Economics* **13**(3), 401–25.

Bartel, A. P. (2000), 'Measuring the employer's return on investments in training: Evidence from the literature', *Industrial Relations* **39**(3), 502–524.

Bassanini, A., Booth, A., Brunello, G., De Paola, M. and Leuven, E. (2007), Workplace Training in Europe, *in* G. Brunello, P. Garibaldi and E. Wasmer, eds, 'Education and Training in Europe', Oxford University Press, Oxford, chapter 8-13.

Batt, R., Doellgast, V. and Kwon, H. (2005), Service management and employment systems in u.s. and indian call centers, *in* S. M. Collins and L. Brainard, eds, 'Brookings Trade

Forum 2005: Offshoring White-Collar Work—The Issues and Implications', Brookings Institution Press, Washington, D.C., pp. 335–372.

Beyer, J. D. (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, R. (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.

Blau, D. M. and Robins, P. K. (1987), 'Training programs and wages: A general equilibrium analysis of the effects of program size', *Journal of Human Resources* **22**(1), 113–125.

Blundell, R., Dearden, L. and Meghir, C. (1996), *The determinants and effects of work-related training in Britain*, number R50, IFS Reports, Institute for Fiscal Studies.

Booth, A. L. (1991), 'Job-related formal training: Who receives it and what is it worth?', *Oxford Bulletin of Economics and Statistics* **53**(3), 281–294.

Booth, A. L. (1993), 'Private sector training and graduate earnings', *Review of Economics and Statistics* **75**(1), 164–170.

Booth, A. L. and Bryan, M. L. (2005), 'Testing some predictions of human capital theory: New training evidence from britain', *Review of Economics and Statistics* **87**(2), 391–394.

Booth, A. L., Francesconi, M. and Zoega, G. (2003), 'Unions, work-related training, and wages: Evidence for british men', *Industrial and Labor Relations Review* **57**(1), 68–91.

Breuer, K., Nieken, P. and Sliwka, D. (2013), 'Social ties and subjective performance evaluations: An empirical investigation', *Review of Managerial Science* **7**(2), 141–157.

Brown, J. 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, S. and Pereira, P. T. (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).

Campos, F., Frese, M., Goldstein, M., Iacovone, L., Johnson, H. C., McKenzie, D. and Mensmann, M. (2017), 'Teaching personal initiative beats traditional training in boosting small business in west africa', *Science* **357**(6357), 1287–1290.

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

Cornelissen, T., Dustmann, C. and Schönberg, U. (2017), 'Peer effects in the workplace', *American Economic Review* **107**(2), 425–56.

De Grip, A. and Sauermann, J. (2012), 'The effects of training on own and co-worker productivity: Evidence from a field experiment', *Economic Journal* **122**(560), 376–399.

De Grip, A., Sauermann, J. and Sieben, I. (2016), 'Tenure-performance profiles and the role of peers: Evidence from personnel data', *Journal of Economic Behavior & Organization* **126**, 39–54.

Dearden, L., Reed, H. and Van Reenen, J. (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, J. J. and Handa, S. (2006), 'An assessment of propensity score matching as a non-experimental impact estimator: Evidence from mexico's progresa program', *Journal of Human Resources* **41**(2), 319–345.

Dimitriadis, S. and Koning, R. (2019), The value of communication:evidence from a field experiment with entrepreneurs in togo, Technical report, SSRN Working Paper Series No. 3459643.

Evertsson, M. (2004), 'Formal on-the-job training: A gender-typed experience and wage-related advantage?', *European Sociological Review* **20**(1), 79–94.

Falk, A. and Ichino, A. (2006), 'Clean evidence on peer effects', *Journal of Labor Economics* **24**(1), 39–58.

Fraker, T. and Maynard, R. (1987), 'The adequacy of comparison group designs for evaluations of employment-related programs', *The Journal of Human Resources* **22**(2), 194–227.

Freifeld, L. (2018), *2018 Training Industry Report*, Training Magazine.

Görlitz, K. (2011), 'Continuous training and wages: An empirical analysis using a comparison-group approach', *Economics of Education Review* **30**(4), 691–701.

Goux, D. and Maurin, E. (2000), 'Returns to firm-provided training: evidence from french worker-firm matched data', *Labour Economics* **7**(1), 1–19.

Griffen, A. S. and Todd, P. E. (2017), 'Assessing the performance of nonexperimental estimators for evaluating head start', *Journal of Labor Economics* **35**(S1), S7–S63.

Groh, M., Krishnan, N., McKenzie, D. and Vishwanath, T. (2012), *Soft skills or hard cash? The impact of training and wage subsidy programs on female youth employment in Jordan*, The World Bank.

Haelermans, C. and Borghans, L. (2012), 'Wage effects of on-the-job training: A meta-analysis', *British Journal of Industrial Relations* **50**(3), 502–528.

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica* **47**(1), 153–161.

Heckman, J. J., Ichimura, H., Smith, J. A. and Todd, P. (1998), 'Characterizing selection bias using experimental data', *Econometrica* **66**(5), 1017–98.

Hill, E. T. (1995), 'Labor market effects of women's post-school-age training.', *Industrial and Labor Relations Review* **49**(1), 138–149.

Imbens, G. W. and Wooldridge, J. M. (2009), 'Recent developments in the econometrics of program evaluation', *Journal of Economic Literature* **47**(1), 5–86.

Kawaguchi, D. (2006), 'The incidence and effect of job training among japanese women', *Industrial Relations: A Journal of Economy and Society* **45**(3), 469–477.

Konings, J. and Vanormelingen, S. (2015), 'The impact of training on productivity and wages: Firm-level evidence', *Review of Economics and Statistics* **97**(2), 485–497.

Krueger, A. and Rouse, C. (1998), 'The effect of workplace education on earnings, turnover, and job performance', *Journal of Labor Economics* **16**(1), 61–94.

LaLonde, R. J. (1986), 'Evaluating the econometric evaluations of training programs with experimental data', *American Economic Review* **76**(4), 604–620.

Leuven, E. and Oosterbeek, H. (2004), 'Evaluating the effects of a tax deduction on training', *Journal of Labor Economics* **22**(2), 461–488.

Leuven, E. and Oosterbeek, H. (2008), 'An alternative approach to estimate the wage returns to private-sector training', *Journal of Applied Econometrics* **23**(4), 423–434.

Lillard, L. and Tan, H. (1992), 'Private sector training: Who get's it and what are its effects?', *Research in Labor Economics* **13**(1).

Liu, X. and Batt, R. (2007), 'The economic pay-offs to informal training: Evidence from routine service work', *Industrial and Labor Relations Review* **61**(1), 75–89.

Loewenstein, M. A. and Spletzer, J. R. (1998), 'Dividing the costs and returns to general training', *Journal of Labor Economics* **16**(1), 142–171.

Lynch, L. M. (1992), 'Private-sector training and the earnings of young workers', *American Economic Review* **82**(1), 299–312.

Marcotte, D. E. (2000), 'Continuing education, job training, and the growth of earnings inequality', *Industrial and Labor Relations Review* **53**(4), 602–623.

Mas, A. and Moretti, E. (2009), 'Peers at work', *American Economic Review* **99**(1), 112–145.

McKenzie, D., Stillman, S. and Gibson, J. (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.

Murthy, N. N., Challagalla, G. N., Vincent, L. H. and Shervani, T. A. (2008), 'The impact of simulation training on call center agent performance: A field-based investigation', *Management Science* **54**(2), 384–399.

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

Parent, D. (1999), 'Wages and mobility: The impact of employer-provided training', *Journal of Labor Economics* **17**(2), 298–317.

Parent, D. (2003), 'Employer-supported training in canada and its impact on mobility and wages', *Empirical Economics* **28**(3), 431–459.

Pischke, J.-S. (2001), 'Continuous training in germany', *Journal of Population Economics* **14**(3), 523–548.

Pischke, J.-S. (2007), Comment on 'workplace training in europe', by andrea bassanini et al., *in* G. Brunello, P. Garibaldi and E. Wasmer, eds, 'Education and Training in Europe', Oxford University Press, Oxford, pp. 330–342.

Prada, M., Rucci, G. and Urzua, S. (2019), Training, Soft Skills and Productivity: Evidence from a Field Experiment, IZA Discussion Paper 12447.

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

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

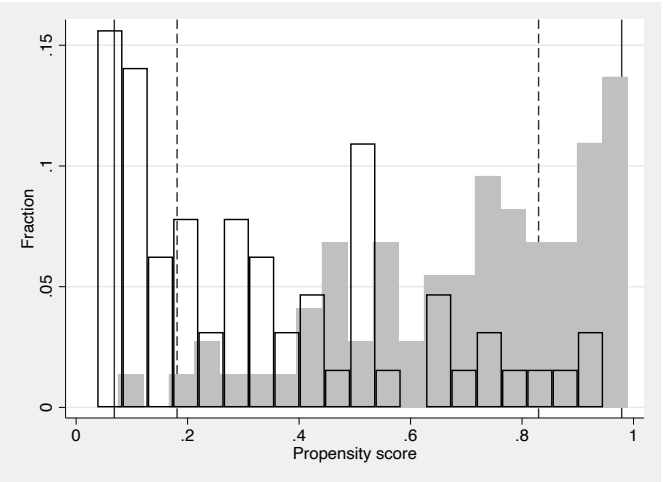
Sieben, I., De Grip, A., Van Jaarsveld, D. and Sørensen, O. (2009), 'Technology, selection, and training in call centers', *Industrial and Labor Relations Review* **62**(4), 553–572.

Vignoles, A., Galindo-Rueda, F. and Feinstein, L. (2004), 'The labour market impact of adult education and training: A cohort analysis', *Scottish Journal of Political Economy* **51**(2), 266.

Wilde, E. T. and Hollister, R. (2007), 'How close is close enough? evaluating propensity score matching using data from a class size reduction experiment', *Journal of Policy Analysis and Management* **26**(3), 455–477.

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).
Görlitz (2011)	Wages	1.9% 0.5% (n.s.)	"Endog." comparison "Exog." comparison	German survey data among individuals employed in December 2006. Approach cf. to Leuven and Oosterbeek (2008) .
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.
Murthy et al. (2008)	Worker productivity	8.6-20.9%	OLS (field experiment)	Call agents in two US firms (simulation vs. role-play training)
O'Connell and Byrne (2010)	Wages	3.6%	OLS	Irish Survey of Employees' 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 9). 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) Non-exp. control group G1	(2) G2	(3) Difference G1-G2	(4) Difference G1-Exp. TG	(5) Difference 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; Column (3) shows differences between G1 and G2, Column (4) differences between G1 and the experimental treatment group, and Column (5) differences between G2 and the experimental treatment group, respectively. Parentheses and asterisks in Columns (3) to (5) are from a two-sided t-test on the respective differences (*** p<0.01, ** p<0.05, * p<0.1).

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**	0.1386***	0.1346***
	(0.0200)	(0.0207)	(0.0216)	(0.0427)	(0.0174)	(0.0124)
Adjusted R-squared	0.1067	0.1270	0.1357	0.1276	0.4897	0.1780
*B: TG-G2 (*N=100, n=2,190*)*						
Treatment dummy	0.2205***	0.2179***	0.1769***	0.1261***	0.1143***	0.1252***
	(0.0402)	(0.0417)	(0.0193)	(0.0211)	(0.0138)	(0.0098)
Adjusted R-squared	0.0832	0.0996	0.1305	0.1421	0.4760	0.0472
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
Pre-treatment performance	No	No	No	No	Yes	No
Worker FE	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***	0.1041***	0.1267***
	(0.0177)	(0.0138)	(0.0095)
Adjusted R-squared	0.0276	0.4240	0.0732
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***	0.1169***	0.1357***
	(0.0272)	(0.0211)	(0.0114)
Adjusted R-squared	0.0865	0.3804	0.0756
Number of agents	64	62	64
Number of observations	1,722	1,684	1,722
C: Selection bias			
$\widehat{\tau}_N - \widehat{\tau}_E$	0.0715	0.0129	0.0090
S.E. ($\widehat{\tau}_N - \widehat{\tau}_E$)	(0.0263)	(0.5370)	(0.3527)
Bias in % ($(\widehat{\tau}_N - \widehat{\tau}_E)/\widehat{\tau}_E$*100)	72.3	12.4	7.1
Control variables	No	No	No
Tenure (linear + squared)	No	No	No
Common trend	No	No	No
Pre-treatment performance	No	Yes	No
Worker FE	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.