Journal of the
Royal Statistical Society

SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2012)
175, *Part* 2, *pp.* 587–617

Evaluating continuous training programmes by using the generalized propensity score

Jochen Kluve,

Humboldt University, Berlin, Rheinisch-Westfälisches Institut für Wirtschaftsforschung, Essen, and Institute for the Study of Labor, Bonn, Germany

Hilmar Schneider,

Institute for the Study of Labor, Bonn, Germany

Arne Uhlendorff

University of Mannheim, Institute for the Study of Labor, Bonn, and Deutsches Institut für Wirtschaftsforschung, Berlin, Germany

and Zhong Zhao

Renmin University of China and Institute for the Study of Labor, Bonn, Germany

[Received March 2010. Final revision June 2011]

Summary. The paper assesses the heterogeneity of treatment effects arising from variation in the duration of training. We use German administrative data that have the extraordinary feature that the amount of treatment varies continuously from 10 days to 395 days (i.e. 13 months). This feature allows us to estimate a continuous dose–response function that relates each value of the dose, i.e. days of training, to the individual post-treatment probability of employment (the response). The dose–response function is estimated after adjusting for covariate imbalance by using the generalized propensity score, which is a recently developed method for covariate adjustment under continuous treatment regimes. Our data have the advantage that we can consider both the actual and the planned durations of training as treatment variables: if only actual durations are observed, treatment effect estimates may be biased because of endogenous exits. Our results indicate an increasing dose–response function for treatments of up to 120 days, which then flattens out, i.e. longer training programmes do not seem to add an additional treatment effect.

Keywords: Continuous treatment; Generalized propensity score; Programme evaluation; Training

1. Introduction

Over recent years there has been an increasing amount of research on the effectiveness of labour market training programmes in many countries. Training programmes represent the ‘classic’ type of so-called active labour market programmes, owing to their objective of enhancing participants’ employment prospects by increasing their human capital. Whereas the evidence on early training programmes in the 1970s and 1980s showed relatively optimistic results, the more recent research from the 1990s and 2000s—generally based on much better data and

Address for correspondence: Jochen Kluve, School of Business and Economics, Humboldt Universität zu Berlin, Spandauer Strasse 1, 10178 Berlin, Germany.
E-mail: jochen.kluve@hu-berlin.de

0964–1998/12/175587

advanced econometric methods—points to the result that training programmes seem to be modestly effective at best (Heckman *et al.*, 1999; Kluve, 2010). Adding to this general finding, one recent line of research shows that positive treatment effects may only materialize in the long run, and that programme effectiveness can show a considerable dynamic ranging from often severe short-term locking-in effects to long-term gains in employment prospects (e.g. Lechner *et al.* (2011)).

In the previous training literature, the focus is typically on binary treatment effects and there is no differentiation of training programmes by their duration. The provision of training, however, can be quite heterogeneous both in terms of content of training and in terms of duration of training. In this paper we contribute to the literature on training programmes by focusing on the heterogeneity of treatment effects that may arise from variation in the duration of treatment. We implement this analysis on the basis of data on training programmes in Germany. The key feature of the data is the fact that the duration of treatment varies almost continuously from approximately 1 week up to approximately 13 months. We focus on programmes in which no specific vocational degree is acquired as part of the requirements of the programme—this is the majority of training programmes in Germany. Participants in these programmes learn specific skills that are required for a certain vocation, like computer-aided design for technicians, or receive qualifications that are of general vocational use, like general computer skills. In this paper we compare the effect of being trained within the same type of programme, but with different durations, on the subsequent probability of employment.

The evaluation question that corresponds to the continuous administering of training is how effective (relative to each other) are training programmes with different durations? This assessment of the heterogeneity of treatment effects along the duration dimension essentially amounts to estimating a dose–response function as proposed in Hirano and Imbens (2004). In this paper we therefore estimate the response—i.e. the probability of employment—that corresponds to specific values of continuous doses—i.e. training of a particular length.

In a setting in which doses are not administered under experimental conditions, estimation of a dose–response function is possible by using the generalized propensity score (GPS). The GPS for continuous treatments is a straightforward extension of the well-established and widely used propensity score methodology for binary treatments (Rosenbaum and Rubin, 1983) and multivalued treatments (Imbens, 2000; Lechner, 2001). The GPS methodology is developed in Hirano and Imbens (2004) and Imai and van Dyk (2004). Similarly to the binary and multivalued treatment propensity score methods it is assumed that—conditional on observable characteristics—the level of treatment received can be considered as random. Hirano and Imbens (2004) showed that the GPS has a balancing property that is similar to the balancing property of the ‘classic’ propensity score. This implies that individuals within the same strata of the GPS should be identical in terms of their observable characteristics, independent of their level of treatment. Compared with propensity score methods for multivalued treatments, the GPS has the advantage that we do not have to discretize the continuously distributed duration of training and thus can make use of more comprehensive information. To our knowledge, our paper along with parallel work by Flores *et al.* (2011) constitute the first applications of the GPS in the context of evaluating active labour market programmes.

In implementing the GPS approach, our data have the advantage that we can consider both the actual and the planned durations of training as treatment variables: if only actual durations are observed, treatment effect estimates may be biased because of endogenous exits. This could be the case, for instance, if observed durations are shorter than the initially planned durations, because people exit from the programme early if they find a job. The bias could also point the other way, if a substantial fraction of programme participants drop out early. We investigate these issues by taking into account both the actual and the planned durations of individual programme participants.

The paper is organized as follows. Section 2 gives details on the data and the treatment that we study. Section 3 describes the methodology of estimating a dose–response function to evaluate a continuous policy measure, adjusting for the GPS. The fourth section contains the empirical implementation. It discusses the plausibility of the unconfoundedness assumption, it details the GPS estimation, the common support condition and the balancing of covariates, and it presents the results from estimating the dose–response function. Section 5 contains several robustness checks. Section 6 concludes.

The programs that were used to analyse the data can be obtained from

```
http://www.blackwellpublishing.com/rss
```

2. Institutional setting and data

2.1. Public training programmes in Germany

The most important German Government labour market policy that is relevant to our paper is social code III (*Sozialgesetzbuch III*) that was enacted in 1998. The focus group for social code III consists of people who are unemployed or under threat of unemployment. The code has emphasized the use of an active labour market policy and aims to reduce unemployment. The Federal Employment Agency through its 10 regional directorates and 180 local employment agencies (with around 660 branch offices) is responsible for implementation of the federal labour market policy at national, regional and local level. See Wunsch (2005) for a detailed description of German labour market policy and related institutions.

Training programmes are one of the most important components of an active labour market policy in Germany with an annual budget of around €7 billion (2002 figures; see Eichhorst and Zimmermann (2007)). Access to training programmes is not a legal entitlement but is based on the decision of the caseworker. If a caseworker has decided that her client needs to go through a training programme, the caseworker also specifies the type, the content and the duration of the training and refers the client to a designated training provider. During the process, the factors that the caseworker takes into consideration include the aptitude of her client for a certain job, the likelihood of succeeding in a specific training programme, the local labour market condition, the cost of training and to some extent the available training slots in the contracted training institutions. It is thus reasonable to assume that once we condition on the large set of observable characteristics, including previous labour market outcomes, the decision about the length of the programme is independent of the future labour market outcomes of the participants.

Among the programmes that are considered here we can distinguish between classroom-oriented training (type 1) and more practically oriented programmes with only a few theoretical parts (type 2). However, the duration as well as the effectiveness of both types are very similar and therefore we pool both programmes to increase our sample size. Participants in the programmes that we consider learn specific skills required for a certain vocation (e.g. computer-aided design for a technician or tracer, '*berufsbezogene Weiterbildung*') or receive qualifications that are of general vocational use (e.g. Microsoft Office and computer skills, '*berufsübergreifende Weiterbildung*'). Numerically, these types constitute the most important among all publicly financed training programmes: in 2000, roughly 70% of all participants in training were assigned to these two types (Schneider and Uhlendorff, 2006; Institute for the Study of Labor *et al.*, 2007). Whereas the programmes that we focus on do not lead to a vocational degree, participants may receive a certificate about the type and the content of the training. Programmes leading to the acquisition of a degree are not considered, since the degree requirement generates discontinuities in the distribution of treatment durations, and the objective of the analysis in this paper is to estimate the employment outcomes that are associated with each level of a continuous treatment. Programmes leading to a vocational degree have a duration of around 2 years—see for example Lechner *et al.* (2011) for an analysis of these long-term programmes.

2.2. Data

We use a sample of a particularly rich administrative data set, the integrated employment biographies of the German Federal Employment Agency (Bundesagentur für Arbeit). The data contain detailed daily information on employment subject to social security contributions, including occupational and sectoral information, receipt of transfer payments during periods of unemployment, job search activity and participation in different programmes of an active labour market policy. Furthermore, the integrated employment biographies comprise a large variety of covariates like age, education, disability, nationality and regional indicators.

Our sample of participants consists of about 265 male unemployed people per quarter entering the programme during the years 2000, 2001 and 2002, i.e. we observe approximately 3180 programme participants. The system of publicly financed training in Germany underwent several alterations during labour policy reforms in 2003 (see Jacobi and Kluve (2007)). We therefore restrict our analysis to pre-reform training programmes to avoid possible distortions in measuring programme effectiveness. The data were generated as part of a research project for the German government evaluating these labour policy reforms. To capture trends over time and to separate this from the reform effect in 2003, the original sample consists of entry cohorts over time with stable cohort size per quarter.

The data allow us to draw conclusions on the average participant starting a programme during this time period. The core feature of the training programmes that we analyse is the fact that the provision of treatment is a continuous variable. For all participants we know the initial length of the treatment that they were assigned to (i.e. the planned duration), as well as how long they actually stayed in the treatment (i.e. the actual duration).

One *caveat* is that we do not have detailed information on the content of the training. Although we restrict our analysis to two relatively homogeneous types of training programmes, we cannot rule out that the content of the training could vary with the duration of training. Our estimates may reflect the composite effect of duration of training and content of training. Nonetheless, our paper still has strong policy implications and reflects the reality that content of training and duration of training are usually offered together as a bundle to the programme applicants.

We discard observations with duration of treatment below 10 days, since such short durations arguably do not imply a serious attempt at finishing the programme. Durations above 395 days are also discarded, since only very few observations are available. We do not consider durations of length 0, i.e. no non-treated individuals are included (though we do estimate the baseline counterfactual outcome of those non-treated separately to facilitate interpreting our main results). Instead, we focus on the average responses of those individuals who did receive some treatment. Fig. 1 shows the distribution of duration of treatment, for both the actual and the planned durations. Both distributions cover the full range of durations of training, and for both distributions two peaks exist at durations of 180 days and 360 days. Fig. 1 also shows that actual durations tend to be slightly shorter than planned durations.

The responses, i.e. the outcome variables of interest, are

(a) the probability of employment at time 1 year after exit from the programme and
(b) the probability of employment at time 2 years after entry into the programme.

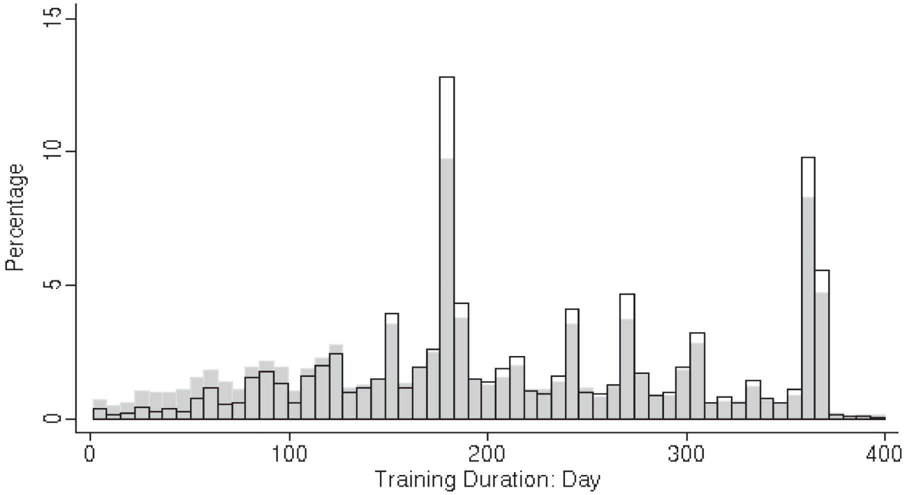

Fig. 1. Distributions of actual (■) and planned (□) durations of training

Table 1 presents summary statistics of the two outcome variables and the covariates, for the full sample (columns (1) and (2)) as well as for three subsamples, 'early exits' (i.e. actual duration shorter than planned duration: columns (3) and (4)), 'late exits' (i.e. actual duration longer than planned duration: columns (5) and (6)) and 'exits as planned' (i.e. actual duration equal to planned duration: columns (7) and (8)). The share of individuals who stayed in the programme exactly as long as planned is quite high (68.7%). In the case in which actual and planned durations differ, early exits are much more common than late exits (22.1% and 9.2% of observations respectively), which is a pattern that has already been observed in Fig. 1.

As Table 1 shows, the data contain a large number of covariates. In particular, we can use information on numerous variables that have been identified in the programme evaluation literature to be important determinants of selection into a programme: this comprises detailed data on citizenship and educational background, including vocational education. Moreover, we have detailed information on pretreatment employment histories, covering the number of days spent in employment and unemployment during the 4 years preceding treatment and the previous employment states measured at 12 points in time $(4, 8, \ldots, 48$ weeks before entering the programme) as well as regional indicators. When using non-experimental estimators, previous studies, e.g. Heckman *et al.* (1998), Diaz and Handa (2006) and Mueser *et al.* (2007), have emphasized the importance of having treated and comparison groups from the same local labour market, applying the same survey instrument for both groups and having rich information on individuals' recent labour market history. First, given that our data come from one administrative source, the requirement of a homogeneous survey instrument is certainly satisfied. Second, the richness of the covariates makes the weak unconfoundedness assumption plausible (see also below). Third, our sample is a random sample of participants from West Germany, such that, to control for local labour market conditions, we condition on the local rate of unemployment and the regional type, reflecting the general performance of the regional labour market (for a description of the various regional types see Blien *et al.* (2004)). This and the fact that we analyse a national programme ensure that different effects of different lengths of programme are not driven by different regional administrations or different local labour market conditions.

Table 1. Summary statistics†

Statistic	*Full sample*		*Early exits*		*Late exits*		*Exits as planned*	
	Mean, (1)	*Standard deviation, (2)*	*Mean, (3)*	*Standard deviation, (4)*	*Mean, (5)*	*Standard deviation, (6)*	*Mean, (7)*	*Standard deviation, (8)*
Age	37.22	10.36	36.30	10.54	37.00	10.40	37.55	10.27
Disability								
Disability low degree	0.07	—	0.09	—	0.04	—	0.07	—
Disability medium degree	0.01	—	0.00	—	0.00	—	0.01	—
Disability high degree	0.01	—	0.00	—	0.00	—	0.01	—
Citizenship								
Foreigner, European Union	0.02	—	0.02	—	0.01	—	0.02	—
Foreigner, non-European-Union	0.10	—	0.11	—	0.14	—	0.10	—
Educational attainment								
No graduation	0.12	—	0.14	—	0.09	—	0.12	—
1st stage of secondary level	0.48	—	0.53	—	0.48	—	0.47	—
2nd stage of secondary level	0.26	—	0.23	—	0.29	—	0.26	—
Advanced technical college entrance qualification	0.04	—	0.03	—	0.05	—	0.04	—
General qualification for university entrance	0.10	—	0.06	—	0.09	—	0.11	—
Vocational attainment								
No vocational degree	0.34	—	0.43	—	0.32	—	0.32	—
In-plant training	0.53	—	0.48	—	0.56	—	0.55	—
Off-the-job training, vocational school, technical school	0.06	—	0.05	—	0.05	—	0.06	—
University, advanced technical college	0.07	—	0.04	—	0.07	—	0.07	—
Employment history								
Previous unemployment duration (months)	9.38	7.66	9.14	7.55	8.51	7.39	9.57	7.72
Duration of last employment (months)	20.74	30.26	17.52	27.22	21.71	32.52	21.65	30.82
Log(wage) of last employment	3.61	1.17	3.59	1.12	3.47	1.32	3.63	1.16
No last employment observed	0.08	—	0.08	—	0.11	—	0.08	—
Share of days in employment, 1st year before programme	0.19	—	0.19	—	0.21	—	0.18	—
Share of days in employment, 2nd year before programme	0.38	—	0.36	—	0.40	—	0.38	—
Share of days in employment, 3rd year before programme	0.43	—	0.41	—	0.41	—	0.43	—
Share of days in employment, 4th year before programme	0.45	—	0.42	—	0.44	—	0.46	—

(*continued*)

Table 1 (*continued*)

Statistic	*Full sample*		*Early exits*		*Late exits*		*Exits as planned*	
	Mean, (1)	*Standard deviation, (2)*	*Mean, (3)*	*Standard deviation, (4)*	*Mean, (5)*	*Standard deviation, (6)*	*Mean, (7)*	*Standard deviation, (8)*
Employment history								
Share of days in unemployment, 1st year before programme	0.67	—	0.68	—	0.64	—	0.67	—
Share of days in unemployment, 2nd year before programme	0.39	—	0.43	—	0.36	—	0.39	—
Share of days in unemployment, 3rd year before programme	0.34	—	0.37	—	0.33	—	0.33	—
Share of days in unemployment, 4th year before programme	0.30	—	0.33	—	0.27	—	0.29	—
Employment 4 weeks before programme entry	0.07	—	0.07	—	0.10	—	0.07	—
Employment 8 weeks before programme entry	0.14	—	0.16	—	0.17	—	0.14	—
Employment 12 weeks before programme entry	0.20	—	0.21	—	0.21	—	0.19	—
⋮								
Employment 48 weeks before programme entry	0.42	—	0.39	—	0.45	—	0.42	—
Outcome variables								
Employment 2 years after programme entry	0.35	—	0.35	—	0.33	—	0.38	—
Employment 1 year after programme exit	0.34	—	0.35	—	0.34	—	0.33	—
Number of observations	3162		700		291		2171	

†Source: integrated employment biographies of the German Federal Employment Agency. The sample consists of male participants in training programmes in West Germany for the years 2000–2002. The subsample ‘early exits’ contains individuals with actual training duration shorter than the planned duration; the subsample ‘late exits’ refers to actual duration longer than the planned duration and the subsample ‘exits as planned’ contains individuals for whom the planned duration equals the actual duration. All time varying characteristics are measured at the beginning of the training.

Table 1 also shows that the covariate distributions are very similar across all (sub)samples, which indicates that selection into the different subsamples on the basis of observable characteristics is not strong. In the remainder of this paper we shall therefore focus on presenting results for the actual durations of the full sample, complementing these results with planned durations and the sample of actual equal to planned durations as appropriate. Looking at the full sample, the participants are on average 37 years old, around 9% of them are handicapped and 12% do not have German citizenship. The participants are on average relatively low skilled: more than 60% did not progress further than the first stage of secondary level education, around 35% do not have any vocational degree and only a minority (7%) have obtained a university degree.

Before entering training the participants were on average unemployed for 9 months, and their previous employment lasted for about 21 months. The individuals for whom we observe a wage for their last employment earned around €50 per day. For the previous employment history we construct eight variables describing the share of time spent in employment and unemployment during each of the four years before entering the programme. There is a clearly increasing trend in the average probability of being unemployed over time as the individuals move closer to enrolment in the programme. This is also reflected in the decreasing share of individuals who are employed measured at 12 points in time during the year before the programme starts.

Looking at the two outcome variables, both 2 years after entry to the programme and 1 year after the programme ended around 35% of the participants are employed. Fig. 2 contains two panels plotting unadjusted outcomes—i.e. the probability of employment 2 years after entry to the programme as well as the probability of employment 1 year after exit from the programme—against the actual duration of training. Figs 2(a) and 2(b) generally show an increasing trend: after an initial dip in the probability of employment during the first weeks in the programme, employment rates seem to increase with the length of participation.

3. Removal of bias by using the generalized propensity score

Research in programme evaluation in recent years has made comprehensive use of matching methods (see *inter alia* the overview in Augurzky and Kluve (2007) and a symposium on the econometrics of matching in *The Review of Economics and Statistics*, volume 86 (2004), part 1, pages 1–194, in particular the article by Imbens (2004)). In the absence of experimental data, which is largely the case, the popularity of matching is due to its intuitively appealing technique of mimicking an experiment *ex post*. The standard case, which is also appropriate for the majority of applications, considers a binary treatment. One of the key results that have made matching such an attractive empirical tool was developed in Rosenbaum and Rubin (1983), who showed that, rather than conditioning on the full set of covariates, conditioning on the propensity score—i.e. the probability of receiving the treatment given the covariates—is sufficient to balance treatment and comparison groups.

Subsequently, the literature has extended propensity score methods to the cases of multivalued treatments (Imbens, 2000; Lechner, 2001) and, more recently, continuous treatments (Imbens, 2000; Hirano and Imbens, 2004; Imai and van Dyk, 2004). In this paper, we build on the approach that was developed by Hirano and Imbens (2004) who proposed estimating the entire dose–response function of a continuous treatment. This approach fits perfectly with the objective of our paper, since we are interested in the response—i.e. the post-treatment probability of employment—that is associated with each value of the continuous dose, i.e. the days spent in training. Alternatively we could discretize the continuously distributed treatment variable

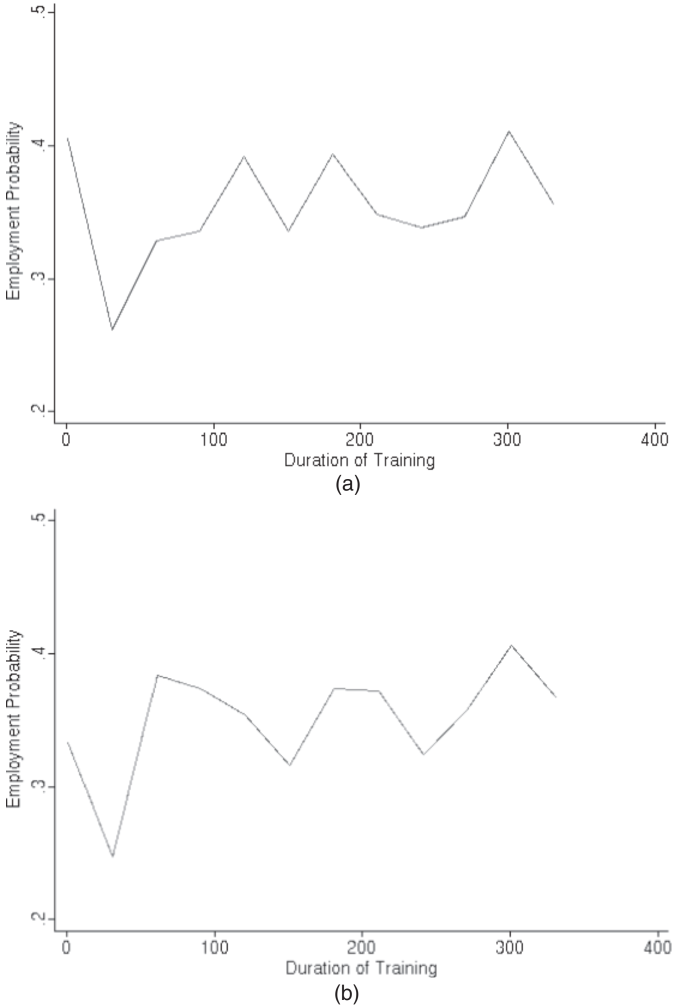

Fig. 2. Unadjusted probability of employment at time 2 years after entry into the programme: (a) actual duration of training; (b) planned duration of training

and apply propensity score methods for multivalued treatments. The GPS, though, has the advantage that it makes use of the entire information contained in the distribution of duration of treatment.

3.1. Generalized propensity score methodology

Hirano and Imbens (2004) developed the GPS methodology in the context of the potential outcomes model for estimation of causal effects of treatments. In what follows we closely follow their presentation. We have a random sample of training participants, indexed by $i=1,\ldots,N$. For each unit i there is a set of potential outcomes $Y_i(t)$—the employment state—for $t \in \Im$, which is referred to as the unit level dose–response function. In the continuous case, $\Im$ is an interval $[t_0, t_1]$, whereas in the binary case it would be $\Im=\{0,1\}$. Our objective is to estimate the average dose–response function $\mu(t)=E[Y_i(t)]$. For each participant i, we observe a vector of covariates X_i, the duration T_i of the training that individual i actually receives, with $T_i \in [t_0, t_1]$, and the potential outcome corresponding to the level of treatment received, $Y_i = Y_i(T_i)$. In the remainder of this section the subscript i will be omitted to simplify the notation.

The key assumption of Hirano and Imbens (2004) generalizes the *unconfoundedness* assumption for binary treatments that was made by Rosenbaum and Rubin (1983) to the continuous case:

$$Y(t) \perp T \mid X \qquad \text{for all } T \in \Im. \tag{1}$$

Hirano and Imbens (2004) referred to this as *weak unconfoundedness*, since it only requires conditional independence to hold for each value of the treatment, rather than joint independence of all potential outcomes. Calling $r(t,x) = f_{T|X}(t|x)$ the conditional density of the treatment given the covariates, the GPS is defined as

$$R = r(T, X). \tag{2}$$

The GPS has a balancing property that is similar to the balancing property of the propensity score for binary treatments. Within strata with the same value of $r(t,X)$ the probability that $T=t$ does not depend on the value of X, i.e. the GPS has the property that $X \perp \mathbf{1}\{T=t\} \mid r(t,X)$. Hirano and Imbens (2004) emphasized that this is a mechanical implication of the definition of the GPS and does not require weak unconfoundedness. In combination with weak unconfoundedness, however, it implies that assignment to treatment is unconfounded given the GPS, i.e. Hirano and Imbens (2004) proved that, if assignment to treatment is weakly unconfounded given covariates X, then it is also weakly unconfounded given the GPS.

Given this result, it is possible to use the GPS to remove bias that is associated with differences in covariates in three steps. The first step consists of modelling and estimating the GPS. The second step is to estimate the conditional expectation of the outcome as a function of two scalar variables: the treatment level T and the GPS R, i.e.

$$\beta(t,r) = E[Y \mid T=t, R=r]. \tag{3}$$

For the estimation of equation (3) we must assume some functional form of the relationship between the employment state Y, the training duration T and the GPS R. The third step is to estimate the dose–response function at each particular level of the treatment. This is implemented by averaging the conditional expectation function over the GPS at that particular level of the treatment:

$$\mu(t) = E[\beta\{t, r(t,X)\}]. \tag{4}$$

The procedure does not average over the GPS $R = r(T,X)$, but instead it averages over the score evaluated at the treatment level of interest $r(t,X)$. Hirano and Imbens (2004) also emphasized that the regression function $\beta(t,r)$ does not have a causal interpretation, but that $\mu(t)$ corresponds to the value of the dose–response function for treatment value t, which compared with another treatment level t' does have a causal interpretation.

3.2. Implementation

In the practical implementation of the methodology that was outlined in the previous section, we use a normal distribution for the duration of training given the covariates

$$T_i \mid X_i \sim N(\beta_0 + \beta_1' X_i, \sigma^2), \tag{5}$$

which we estimate by ordinary least squares regression. It is possible to assume other distributions than the normal distribution, and to estimate the GPS by other methods such as maximum likelihood, or to use non-parametric methods for partial means. The key point here, however, is to make sure that the covariates are balanced after adjusting for the GPS: as long as sufficient covariate balance is achieved, the exact procedure for estimating the GPS is of secondary importance. We alternatively estimated the GPS on the basis of the logarithm of the duration as a check of robustness, but using duration instead of the logarithm of the duration turned out to be superior in finding GPS specifications that balance the covariates in our sample. The GPS estimated is calculated as

$$\hat{R}_i = \frac{1}{\surd(2\pi\hat{\sigma}^2)} \exp\left\{-\frac{1}{2\hat{\sigma}^2}(T_i - \hat{\beta}_0 - \hat{\beta}_1' X_i)^2\right\}. \tag{6}$$

In the second stage we calculate the conditional expectation function of the probability of employment Y_i given T_i and R_i as a flexible function of its two arguments. Our empirical approach uses the following polynomial approximation:

$$E[Y_i|T_i, R_i] = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \alpha_3 T_i^3 + \alpha_4 R_i + \alpha_5 R_i^2 + \alpha_6 R_i^3 + \alpha_7 T_i R_i + \alpha_8 T_i^2 R_i + \alpha_9 T_i R_i^2. \tag{7}$$

In addition to the specification in equation (7) we also implement several other specifications to allow for sufficiently flexible functional forms. On the one hand we vary the degree of the polynomial specification to test whether our results are robust with respect to a higher degree of flexibility. On the other hand we apply flexible spline models to avoid a parametric specification. Spline models are a non-parametric regression technique. They are not only flexible, but they also have good properties in terms of the mean-squared error fit. Another attractiveness is that they are relatively easy to implement; see Keele (2008). Moffitt (2009) applied a spline model in the context of estimating marginal treatment effects.

For each individual the observed duration of training T_i and estimated GPS $\hat{R}_i$ are used, and the equation is estimated by ordinary least squares. Given the estimated parameters in the second stage, we estimate the average potential outcome at treatment level t as

$$\begin{aligned} E[Y(t)] = \frac{1}{N}\sum_{i=1}^{N}\{&\hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2 + \hat{\alpha}_2 t^3 + \hat{\alpha}_4\, \hat{r}(t, X_i) + \hat{\alpha}_5\, \hat{r}(t, X_i)^2 \\ &+ \hat{\alpha}_6\, \hat{r}(t, X_i)^3 + \hat{\alpha}_7 t\, \hat{r}(t, X_i) + \hat{\alpha}_8 t^2\, \hat{r}(t, X_i) + \hat{\alpha}_9\, t\, \hat{r}(t, X_i)^2\}. \end{aligned} \tag{8}$$

The entire dose–response function can then be obtained by estimating this average potential outcome for each level of the treatment. In our application, we use bootstrap methods to obtain standard errors that take into account estimation of the GPS and the α-parameters, i.e. we bootstrap the entire estimation process.

3.3. Assessing the weak unconfoundedness assumption, common support condition and balancing of covariates

The key assumption for the GPS is the weak unconfoundedness assumption, which is also known as the assumption of selection on observables. As an identifying assumption, it is not statistically testable. One potential case of violating this assumption is the possibility of reverse causality. Individuals might leave the training programme because they have received an acceptable job offer. In this case, finding a job leads to a shorter duration of training. The planned duration is determined before the programme, which is free from endogenous exit and plausibly exogenous once we condition on detailed observed characteristics including previous labour market history. We can use the information on the planned duration as an instrumental variable to test the potential endogeneity due to reverse causality of the actual duration of treatment. If we assume that the planned duration is exogenous, this allows us to take potential endogeneity into account which may occur because of reverse causality or because some individuals leave the programme because they expect no further benefits from continuing the programme. This approach does not allow us to test whether there are unobservable characteristics (unobserved to researchers, but observable to the programme applicants and caseworkers) like motivation or self-esteem which may have an effect on both, the planned and the observed duration of the training programme, and which are relevant for subsequent labour market outcomes. However, the detailed information on previous labour market outcomes that is contained in our data makes it very likely that unobservable characteristics such as motivation or self-esteem are captured by the individual labour market history.

Similarly to standard propensity score matching methods, common support is also a concern in the GPS application. We propose to test the common support condition as follows (Flores *et al.* (2011) use a similar approach): first, we divide the sample into three groups according to the distribution of length of treatment, cutting at the 30th and 70th percentile of the distribution. Then we evaluate the GPS at the group median of the treatment duration variable. For example, we evaluate the GPS for the whole sample at the median treatment duration of group 1, and after that we plot the distribution of the evaluated GPS for group 1 against the distribution of the GPS for the rest of the sample. Like in the case of binary propensity score matching, by inspecting the overlap of these two distributions we can examine the common support condition graphically. In the same fashion, we can test the common support condition of groups 2 and 3 *versus* the rest of the sample. Finally, we restrict our final sample to individuals who are comparable across the three groups simultaneously, i.e. we drop those participants whose GPS is not among the common support region for all three groups.

Finally, in addition to assessing the identifying assumption and common support, in the case of a continuous treatment it is also crucial to evaluate how well adjustment for the GPS works in balancing the covariates, i.e. whether the specification for estimation of expression (5) is adequate. Whereas in the binary case the typical approach is to compare the covariate means for the treated and control units before and after matching, testing for covariate balance is more difficult with continuous treatments. We apply three approaches to test the balancing properties of the GPS.

In the first approach we follow Imai and van Dyk (2004) who proposed to evaluate the balancing of the covariates by regressing each covariate on the treatment variable

(a) without and
(b) with conditioning

on the predicted training duration $E[T|X_i]$. Once we condition on $E[T|X_i]$ we expect that the duration of training is not correlated with the covariate if adjustment for the GPS works in balancing the observable characteristics. We use a log-transformation of the treatment variable since the duration of training is mechanically uncorrelated with each continuous covariate given $E[T|X_i]$ (see appendix B in Imai and van Dyk (2004)).

In the second approach we propose to regress each covariate on the treatment variable

(a) without and
(b) with conditioning

on the distribution of the GPS R_i. R_i is evaluated at different potential values of the treatment level, the 25th, the 50th and the 75th percentile of the training duration. The basic idea is similar to the first approach: once we condition on the distribution of R_i the duration of training should be uncorrelated with the covariate if the GPS properly balances the covariates. For both

Table 2. Effect of duration of treatment on probability of employment at time 2 years after entry into the programme—estimates from a probit model†

Variable	*(1), marginal effect (standard error)*	*(2), marginal effect (standard error)*	*(3), marginal effect (standard error)*	*(4), marginal effect (standard error)*
(a) Only control for duration of treatment				
Treatment duration/100	0.0061	0.0526	0.0630	0.0239
	(0.0085)	(0.0363)	(0.1062)	(0.2488)
Square of treatment duration/100		−0.0112	−0.0175	0.0224
		(0.0085)	(0.0604)	(0.2374)
Cube of treatment duration/100			0.0010	−0.0141
			(0.0099)	(0.0879)
4th power of treatment duration/100				0.0019
				(0.0110)
Pseudo-R^2	0.0001	0.0006	0.0006	0.0006
Number of observations	3162	3162	3162	3162
(b) Control for duration of treatment and other variables				
Treatment duration/100	0.0057	0.0797	0.1912	0.0446
	(0.0100)	(0.0385)	(0.1119)	(0.2595)
Square of treatment duration/100		−0.0182	−0.0850	0.0650
		(0.0092)	(0.0636)	(0.2481)
Cube of treatment duration/100			0.0111	−0.0461
			0.0105	0.0920
4th power of treatment duration/100				0.0072
				(0.0115)
Pseudo-R^2	0.1321	0.1331	0.1334	0.1335
Number of observations	3130	3130	3130	3130

†Dependent variable: employment status at time 2 years after entry into the programme. Source: integrated employment biographies of the German Federal Employment Agency. Marginal effects refer to the effect of a change in the duration of training on the probability of being employed 2 years after entry into the programme, evaluated for the average participant. The sample consists of male participants in training programmes in West Germany for the years 2000–2002. Additional control variables include age, disability, citizenship, educational attainment, vocational attainment and employment history. Duration of treatment in days is divided by 100 to obtain readable effects.

approaches we employ linear regression models in the case of continuous covariates and probit models for binary covariates.

Finally, we follow Hirano and Imbens's (2004) approach of 'blocking on the score' and divide the sample into three groups according to the distribution of length of treatment, cutting at the 30th and 70th percentile of the distribution. Within each group we evaluate the GPS at the median of the treatment variable. Then, in the second step we divide each group into five blocks by the quintiles of the GPS evaluated at the median. Within each of these blocks we calculate the difference in means of covariates with respect to individuals who have a GPS such that they belong to that block but have a treatment level that is different from the one being evaluated. This procedure tests whether for each of these blocks the covariate means of individuals belonging to the particular treatment level group are significantly different from those of individuals with a different treatment level, but similar GPS. A weighted average over the five blocks in each treatment level group can be used to calculate the t-statistic of the differences in means between the particular treatment level group and all other groups. The procedure needs to be repeated for each treatment level group and for each covariate. If adjustment for the GPS properly balances

Table 3. Effect of duration of treatment on probability of employment at time 1 year after exit from the programme—estimates from a probit model†

Variable	*(1), marginal effect (standard error)*	*(2), marginal effect (standard error)*	*(3), marginal effect (standard error)*	*(4), marginal effect (standard error)*
(a) Only control for duration of treatment				
Treatment duration/100	0.0069 (0.0085)	0.0246 (0.0361)	−0.0582 (0.1056)	−0.0962 (0.2479)
Square of treatment duration/100		−0.0043 (0.0085)	0.0454 (0.0602)	0.0842 (0.2367)
Cube of treatment duration/100			−0.0083 (0.0099)	−0.0230 (0.0876)
4th power of treatment duration/100				0.0019 (0.0110)
Pseudo-R^2	0.0002	0.0002	0.0004	0.0004
Number of observations	3162	3162	3162	3162
(b) Control for duration of treatment and other variables				
Treatment duration/100	0.0074 (0.0099)	0.0485 (0.0381)	0.0431 (0.1106)	−0.1124 (0.2581)
Square of treatment duration/100		−0.0101 (0.0091)	−0.0069 (0.0630)	0.1521 (0.2468)
Cube of treatment duration/100			−0.0005 (0.0104)	−0.0611 (0.0915)
4th power of treatment duration/100				0.0076 (0.0115)
Pseudo-R^2	0.1183	0.1186	0.1186	0.1187
Number of observations	3130	3130	3130	3130

†Dependent variable: employment status at time 1 year after exit from the programme. Source: integrated employment biographies of the German Federal Employment Agency. Marginal effects refer to the effect of a change in the duration of training on the probability of being employed 1 year after exit from the programme, evaluated for the average participant. The sample consists of male participants in training programmes in West Germany for the years 2000–2002. Additional control variables include age, disability, citizenship, educational attainment, vocational attainment and employment history. Treatment duration in days is divided by 100 to obtain readable effects.

the covariates, we would expect all those differences in means to be not statistically different from 0. Note, though, that this test might suffer from 'balance test fallacy', i.e. we might observe for some covariates decreased t-statistics which might be driven by increased variances and not by decreased mean differences (see Imai *et al.* (2008) for a general discussion of balance test fallacy in the context of binary matching approaches). It is therefore important to apply multiple approaches to test the balancing property of the GPS.

4. Empirical results

4.1. Estimates from a probit model

As mentioned in Section 2, in this paper we consider two outcome variables: one is the probability of employment at the point in time 2 years after the participants entered the programme. This corresponds to the standard approach in the literature on the effectiveness of training programmes: outcomes of participants and non-participants are compared at specific points in time after the programme started. However, in our analysis participants leave the programme at different points in time. Since the intensity of search for a new job might be low during

Table 4. Effect of duration of treatment on probability of employment at time 2 years after entry into the programme—instrumental variable estimates from a linear probability model†

Variable	*(1), coefficient (standard error)*	*(2), coefficient (standard error)*	*(3), coefficient (standard error)*	*(4), coefficient (standard error)*
(a) Only control for duration of treatment				
Constant	0.3433 (0.0236)	0.3246 (0.0602)	0.3426 (0.1123)	0.0080 (0.1887)
Treatment duration/100	0.0020 (0.0113)	0.0237 (0.0639)	−0.0136 (0.2019)	1.0427 (0.5094)
Square of treatment duration/100		−0.0049 (0.0140)	0.0157 (0.1049)	−0.9798 (0.4466)
Cube of treatment duration/100			−0.0033 (0.0163)	0.3573 (0.1564)
4th power of treatment duration/100				−0.0440 (0.0188)
Adjusted R^2	−0.0002	−0.0001	−0.0006	0.0000
Number of observations	3162	3162	3162	3162
Hausman test: χ^2	0.3000	0.3200	0.3400	11.2500
Hausman test: probability> χ^2	0.8601	0.9563	0.9527	0.0239
(b) Control for duration of treatment and other variables				
Constant	−0.0745 (0.4759)	−0.1768 (0.4787)	−0.2855 (0.4850)	−0.5157 (0.5051)
Treatment duration/100	0.0014 (0.0129)	0.1228 (0.0603)	0.3164 (0.1924)	1.0302 (0.4824)
Square of treatment duration/100		−0.0278 (0.0133)	−0.1347 (0.0999)	−0.8080 (0.4226)
Cube of treatment duration/100			0.0170 (0.0156)	0.2611 (0.1481)
4th power of treatment duration/100				−0.0298 (0.0178)
Adjusted R^2	0.1361	0.1360	0.1350	0.1296
Number of observations	3130	3130	3130	3130
Hausman test: χ^2	0.0500	1.4200	2.2600	6.7600
Hausman test: probability> χ^2	1.0000	1.0000	1.0000	1.0000

†Source: integrated employment biographies of the German Federal Employment Agency. Planned duration is used as an instrumental variable for actual duration; the dependent variable is the probability of being employed 2 years after entry into the programme. The sample consists of male participants in training programmes in West Germany for the years 2000–2002. The Hausman tests refer to the differences of the instrumental variable estimates and ordinary least squares estimates. Additional control variables in panel (b) include age, disability, citizenship, educational attainment, vocational attainment and employment history. Duration of treatment in days is divided by 100 to obtain readable coefficients.

programme participation, it seems to be reasonable alternatively to compare individuals with different durations of training at the same point in time since they left the programme. This ensures that participants with different durations of training have the same time to find a job after leaving the programme. Therefore, we additionally consider the probability of employment at the point in time 1 year after the participants exited the programme. Before presenting results for the GPS, we explore first the relationship between the post-treatment probability of employment and the duration of treatment by using a probit model. Tables 2 and 3 investigate the relationship between the probability of employment at 2 years after entering the programme and 1 year after exit from the programme respectively, with the duration of treatment.

Table 5. Effect of duration of treatment on probability of employment at time 1 year after exit from the programme—instrumental variable estimates from a linear probability model†

Variable	*(1), coefficient (standard error)*	*(2), coefficient (standard error)*	*(3), coefficient (standard error)*	*(4), coefficient (standard error)*
(a) Only control for duration of treatment				
Constant	0.3232	0.2993	0.3603	0.2232
	(0.0235)	(0.0600)	(0.1120)	(0.1878)
Treatment duration/100	0.0102	0.0379	−0.0886	0.3443
	(0.0113)	(0.0637)	(0.2014)	(0.5069)
Square of treatment duration/100		−0.0063	0.0636	−0.3444
		(0.0140)	(0.1046)	(0.4444)
Cube of treatment duration/100			−0.0111	0.1367
			(0.0163)	(0.1556)
4th power of treatment duration/100				−0.0180
				(0.0187)
Adjusted R^2	−0.0002	−0.0005	−0.0005	0.0000
Number of observations	3162	3162	3162	3162
Hausman test: χ^2	0.2000	0.4200	0.1600	1.4500
Hausman test: probability $> \chi^2$	0.6508	0.8120	0.9249	0.4834
(b) Control for duration of treatment and other variables				
Constant	−0.1324	−0.2332	−0.2227	−0.2637
	(0.4792)	(0.4826)	(0.4888)	(0.5076)
Treatment duration/100	0.0151	0.1346	0.1993	0.3405
	(0.0130)	(0.0608)	(0.1939)	(0.4851)
Square of treatment duration/100		−0.0274	−0.0631	−0.1963
		(0.0134)	(0.1008)	(0.4250)
Cube of treatment duration/100			0.0057	0.0540
			(0.0157)	(0.1489)
4th power of treatment duration/100				−0.0059
				(0.0179)
Adjusted R^2	0.1187	0.1163	0.1154	0.1145
Number of observations	3130	3130	3130	3130
Hausman test: χ^2	1.0200	3.8700	4.2100	4.0800
Hausman test: probability$> \chi^2$	1.0000	1.0000	1.0000	1.0000

†Source: integrated employment biographies of the German Federal Employment Agency. Planned duration is used as an instrumental variable for actual duration; the dependent variable is the probability of being employed 1 year after exit from the programme. The sample consists of male participants in training programmes in West Germany for the years 2000–2002. The Hausman tests refer to the differences of the instrumental variable estimates and ordinary least squares estimates. Additional control variables in panel (b) include age, disability, citizenship, educational attainment, vocational attainment and employment history. Duration of treatment in days is divided by 100 to obtain readable coefficients.

We report marginal effects instead of estimated coefficients for our probit model to facilitate understanding of the relationship between probability of employment and duration of treatment. First, Tables 2 and 3 show that there is a positive correlation between the probability of employment and duration of treatment, and a negative correlation between the probability of employment and the square of the duration of treatment with or without controlling for additional variables. However, the estimated coefficients of the duration of treatment are small, and the explanatory power of the duration of treatment is low (see the low pseudo-R^2 in panel (a) of Tables 2 and 3). These suggest that the effect of duration of treatment on the probability of employment is small or negligible. Results from a linear probability model (which are not reported here) are similar to those from the probit model. In Figs O5a–O6b in the on-line appendix we report the effects from the cubic specification of the probit models, which illustrate the average response levels to the corresponding duration of treatment.

However, it is worth noting that a regression-type analysis such as the probit or linear probability models may rely on extrapolation, compare incomparable observations and have a greater risk of misspecifying the model. All of these could potentially bias the estimates. Propensity score methods can alleviate these potential problems to some extent. For a discussion of the advantages of matching methods compared with parametric regressions see for example Ho *et al.* (2007).

4.2. Weak unconfoundedness assumption

The GPS approach is based on weak unconfoundedness as an identifying assumption. The factors which the caseworker takes into account when assigning clients to a training programme centre on the clients' aptitude for a certain job and the likelihood of succeeding in a particular training programme (Section 2). The information that the caseworker bases her decision on is largely congruent with the information that is contained in our data, in particular previous educational attainment and vocational education, and detailed labour market histories capturing all spells of employment and unemployment from the previous 4 years. Controlling for these covariates, we are thus confident that the influence of any remaining unobserved factors is negligible and the application of the GPS justified. Moreover, to investigate the concern that duration of treatment may possibly be endogenous, Tables 4 and 5 report instrumental variables estimates by using planned duration as instrumental variable. Comparing these instrumental variable estimates with the ordinary least squares estimates based on actual durations (which are not reported here), we find that in 15 out of 16 specifications they are not significantly different (see the results of the Hausman test in Tables 4 and 5). This suggests that the actual duration of training does not suffer strongly from endogeneity.

4.3. Generalized propensity score estimation, common support and covariate balance

Our implementation of the GPS follows the procedure that was outlined in Hirano and Imbens (2004) and adapted to our context as presented in Section 3 above. We first estimate the conditional distribution of the length of the training programme (treatment) by applying ordinary least squares (the estimation results are reported in Table O1 in the on-line appendix). To test the common support condition for the actual duration, following the approach that was outlined in Section 3.3, we divide the sample into three groups, which are defined by cutting the distribution of treatment duration at the 30th and 70th percentiles. We then first evaluate the GPS of the whole sample at the median actual duration of treatment of group 1, i.e. 84 days. After that we plot the distribution of this GPS (i.e. evaluated at the median actual duration of the first group) for group 1 and for the other two groups taken together: a procedure which results in Fig. 3(a). We repeat the same procedure for group 2 and group 3, which gives us Figs 3(b) and 3(c). In the second step we delete those individuals from our sample who fall outside the common support region. For the actual duration we delete around 2% of our sample. The distributions for the GPS after deleting the non-overlap are reported in Figs 3(d)–3(f) (corresponding common support graphs for planned duration of training and actual equal to planned duration are reported in the on-line appendix).

To assess the balancing property of the GPS (see Section 3.3), we first regress every observable characteristic on the logarithm of the duration of training and compare the coefficients for specifications without and with conditioning on the predicted individual duration $E[T|X_i]$. The results for the actual duration of training are reported in the first four columns of numbers of Table 6.

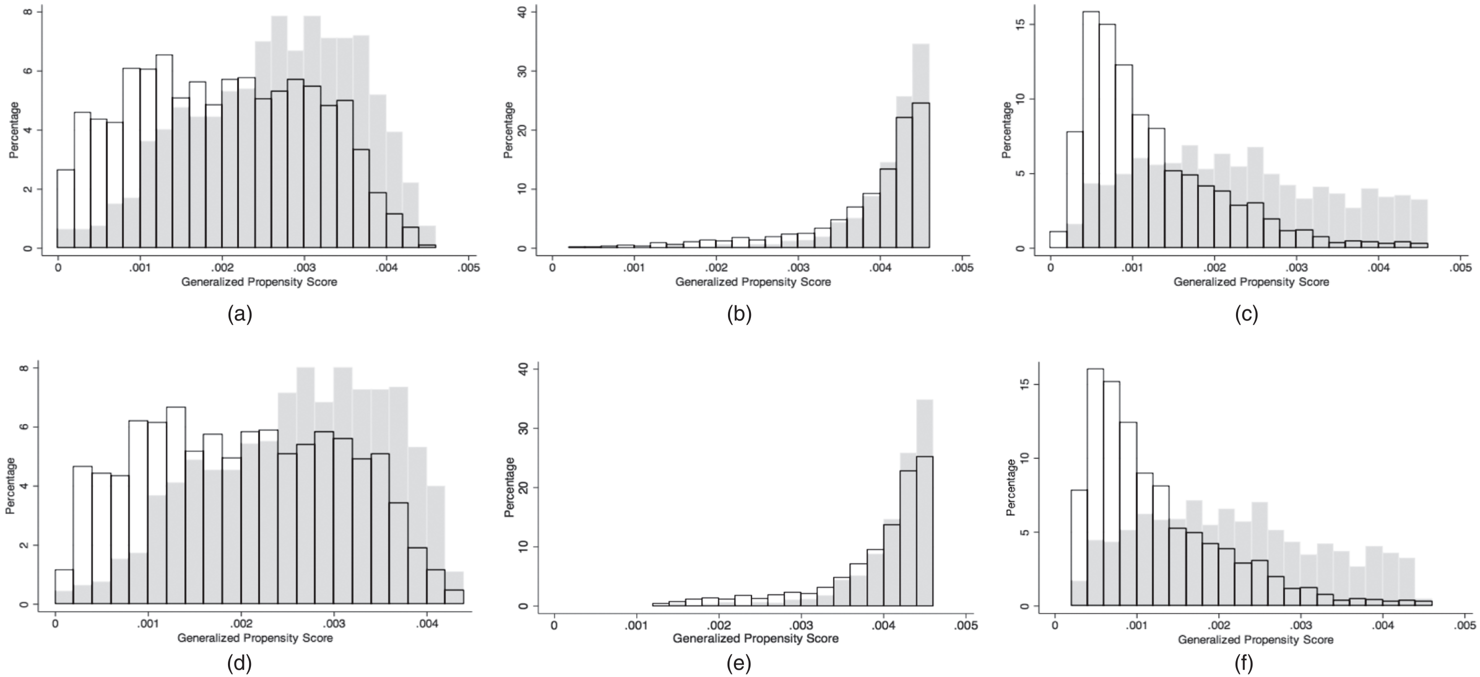

Fig. 3. Common support condition—actual duration of training: before deleting non-overlap for (a) base groups 1 (■) and groups 2 and 3 (□), (b) base group 2 (■) and groups 1 and 3 (□), and (c) base group 3 (■) and groups 1 and 2 (□); after deleting non-overlap for (d) base group 1 (■) and groups 2 and 3 (□), (e) base groups 2 (■) and groups 1 and 3 (□), and (f) base group 3 (■) and groups 1 and 2 (□)

Without conditioning on the predicted duration of training we observe many significant coefficients. For example, age is strongly and positively correlated with observed duration of training. However, once we condition on $E[T|X_i]$, the coefficient for age decreases from 1.17 to 0.11 and is clearly insignificant. This also holds for previous employment outcomes, e.g. the negative correlation between being employed 12 weeks before entering the programme decreases and becomes insignificant. We do not observe any significant correlations for the covariates once we condition on $E[T|X_i]$ and the corresponding coefficients clearly decrease, which indicates that the GPS properly balances the observable characteristics in our sample.

If we compare the regressions of the covariates on the duration of training without and with conditioning on the distribution of R_i, which are reported in the last four columns of Table 6, we come to the same conclusion. All previously significant coefficients become insignificant once we condition on R_i and the corresponding coefficients clearly decrease. This picture is the same for the sample with actual equal to planned duration (see Table O2 in the on-line appendix).

As an additional test for the balancing property of the GPS we apply the approach of blocking on the score that was suggested by Hirano and Imbens (2004). For actual durations, group 1 includes individuals with a treatment level between 11 and 137 days, group 2 ranges from 138 to 247 days and group 3 from 248 to 395 days. For each of the covariates we test whether the difference in means of one group compared with the other two groups is significantly different. Without adjustment the clear majority of t-statistics are greater than 1.96, indicating a clearly unbalanced distribution of covariates. In the second step, we calculate the corresponding t-statistics for the GPS-adjusted sample. In contrast with the unadjusted sample, we observe no t-statistics larger than 1.96 for both the actual duration and the sample with equal planned and actual durations (reported in Tables O3 and O4 in the on-line appendix). These results indicate that the balance of the covariates is clearly improved by adjustment for the GPS and confirm the results of the two other balancing tests.

4.4. Results from estimating the dose–response function

The final step of our empirical analysis consists in estimating the GPS-adjusted dose–response function. The estimation results of the dose–response function for actual duration of training are reported in Table 7. The estimated coefficients in the dose–response function have no direct causal interpretation, and whether all the estimated coefficients associated with the GPS terms are equal to 0 can indicate whether the covariates introduce any bias, as stressed in Hirano and Imbens (2004). We find that, in the case of the outcome variable measured at 2 years after entry into the programme, the coefficients of the GPS terms are more significant than in the case of the outcome variable measured at 1 year after exit from the programme. This implies for the former that it is more important to apply the GPS technique to remove potential bias introduced by the covariates.

Our main results for both outcome variables are presented in Fig. 4 (probability of employment at time 2 years after entry into the programme) and Fig. 5 (probability of employment at time 1 year after exit from the programme), where each figure consists of two parts showing results for

(a) the actual duration and
(b) for the sample of individuals for which actual duration equals planned duration.

Figs 4 and 5 also include the counterfactual non-participant probability of employment baseline (estimated by using standard binary propensity score matching with a sample of non-participants), which indicates that training effects are generally positive. This finding is in line

Table 6. Covariate balance with and without adjustment—actual duration of training†

Covariate	*Unconditional effect of log(duration)*		*Effect of log(duration) conditional on $E[T\|X_i]$*		*Unconditional effect of duration*		*Effect of duration conditional on R_i*	
	Coefficient	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*
Age	*1.1782*	*0.2835*	0.1134	0.3000	*0.8562*	*0.1883*	0.0303	0.2082
Disability								
No disability	*−0.0215*	*0.0081*	−0.0036	0.0083	*−0.0145*	*0.0051*	−0.0017	0.0053
Disability low degree	0.0126	0.0075	0.0030	0.0079	0.0077	0.0048	0.0009	0.0050
Disability medium degree	*0.0063*	*0.0016*	0.0005	0.0006	*0.0035*	*0.0009*	0.0000	0.0002
Disability high degree	0.0021	0.0023	−0.0005	0.0020	0.0018	0.0014	−0.0001	0.0012
Citizenship								
German	0.0146	0.0086	−0.0002	0.0092	*0.0127*	*0.0059*	0.0004	0.0065
Foreigner, European Union	*−0.0073*	*0.0029*	−0.0015	0.0026	*−0.0057*	*0.0022*	−0.0006	0.0020
Foreigner, non-European-Union	−0.0063	0.0081	0.0025	0.0087	−0.0070	0.0055	0.0005	0.0061
Educational attainment								
No graduation	*−0.0317*	*0.0086*	−0.0057	0.0090	*−0.0239*	*0.0061*	−0.0001	0.0065
1st stage of secondary level	*−0.0912*	*0.0138*	0.0165	0.0154	*−0.0830*	*0.0093*	0.0061	0.0108
2nd stage of secondary level	*0.0439*	*0.0122*	−0.0003	0.0129	*0.0347*	*0.0079*	−0.0024	0.0088
Advanced technical college entrance qualification	*0.0187*	*0.0058*	−0.0010	0.0052	*0.0135*	*0.0034*	−0.0004	0.0032
General qualification for university entrance	*0.0706*	*0.0083*	−0.0054	0.0052	*0.0489*	*0.0046*	−0.0017	0.0037
Vocational attainment								
No vocational degree	*−0.0933*	*0.0130*	−0.0013	0.0141	*−0.0763*	*0.0089*	0.0010	0.0101
In-plant training	0.0190	0.0137	0.0100	0.0147	0.0084	0.0091	0.0008	0.0104
Off-the-job training, vocational school, technical school	*0.0204*	*0.0067*	−0.0067	0.0060	*0.0187*	*0.0041*	−0.0015	0.0037
University, advanced technical college	*0.0641*	*0.0063*	0.0003	0.0015	*0.0376*	*0.0035*	−0.0001	0.0006

(*continued*)

Table 6 (*continued*)

Covariate	*Unconditional effect of log(duration)*		*Effect of log(duration) conditional on $E[T\|X_i]$*		*Unconditional effect of duration*		*Effect of duration conditional on R_i*	
	Coefficient	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*
Employment history								
Previous unemployment duration	0.2770	0.2084	0.2234	0.2239	0.0541	0.1385	0.0420	0.1550
Duration of last employment	*64.0300*	*25.0149*	11.5619	26.7505	*41.5635*	*16.6230*	0.5909	18.5098
Log(wage) of last employment	−0.0408	0.0317	−0.0069	0.0341	−0.0303	0.0211	−0.0054	0.0236
No last employment observed	*0.0227*	*0.0079*	0.0056	0.0081	*0.0137*	*0.0049*	0.0019	0.0054
Share of days in employment, 1st year before programme	−0.0561	0.0308	−0.0194	0.0330	−0.0308	0.0205	−0.0085	0.0229
Share of days in employment, 2nd year before programme	−0.0301	0.0454	−0.0500	0.0487	0.0118	0.0302	−0.0073	0.0338
Share of days in employment, 3rd year before programme	0.0022	0.0471	−0.0165	0.0505	0.0155	0.0313	−0.0027	0.0350
Share of days in employment, 4th year before programme	0.0742	0.0476	0.0076	0.0511	0.0555	0.0316	0.0030	0.0353
Share of days in unemployment, 1st year before programme	0.0108	0.0347	0.0127	0.0373	−0.0034	0.0231	0.0026	0.0258
Share of days in unemployment, 2nd year before programme	*−0.0904*	*0.0426*	0.0329	0.0453	*−0.0988*	*0.0283*	−0.0001	0.0314
Share of days in unemployment, 3rd year before programme	*−0.1027*	*0.0422*	0.0293	0.0448	*−0.1056*	*0.0280*	−0.0012	0.0310
Share of days in unemployment, 4th year before programme	*−0.1688*	*0.0411*	−0.0043	0.0434	*−0.1354*	*0.0273*	−0.0066	0.0300
Employment 4 weeks before programme entry	−0.0051	0.0069	−0.0008	0.0074	−0.0024	0.0047	0.0009	0.0052
Employment 8 weeks before programme entry	*−0.0320*	*0.0092*	−0.0119	0.0098	*−0.0188*	*0.0064*	−0.0033	0.0071
Employment 12 weeks before programme entry	*−0.0257*	*0.0108*	−0.0097	0.0115	*−0.0153*	*0.0073*	−0.0041	0.0081
Employment 16 weeks before programme entry	−0.0151	0.0119	−0.0028	0.0127	−0.0108	0.0080	−0.0032	0.0089
Employment 20 weeks before programme entry	−0.0026	0.0125	−0.0016	0.0134	−0.0011	0.0083	−0.0026	0.0093
Employment 24 weeks before programme entry	−0.0067	0.0128	−0.0079	0.0137	0.0002	0.0085	−0.0030	0.0095
Employment 28 weeks before programme entry	−0.0055	0.0131	−0.0086	0.0141	0.0014	0.0087	−0.0034	0.0098
Employment 32 weeks before programme entry	−0.0004	0.0134	−0.0098	0.0143	0.0062	0.0089	−0.0025	0.0100
Employment 36 weeks before programme entry	0.0009	0.0134	−0.0084	0.0144	0.0066	0.0089	−0.0021	0.0100
Employment 40 weeks before programme entry	0.0008	0.0134	−0.0104	0.0144	0.0077	0.0089	−0.0010	0.0100

(*continued overleaf*)

Table 6 (*continued*)

Covariate	*Unconditional effect of log(duration)*		*Effect of log(duration) conditional on $E[T\|X_i]$*		*Unconditional effect of duration*		*Effect of duration conditional on R_i*	
	Coefficient	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*
Employment history								
Employment 44 weeks before programme entry	−0.0130	0.0135	−0.0174	0.0145	0.0016	0.0090	−0.0015	0.0101
Employment 48 weeks before programme entry	−0.0158	0.0135	−0.0151	0.0145	−0.0017	0.0090	−0.0011	0.0101
Regional indicators								
Regional type 1	*0.1043*	*0.0124*	−0.0144	0.0125	*0.0854*	*0.0078*	−0.0026	0.0059
Regional type 2	*0.0589*	*0.0115*	0.0060	0.0119	*0.0402*	*0.0072*	0.0026	0.0077
Regional type 3	*−0.0991*	*0.0129*	0.0078	0.0140	*−0.0885*	*0.0089*	0.0024	0.0100
Regional type 4	0.0069	0.0059	−0.0022	0.0060	0.0067	0.0038	0.0007	0.0039
Regional type 5	*−0.0612*	*0.0101*	−0.0008	0.0102	*−0.0532*	*0.0071*	−0.0017	0.0076
Regional unemployment rate	*0.0180*	*0.0015*	−0.0008	0.0013	*0.0152*	*0.0010*	−0.0006	0.0008

†Source: integrated employment biographies of the German Federal Employment Agency. The second and third columns contain the results of regressing the corresponding observable characteristic on the logarithm of the training duration. The fourth and fifth columns contain the same results conditional on the predicted individual duration $E[T|X_i]$. The sixth and seventh columns report the results of regressing the corresponding observable characteristic on the duration of training, and the last two columns contain the same results conditional on the distribution of the GPS R_i. Italic numbers indicate significance at the 5% level. Duration in days is divided by 100.

Table 7. Estimated dose–response functions—actual duration of training†

	Actual duration	
	Coefficient	*Standard error*
(a) Outcome variable: employment status at time 2 years after entry into the programme		
GPS	−2.4486	1.2408
GPS^2	8.0887	4.8588
GPS^3	−8.0235	5.8881
Program duration	0.0996	0.1769
Program duration^2	−0.0420	0.0822
Program duration^3	0.0022	0.0133
GPS∗Program duration	0.4548	0.6005
GPS^2∗Program duration	−0.9821	0.6654
GPS∗Program duration^2	0.0088	0.1281
Constant	0.4534	0.1105
Adjusted R^2	0.0000	
Number of observations	3070	
(b) Outcome variable: employment status at time 1 year after exit from the programme		
GPS	−0.4982	1.1947
GPS^2	2.3839	4.6814
GPS^3	−3.0652	5.6882
Program duration	−0.0205	0.1683
Program duration^2	0.0349	0.0792
Program duration^3	−0.0082	0.0129
GPS∗Program duration	0.0059	0.5563
GPS^2∗Program duration	−0.1736	0.6275
GPS∗Program duration^2	0.0163	0.1180
Constant	0.3595	0.1067
Adjusted R^2	−0.0019	
Number of observations	3069	

†Source: integrated employment biographies of the German Federal Employment Agency. The sample consists of male participants in training programmes in West Germany for the years 2000–2002. The dependent variable is the probability of being employed 2 years after entry into the programme (panel (a)) and 1 year after exit from the programme (panel (b)). The duration in days is divided by 100. The coefficients of the GPS are reported in Table O1 in the on-line appendix.

with Rinne *et al.* (2011), who implemented a binary treatment comparison training evaluation also using the integrated employment biographies data.

Figs 4(a) and 5(a) show a similar shape for the dose–response function of the two employment outcomes based on actual duration of training. Both graphs indicate a slowly and monotonously increasing employment response to different levels of the treatment until around 240 days (i.e. 8 months). Although steadily increasing, the shape of the dose–response function is relatively flat and implies an increase in the post-training probability of employment of approximately 2 percentage points from a treatment dose of 50 days (employment response approximately 0.34) to a treatment dose of 200 days (employment response approximately 0.36). The confidence bands, however, suggest that this increase might actually be zero. Durations of training that are larger than 240 days do not seem to lead to a further increase in the probability of employment, as indicated by the shape of the dose–response function and the confidence bands.

Looking at the dose–response function of the sample of participants with equal actual and planned durations of training, this pattern becomes more pronounced (Figs 4(b) and 5(b)). For

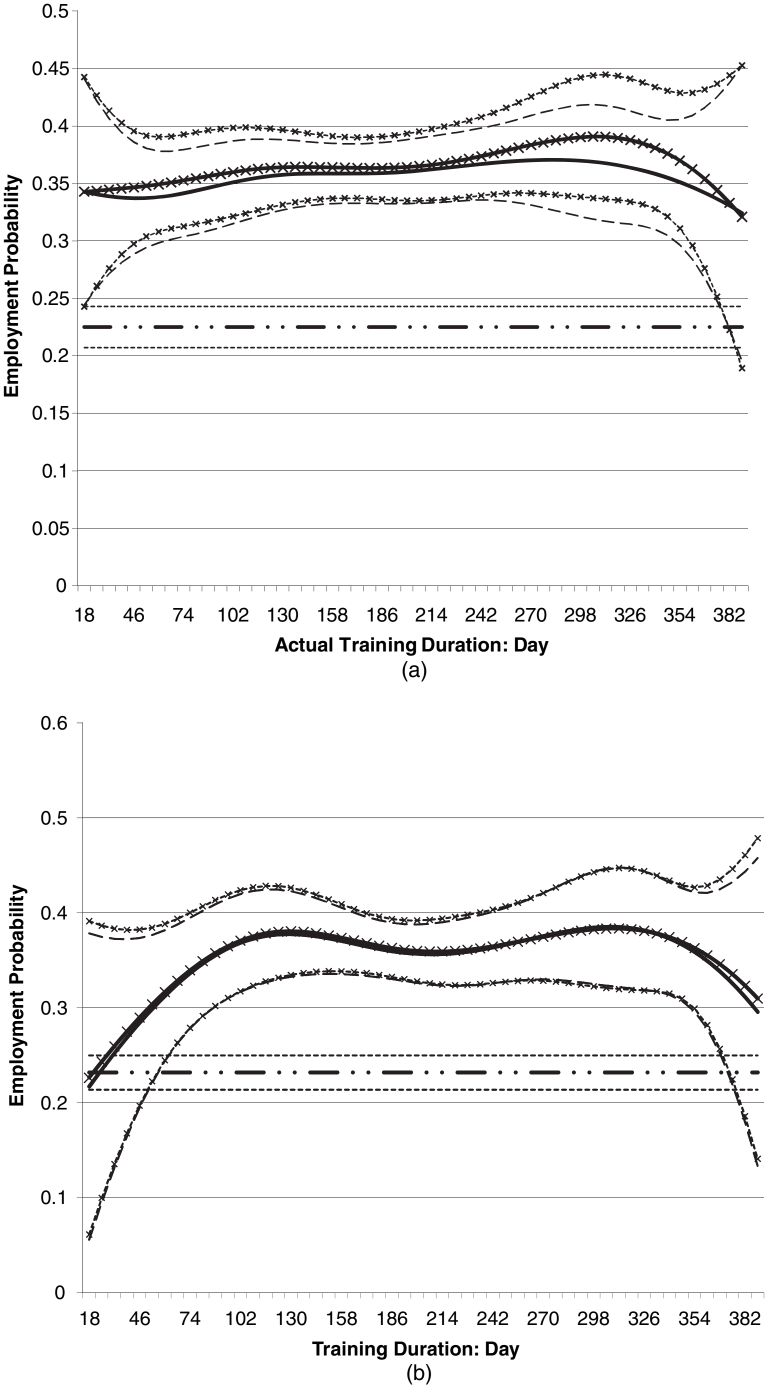

Fig. 4. Dose–response function for the probability of employment at time 2 years after entry into the programme (———, dose–response; – – –, lower and upper bounds; · · — · ·, non-participant baseline; -------, lower and upper bounds; **— x —** , dose–response for subsample (i.e. dose–response function for a subgroup of individuals who went through a training pogramme exactly once); – – x – –, lower and upper bounds for subsample): (a) actual duration of training; (b) actual equal to planned duration of training

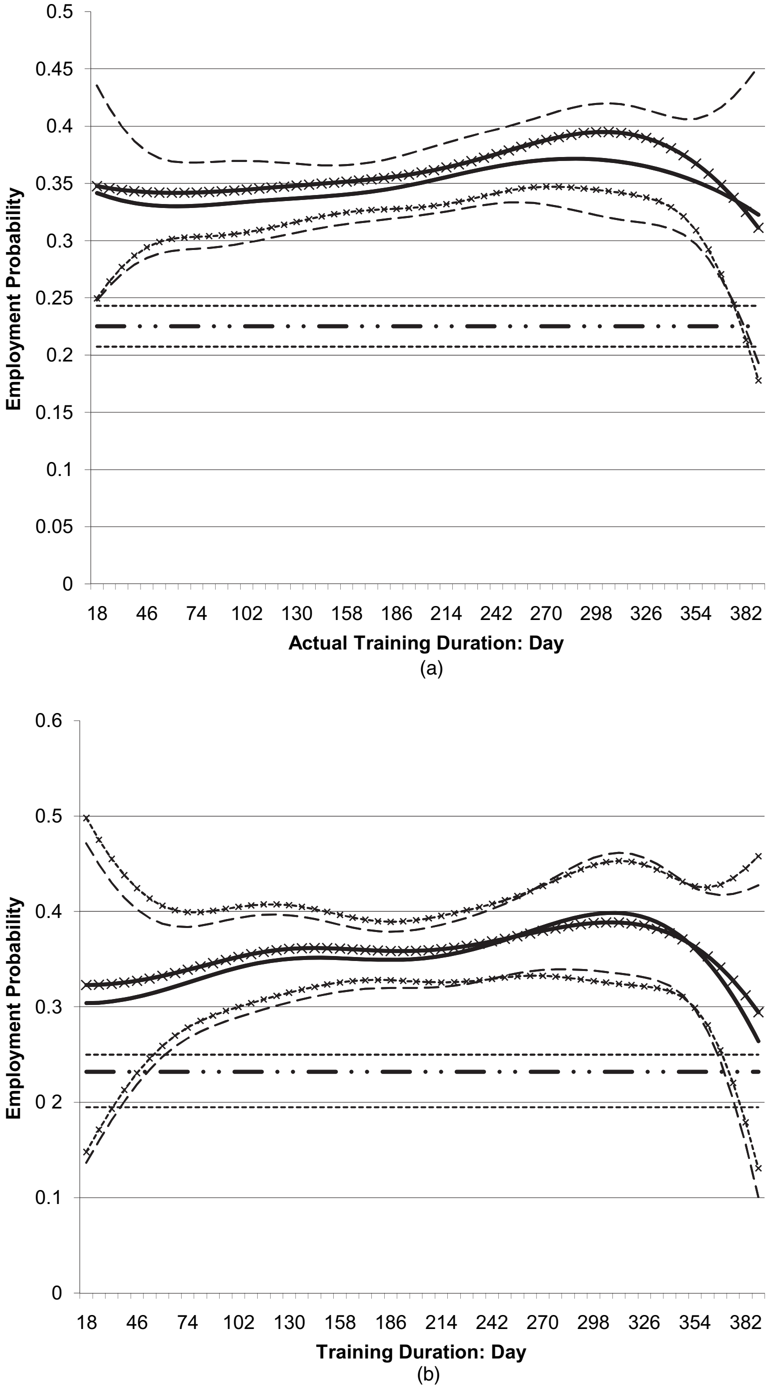

Fig. 5. Dose–response function for the probability of employment at time 1 year after exit from the programme (———, dose–response: – – –, lower and upper bounds; · · — · ·, non-participant baseline; **—x—**, dose–response for subsample (i.e. dose–response function for a subgroup of individuals who went through a training pogramme exactly once); - -x- -, lower and/or upper bound for subsample): (a) actual duration of training; actual equal to planned duration of training

the employment outcome at time 2 years after entry into the programme in Fig. 4(b) there is a clear indication that the increase in the probability of employment that is induced by the duration of treatment occurs mainly during approximately the first 120 days. For longer durations of training the dose–response function is flat and no additional employment impulse seems to be brought about by the treatment. Fig. 5(b) further confirms this general pattern, pointing to a core increase in the probability of employment for programmes up to 120–150 days and no additional effect from a maximum of 200 days onwards (Figs O3 and O4 in the on-line appendix show very similar dose–response functions for planned duration of training).

In addition to the graphs displaying the dose–response function we calculate the pairwise differences in probabilities of employment between different durations of training and bootstrap standard errors from 2000 replications. For example, we calculate the difference in our outcome variables between 18 days of training and 39 days of training, between 18 days of training and 60 days of training etc. This allows us to test whether the effect of different durations of training is significantly different from zero.

Table 8 shows, for example, that for participants with actual equal to planned duration the difference in the probability of employment at time 2 years after entry into the programme between a duration of 18 days and a duration of 102 days is significant at the 10% level. The point estimate suggests an increase of around 15 percentage points (pairwise comparisons of alternative specifications, i.e. for actual durations of training and for the second outcome, are reported in the on-line appendix, Tables O5–O7). If we compare the durations of training of 102 days and 60 days, the estimates indicate an increased probability of employment by 5.6 percentage points, which is significantly different from zero at the 5% level. Comparing the employment effect of training of around 100 days with training of around 300 days indicates a small and insignificant difference. However, for treatment effects 1 year after exit, we do not find significant differences between the different levels of intensity of treatment. Similarly, for the actual duration of training the differential duration impacts also remain largely insignificant, as indicated by the confidence bands in Figs 4(a) and 5(a).

In sum, the GPS-adjusted analysis of the relationship between a continuous training programme and the corresponding probability of employment of participants thus shows a very interesting pattern. Whereas during the first 100–150 days of exposure to treatment increasing the dose yields increasing returns in terms of employment prospects 2 years after entry into the programme, further increasing the dose beyond 150 days appears to bring about no additional treatment effect, i.e. the human capital enhancing features of training are effective during the beginning period (i.e. the initial doses work), but effectiveness fades out after a maximum of approximately 5 months of participation. The initial increase in the probability of employment is much less pronounced for the actual duration of training (Figs 4(a) and 5(a)). This result may be driven by individuals who leave the programme early because they received a job offer. Controlling for early exits by using the sample with planned equal to actual durations leads to estimating a positive effect of approximately the first 120 days of duration of training.

These findings add interesting additional insights into the effectiveness of training programmes for the unemployed in Germany. There are several references analysing effect heterogeneity of training programmes with respect to observable characteristics; see for example Rinne *et al.* (2011). They found only weak evidence for heterogeneous effects with respect to level of skill and age. Other research exploits the difference of long *versus* short training programmes by discretizing the length of training; see for example Lechner *et al.* (2011). They show that longer programmes have a larger and more sustainable positive effect on subsequent prospects for employment. However, in their case the longer duration goes along with a different type of programme—substantive training programmes leading to a vocational degree are compared

Table 8. Duration effects at time 2 years after entry into the programme—actual equal to planned duration of training†

Duration days	*Results for the following durations:*								
	18 days	*60 days*	*102 days*	*144 days*	*186 days*	*228 days*	*270 days*	*312 days*	*354 days*
18	0.000	−0.095	−0.151	−0.158	−0.143	−0.142	−0.158	−0.168	−0.142
		(0.061)	(0.087)	(0.090)	(0.086)	(0.081)	(0.082)	(0.087)	(0.087)
60	0.095	0.000	−0.056	−0.063	−0.047	−0.047	−0.063	−0.072	−0.046
	(0.061)		(0.028)	(0.036)	(0.038)	(0.040)	(0.042)	(0.047)	(0.045)
102	0.151	0.056	0.000	−0.007	0.009	0.009	−0.007	−0.017	0.010
	(0.087)	(0.028)		(0.014)	(0.026)	(0.034)	(0.036)	(0.039)	(0.038)
144	0.158	0.063	0.007	0.000	0.016	0.016	0.000	−0.009	0.017
	(0.090)	(0.036)	(0.014)		(0.016)	(0.026)	(0.029)	(0.035)	(0.035)
186	0.143	0.047	−0.009	−0.016	0.000	0.001	−0.016	−0.025	0.001
	(0.086)	(0.038)	(0.026)	(0.016)		(0.013)	(0.023)	(0.035)	(0.035)
228	0.142	0.047	−0.009	−0.016	−0.001	0.000	−0.016	−0.026	0.000
	(0.081)	(0.040)	(0.034)	(0.026)	(0.013)		(0.018)	(0.035)	(0.037)
270	0.158	0.063	0.007	0.000	0.016	0.016	0.000	−0.009	0.017
	(0.082)	(0.042)	(0.036)	(0.029)	(0.023)	(0.018)		(0.019)	(0.036)
312	0.168	0.072	0.017	0.009	0.025	0.026	0.009	0.000	0.026
	(0.087)	(0.047)	(0.039)	(0.035)	(0.035)	(0.035)	(0.019)		(0.033)
354	0.142	0.046	−0.010	−0.017	−0.001	0.000	−0.017	−0.026	0.000
	(0.087)	(0.045)	(0.038)	(0.035)	(0.035)	(0.037)	(0.036)	(0.033)	

†Source: integrated employment biographies of the German Federal Employment Agency. The column headings are durations of training. Table entries are differences in treatment effects from the duration in the first column compared with treatment effects from the duration in the column headings. Treatment effects are based on the estimated dose–response function. The sample consists of male participants in training programmes in West Germany for the years 2000–2002. Standard errors are bootstrapped based on 2000 replications and are reported in parentheses.

with shorter programmes not leading to a degree. In this paper, we use the GPS framework to exploit the effect heterogeneity of different levels of duration of training within the same types of programme. Our results suggest that an increased duration of training does not automatically go along with higher subsequent probabilities of employment, and that, after a maximum of around 5 months of training, additional time spent in programmes does not have an effect on subsequent prospects for employment. These results indicate that learning specific skills required for a certain vocation or receiving qualifications that are of general vocational use—i.e. the contents of the programmes that were analysed here—have positive expected returns only up to a certain duration.

5. Robustness

In this section, we carry out sensitivity checks for our main specification. Some of the individuals in our sample participate in more than one training programme during their spell of unemployment. Since we consider the duration of the first training programme as the treatment dose, this might not reflect the ‘true’ treatment dose for every unemployed individual in our sample. Therefore the first check is that we restrict our sample to the people who went through a training programme exactly once. In the second check, we try various specifications for the dose–response functions to investigate whether our results are sensitive to the specification that we apply, which includes different degrees of polynomial specifications and flexible spline models to check the functional form specifications.

Figs 4 and 5 above also include dose–response functions for a subsample of our data (labelled ‘dose–response for subsample’ in the graphs). The original data contain information on whether a training participant, after having taken part in the course which we analyse here, participated in another training course at some point in time. These are about 7% of individuals in our sample. We therefore include results for the subsample of observations that participated in exactly the one course for which we have data on planned and actual durations. Regarding the balancing of covariates and the shape of the dose–response functions, results for the subsample are very similar to those for the full sample.

Our main impact estimates are based on a cubic specification for the dose–response function. We also estimate results for the dose–response function for the full sample for quadratic and fourth-degree polynomial specifications as well (see the on-line appendix, Figs O5a,b and O6a,b). These results show that the general shapes and trends of the dose–response functions remain relatively unchanged under different specifications, though there are some differences in detail. Our central finding that the main body of the dose–response functions is rather flat, i.e. longer training programmes do not seem to add an additional treatment effect, is robust. This main result holds true also if we use the logarithm of the duration as the dependent variable for our GPS (see the thin lines in the on-line appendix figures).

Besides experimenting with different degrees of polynomial specifications, we also apply spline models to check the robustness of our results with respect to functional assumptions in our main models. Similarly to a non-parametric regression, which requires the researcher to make decisions regarding for example the type of kernel and proper bandwidth, in the spline models we need to choose the type of the spline, the number of knots and the position of the knots. Among them, the number of knots is the most important; see Keele (2008). Figs 6 and 7 present results from the spline models based on B-splines with 12 knots for the specification based on actual duration. The results are very similar to the results from our main models. The only difference

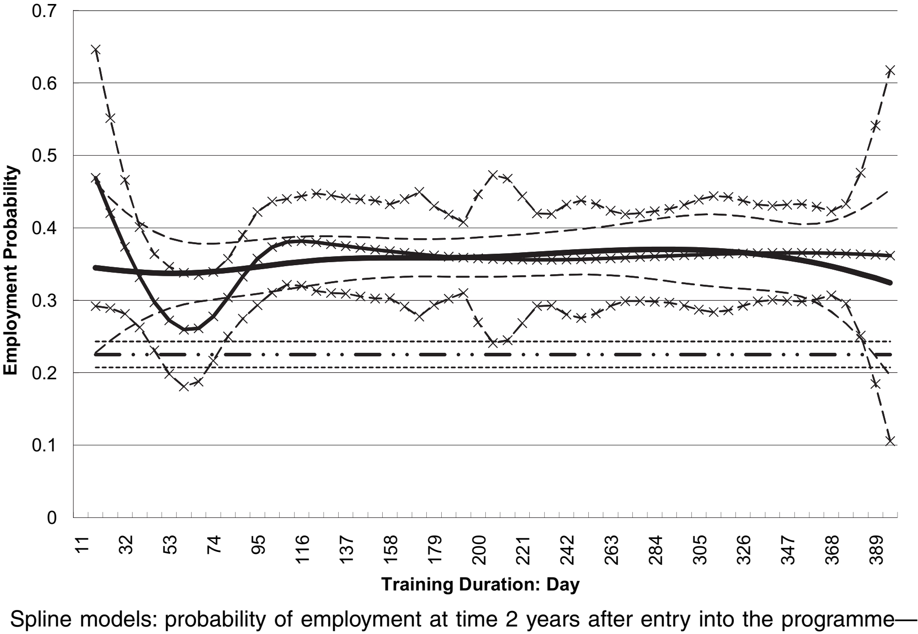

Fig. 6. Spline models: probability of employment at time 2 years after entry into the programme—actual duration of training (———, dose–response; – – – –, lower and upper bounds 1; · · — · ·, non-participant baseline; - - - - - - -, lower and upper bounds 2; —×—, spline (12 knots); – –×– –, lower and upper bounds 3)

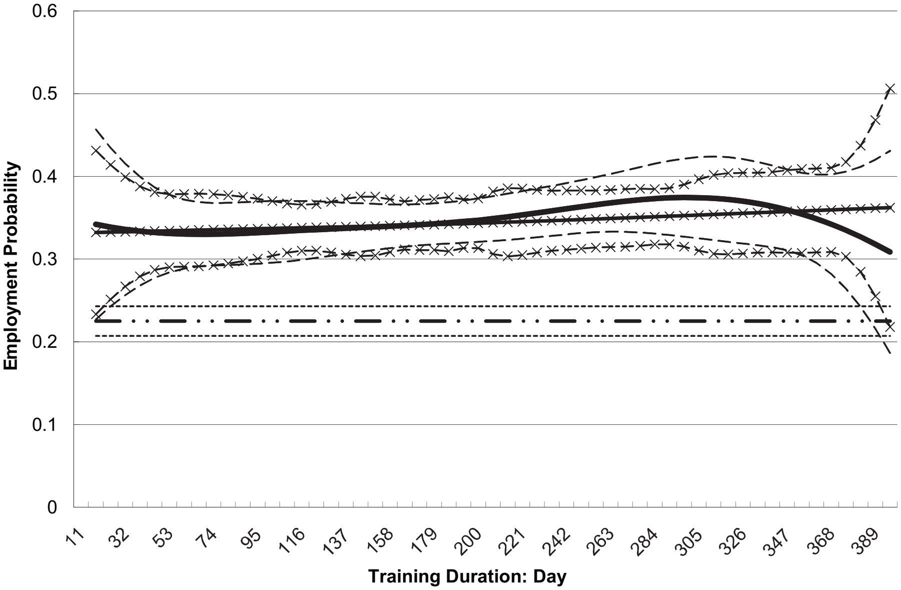

Fig. 7. Spline models: probability of employment at time 1 year after exit from the programme—actual duration of training (———, dose–response; – – –, lower and upper bounds 1; · · — · ·, non-participant baseline; - - - - - - -, lower and upper bounds 2; **—x—**, spline (12 knots); – – × – –, lower and upper bounds 3)

that we observe compared with our main specification is that the results based on the spline model for the outcome at time 2 years after entry into a programme indicate a positive increase from around 50 days to around 100 days of training. However, this positive increase is in line with the results based on the sample of participants for which the actual duration equals the planned duration.

As mentioned earlier, the number of knots is important in the spline models; just like the bandwidth, it controls the smoothness of the fitting. Therefore we varied the number of knots from 6 to 12, and the results are quite similar except at the tails.

6. Conclusion

In this paper, we study the effect of the duration of training programmes on the post-treatment probability of employment, using a particular data set that contains information on duration of training in days for a period of from about 1 week to 13 months. In particular, we are interested in estimating the dose–response function at all possible durations of treatment. We implement this using the recently developed GPS for continuous treatments. Extensive diagnostics on covariate balance, common support and the weak unconfoundedness assumption validate the approach. Moreover, we can consider both the planned and the actual durations as treatment variables, thus avoiding a potential bias in estimating the dose–response function from endogenous exits, which may play a role if only actual durations are observed. We conduct various checks of robustness to solidify our results further.

Our findings indicate that the dose–response function that relates duration of training to the corresponding probability of employment has a relatively flat shape after an initial increase during the first 120 days of training. Indeed, the first 3–5 months of a training programme appear to be the most effective period to improve the probability of employment and to bring about the generally positive effect relative to the non-participant baseline. In the lower segment of the distribution of durations of training, additional doses seem to bring about increases in post-training employment prospects.

After approximately 150 days, however, further participation in the programme does not seem to lead to an increase in the treatment effect, as the dose–response function is basically flat for higher doses. Whether the effect actually even starts to decrease again for the very long durations (longer than 300 days) cannot be said with certainty, as large confidence bands due to a small number of observations exacerbate a precise estimation of this effect. On the basis of our assessment of the dynamics of the individual probability of employment brought about by continuous increases in duration of programmes, there seems to be little justification for training programmes in Germany to last longer than a maximum of about 3–5 months. This conclusion holds for the type of programmes that was analysed in this study, i.e. for training programmes which do not lead to a vocational degree. There is evidence that substantive training programmes with a duration of 2 years and leading to a vocational degree have a large positive effect on subsequent employment outcomes (Lechner *et al.*, 2011).

Acknowledgements

We thank the Joint Editor, an Associate Editor and two referees, Deborah Cobb-Clark, Dan Hamermesh, Oscar Mitnik, Peter Mueser, Donald Parsons, Ulf Rinne, Gerard van den Berg, participants at the Verein für Socialpolitik conference in Munich, the European Cooperation in the Field of Scientific and Technical Research meeting in St Gallen and seminars at the George Washington University, Renmin University of China, Xiamen University, the World Bank and the Washington Statistical Society for valuable discussions and helpful comments, and Jan Stuhler for excellent research assistance. Arne Uhlendorff also thanks DIW DC, where part of this research was pursued during his stay in the autumn of 2007. The usual disclaimer applies.

References

Augurzky, B. and Kluve, J. (2007) Assessing the performance of matching algorithms when selection into treatment is strong. *J. Appl. Econmetr.*, **22**, 533–557.

Blien, U., Hirschenauer, F., Arendt, M., Braun, H. J., Gunst, D. M., Kilcioglu, S., Kleinschmidt, H., Musati, M., Roß, H., Vollkommer, D. and Wein, J. (2004) Typisierung von Bezirken der Agenturen für Arbeit. *Z. Arbeitsmarkt.*, **37**, 146–175.

Diaz, J. J. and Handa, S. (2006) An assessment of propensity score matching as a nonexperimental impact estimator: evidence from Mexico's PROGRESA. *J. Hum. Resour.*, **41**, 319–346.

Eichhorst, W. and Zimmermann, K. F. (2007) And then there were four... how many (and which) measures of active labor market policy do we still need? *Appl. Econ. Q.*, **53**, 243–272.

Flores, C. A., Flores-Lagunes, A., Gonzalez, A. and Neuman, T. C. (2011) Estimating the effects of length of exposure to instruction in a training program: the case of job corps. *Rev. Econ. Statist.*, to be published.

Heckman, J. J., Ichimura, H., Smith, J. and Todd, P. (1998) Characterizing selection bias using experimental data. *Econometrica*, **66**, 1017–1098.

Heckman, J. J., LaLonde, R. J. and Smith, J. A. (1999) The economics and econometrics of active labor market programs. In *Handbook of Labor Economics*, vol. 3 (eds O. Ashenfelter and D. Card). Amsterdam: Elsevier.

Hirano, K. and Imbens, G. W. (2004) The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives* (eds A. Gelman and X. Meng). New York: Wiley.

Ho, D. E., Imai, K., King, G. and Stewart, E. A. (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.*, **15**, 199–236.

Imai, K. and van Dyk, D. A. (2004) Causal inference with general treatment regimes: generalizing the propensity score. *J. Am. Statist. Ass.*, **99**, 854–866.

Imai, K., King, G. and Stuart, E. A. (2008) Misunderstandings between experimentalists and observationalists about causal inference, *J. R. Statist. Soc.* A, **171**, 481–502.

Imbens, G. W. (2000) The role of the propensity score in estimating dose-response functions. *Biometrika*, **87**, 706–710.

Imbens, G. W. (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Statist.*, **86**, 4–29.

Institute for the Study of Labor, DIW and infas (2007) Evaluation der Maßnahmen zur Umsetzung der Vorschläge der Hartz-Kommission—Modul 1b: Förderung beruflicher Weiterbildung und Transferleistungen, Endbericht. *Research Report*. Federal Ministry for Labour and Social Affairs, Berlin.

Jacobi, L. and Kluve, J. (2007) Before and after the Hartz reforms: the performance of active labour market policy in Germany. *J. Lab. Markt Res.*, **40**, 45–64.

Keele, L. (2008) *Semiparametric Regression for the Social Sciences*. New York: Wiley.

Kluve, J. (2010) The effectiveness of European Active Labour Market Policy. *Lab. Econ.*, **17**, 904–918.

Lechner, M. (2001) Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies* (eds M. Lechner and F. Pfeiffer). Heidelberg: Physica.

Lechner, M., Miquel, R. and Wunsch, C. (2011) Long-run effects of public sector sponsored training in West Germany. *J. Eur. Econ. Ass.*, **9**, 742–784.

Moffitt, R. (2009) Estimating marginal treatment effects in heterogeneous populations. *Ann. Econ. Statist.*, fall, 239–261.

Mueser, P. R., Troske, K. R. and Gorislavsky, A. (2007) Using state administrative data to measure program performance. *Rev. Econ. Statist.*, **89**, 761–783.

Rinne, U., Schneider, M. and Uhlendorff, A. (2011) Do the skilled and prime-aged unemployed benefit more from training?: effect heterogeneity of public training programs in Germany. *Appl. Econ.*, **43**, 3465–3494.

Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Schneider, H. and Uhlendorff, A. (2006) Die Wirkungen der Hartz-Reform im Bereich der beruflichen Weiterbildung. *J. Lab. Markt Res.*, **39**, 477–490.

Wunsch, C. (2005) Labour market policy in Germany: institutions, instruments and reforms since unification. *Discussion Paper 2005-06*. Department of Economics, University of St Gallen, St Gallen.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article

‘Evaluating continuous training programs using the generalized propensity score: online appendix’.

Please note: Wiley–Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the author for correspondence for the article.