

Matching, differencing on repeat

Working paper

January 2018

Country

Tanzania

Authors

Michele Binci,

Madhumitha Hebbar,

Paul Jasper,

Georgina Rawle



**Oxford Policy
Management**

Matching, differencing on repeat

Propensity score matching and
difference-in-differences with
repeated cross-sectional data:
Methodological guidance and an
empirical application in education

Working paper

January 2018

Country

Tanzania

Authors

Michele Binci,

Madhumitha Hebbar,

Paul Jasper,

Georgina Rawle

About Oxford Policy Management

Oxford Policy Management is committed to helping low- and middle-income countries achieve growth and reduce poverty and disadvantage through public policy reform.

We seek to bring about lasting positive change using analytical and practical policy expertise. Through our global network of offices, we work in partnership with national decision makers to research, design, implement, and evaluate impactful public policy.

We work in all areas of social and economic policy and governance, including health, finance, education, climate change, and public sector management. We draw on our local and international sector experts to provide the very best evidence-based support.

Oxford Policy Management Limited
Registered in England: 3122495

Level 3, Clarendon House
52 Cornmarket Street
Oxford, OX1 3HJ
United Kingdom

Tel: +44 (0) 1865 207 300
Fax: +44 (0) 1865 207 301
Email: admin@opml.co.uk
Website: www.opml.co.uk
Twitter: [@OPMglobal](https://twitter.com/OPMglobal)
Facebook: [@OPMglobal](https://www.facebook.com/OPMglobal)
YouTube: [@OPMglobal](https://www.youtube.com/OPMglobal)
LinkedIn: [@OPMglobal](https://www.linkedin.com/company/OPMglobal)

Abstract

When evaluating programme impact in a context where a randomised control trial is either infeasible or not appropriate, the quasi-experimental approach of Propensity Score Matching (PSM) is often used to construct a counterfactual. However, if there are imbalances remaining after PSM, selection bias may persist. Increasingly, researchers combine PSM and Difference-in-Differences (DID) to counter such imbalances. While there is guidance on applying this combined approach using panel data, applications of this approach in repeated cross-section settings are less frequent. In this paper, we present an innovative approach to combining PSM and DID when only cross-sections of data are available. We illustrate the methodology in the evaluation of EQUIP-T, a UK Department for International Development-funded education intervention in Tanzania. EQUIP-T is a four-year programme focused on improving teacher performance, school leadership, and community participation, aiming to increase the quality of primary education and improve pupil learning outcomes. This study is likely to represent the first practical application of this PSM with DID procedure for a repeated cross-section in an education evaluation. This paper will review the implementation of the methodology in the context of the EQUIP-T programme. It will also discuss strengths, appropriate contexts, and caveats to the approach, considering unobservable characteristics, time-variant imbalances, implementation of concurrent programmes, and challenges in calculating standard errors. In the first approach, the Average Treatment Effect on the Treated (ATT) was compared across time, between baseline and midline. In the second, PSM was used to match treatment units (pupils and teachers in EQUIP-T schools) over time to construct a pseudo panel from repeated cross-sections to estimate overall ATT. In the absence of panel data, the conventional PSM approach of matching individuals at baseline and then calculating impact at endline is not possible. The innovative pseudo panel approach addresses this, following a suggestion by Blundell and Costa Dias (2000, p. 451). Impact estimates on pupil tests are presented. Pupils' test results were classified into one of five curriculum-linked performance bands in Swahili and in Mathematics. The PSM-DID analysis finds strong evidence that EQUIP-T has reduced the proportion of pupils in the bottom performance band for Swahili in programme schools. These results remain strong and highly significant across both our PSM with DID strategies.

1 Introduction

Randomised control trials (RCTs) are a widely used experimental design in counterfactual-based evaluations, as they are considered one of the most robust approaches by which to minimise the risk of confounding factors when measuring the impact of interventions. However, it is not always feasible or appropriate to evaluate a policy change or a programme using an RCT. For example, many health and education programmes are purposively targeted at the poorest geographical regions. In such settings, comparing recipients and non-recipients leads to biased estimates of impact as the two groups are likely to be dissimilar in terms of their individual characteristics. Consequently, the evaluator is faced with the challenge of selection bias, i.e. disentangling programme-driven differences from pre-existing differences between the two groups. Therefore, the fundamental problem that any alternative counterfactual-based evaluation design must address is constructing a comparison group such that selection bias is removed.

There are several quasi-experimental designs employed in the literature to construct a valid counterfactual, and it is the type of data available and the programme allocation rule that typically determine the choice of evaluation method (Blundell and Dias, 2009). In this paper, we consider a scenario where the programme is purposively targeted and data are available for both recipients and non-recipients independently sampled at two points in time – baseline (i.e. pre-programme) and midline (i.e. post-programme). The two methods often used to obtain an unbiased estimate of impact under these conditions are Propensity Score Matching (PSM) and Difference-in-Differences (DID). PSM is a method wherein recipients and non-recipients are matched on the estimated probability of participation based on their observable characteristics, thereby creating a comparable counterfactual. DID is an analytical approach where imbalances between the two groups are differenced over two waves of data to isolate attributable impact. The combination of these two approaches, the PSM-DID estimator, is also becoming increasingly popular as their combined strengths offset their individual weaknesses. The main appeal of this combined approach is that it mitigates both selection on observables (PSM) and selection on unobservables (DID). However, the bulk of the empirical literature applying PSM-DID procedures uses panel data, and detailed guidance on implementing the PSM-DID estimator in the absence of panel data is limited. This is a gap to be filled as repeated cross-sectional data is more common than panel data in policy/programme evaluations. Therefore, the aim of this paper is to discuss practical considerations to be made when implementing PSM-DID using repeated cross-sectional data and illustrate its implementation using the evaluation of an education intervention in Tanzania.

This paper is organised as follows: Section 2 presents a brief literature review. Section 3 provides an overview of the programme, evaluation design, and data sources. Section 4 discusses our estimation strategy and its implementation. Section 4 presents results. Section 5 outlines some of the limitations, and Section 6 concludes the paper.

2 Related literature

Heckman *et al.* (1997) were first to demonstrate that the PSM-DID estimator removes selection on both observables and unobservables. In one of the first empirical applications, Smith and Todd (2005) apply PSM, DID, and PSM-DID estimators to estimate the impact of a labour market training programme and find that the PSM-DID estimator is the most robust among the three estimators. An early suggestion on how these methods could be adapted to repeated cross-sectional data is found in Blundell and Costa Dias (2000). However, practical applications remain rare, especially in policy evaluations. Aerts and Schmidt (2008) constitute the most comprehensive application of this estimation strategy; they obtain PSM, DID, and PSM-DID estimators to assess whether public research and development (R&D) subsidies crowd out private R&D investment in Flanders and Germany. Since funded firms are likely to differ from non-funded firms, they use matching procedures to construct a valid counterfactual. They implement three matching processes: For each treatment firm i in the post-treatment period (T_1), a statistically similar control firm h is found in the same period. For each treated firm i and non-treated firm h in T_1 , a comparable firm, i.e. k and j respectively, is found in the pre-treatment period (T_0). Since matching does not counter unobserved time-invariant heterogeneity, they combine this with DID in the second step. Therefore, the temporal difference between firm i and j is subtracted from the temporal difference between firm h and k to estimate the treatment effect.

Similar methods are applied by a small number of other studies, including: Hong (2013) to estimate the effect of file sharing technology on record sales; Bönke *et al.* (2013) to estimate the extent to which fiscal equalisation schemes lead to states under-exploiting their tax base in a federation; Hashim and Strong (2015) to examine whether a form of risk assessment reduces target price errors made by equity analysis; and Ordine and Rose (2016) to study the effects of a labour market deregulation policy in Italy. With the exception of Aerts and Schmidt (2008), the focus of all other papers is on advancing the thematic literature further. While Aerts and Schmidt (2008) present an overview of the empirical implementation, a more nuanced discussion on the implementation of the PSM-DID estimator is absent. Our paper seeks to fill this gap in the methodological literature by delineating the key elements underpinning the method's implementation.

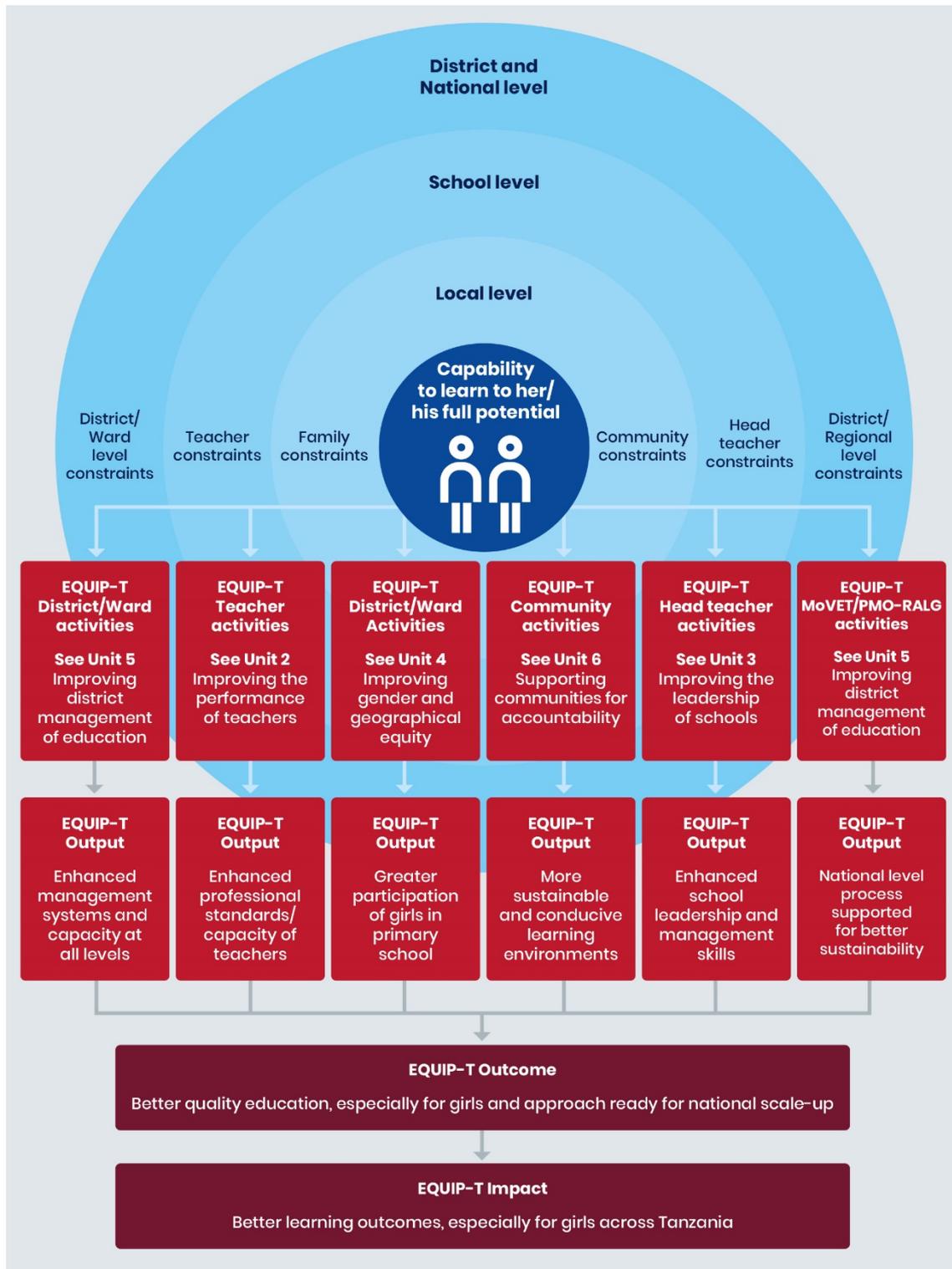
3 Programme and evaluation context

3.1 The programme

The Education Quality Improvement Programme Tanzania (EQUIP-T) is a four-year, UK Department for International Development-funded, Government of Tanzania programme focused on **improving teacher performance, school leadership, community participation, and district management of schools**, aiming to increase the quality of primary education and improve pupil learning outcomes. The programme started in 2014 in five of mainland Tanzania's 26 regions. Two years later it expanded into seven regions. The programme has now been extended to 2020 with a further two regions added.

The programme was designed based on a theory of change (ToC) captured in Figure 1. It identifies six groups of constraints acting on pupils' capability to learn to their full potential.

Figure 1: EQUIP-T programme ToC



Source: Cambridge Education (2014).

The programme's overarching theory is that, by reducing or removing these constraints, the quality of education and pupil learning will improve. The programme has grouped its interventions into five components (reduced from the six shown in Figure 1), each related to a set of constraints. Each component is linked to a programme output. Gender is a cross-cutting theme, and gender-specific interventions are included under each component. The five outputs are:

- Output 1: enhanced professional capacity and performance of teachers;
- Output 2: enhanced school leadership and management skills;
- Output 3: strengthened systems that support the district and regional management of education;
- Output 4: strengthened community participation and demand for accountability; and
- Output 5: strengthened learning and dissemination of results.

Together, changes in these five outputs are intended to reduce constraints on pupil learning and thereby contribute to better-quality education (outcome) and ultimately improved pupil learning (impact).

3.2 The evaluation design

The EQUIP-T Managing Agent purposively selected the regions and districts into the programme on the basis of these being disadvantaged in terms of education and other social and economic indicators. The purposive allocation makes a randomised evaluation of EQUIP-T non-viable, and therefore a quasi-experimental evaluation with matched control schools was designed. This approach creates an ideal setting in which to implement the PSM-DID estimator.

3.3 The data

This paper relies on the baseline (2014) and the midline (2016) data that were collected from a panel of 100 treatment schools and 100 comparison schools in Tanzania. Within schools, data were elicited from head teachers, teachers, pupils, and parents. The total sample comprises just under 3,000 Grade 3 pupils and over 800 teachers across treatment and comparison schools in each survey wave. However, data on teachers and pupils are not longitudinal, i.e. a new cross-section of teachers and pupils is sampled at each survey wave, thus resulting in two cross-sections of data.

4 Estimation strategy

4.1 PSM

The key problem that matching attempts to solve is the problem of selection bias. As geographical areas were purposively selected to receive the intervention, pupils and teachers from schools that did receive EQUIP-T support could be systematically different from individuals in control schools that did not receive such support. Simple comparisons of indicators across such dissimilar groups would be invalid and biased to infer programme impact.

Matching tackles this problem by constructing a control group of pupils (or teachers) who are similar to treated pupils (or teachers) in terms of a number of relevant characteristics. Essentially, matching limits the impact estimation sample to statistically identical treatment and control observations. This is done by matching and comparing outcomes for units in the treatment group with control units that are as similar as possible to each other according to a set of relevant observable characteristics, i.e. comparing like with like only. In this study we use PSM, which is one among the many established matching methods.

PSM is a two-stage analytical approach that employs a propensity score as a 'comparator metric' that summarises the information of the set of relevant characteristics, i.e. the ones that drive selection bias. This propensity score can also be interpreted as an estimation of the hypothetical probability of any individual being in the treatment group, given its characteristics. The first stage of any PSM analysis is to compute a valid propensity score for each unit of observation. The second stage is to then compare outcome indicators of interest across units (i.e. teachers or pupils in this case) with similar propensity scores. Note that because outcome indicators from treatment units are compared to outcome indicators from specific control units based on the propensity score, the estimated average treatment effect will be valid for the group of treatment observations only. This means that PSM allows the estimation of an Average Treatment Effect on the Treated (ATT). Extrapolating this estimate beyond the population for which the treatment sample is representative is not possible.

4.1.1 PSM first stage model selection

The validity of any PSM approach also depends on how well it reduces any imbalance, and thereby selection bias, between treatment and control groups. Achieving balance means that if matched appropriately treatment and control groups' characteristics will not be significantly different from each other. In other words, this means that, across the list of relevant characteristics that are assumed to drive selection bias, the treatment and control groups will be statistically similar to each other.

To estimate the propensity score in the first stage, this study built on the procedure suggested by Imbens and Rubin (2015, p. 281 ff.). The underlying model specification for this procedure is either a logit or probit regression for the first stage. This means

that the propensity scores are estimated by first specifying treatment and control assignment as a binary variable that has the values 0 (for control) and 1 (for treatment). The estimated scores are then modelled as the fitted values that are derived from a logit or probit estimation, with the binary treatment variables as dependent variables and the covariates across which balance is supposed to be achieved as the regressors. These fitted values lie between 0 and 1.

To be more concrete, in the case of a logistic regression specification, the binary response variable is modelled as follows:

$$\Pr(T = 1 | X_i) = \frac{e^{f(X_i)}}{1 + e^{f(X_i)}}, \quad (1)$$

where $\Pr(T = 1 | X_i)$ is the probability of the treatment indicator (T) being equal to one, conditional on the covariates (X_i) for unit i . The function $f(X)$ is normally modelled linearly, i.e. is of the form $f(X) = X\beta$. The coefficients of this function (β) are estimated using maximum likelihood techniques. The fitted values, i.e. the predicted probabilities that follow from this procedure, are the propensity scores for each unit of observation.

The key question for the first stage is which covariates to include in $f(X)$ so that this procedure produces a valid estimate of the propensity score. Building on the procedure described in Imbens and Rubin (2015) for selecting covariates, this study implemented the following four-step approach to make this decision.

1. Select a set of basic covariates based on substantive grounds

The starting point for the PSM analysis was to select variables that were likely to be relevant and valid to be used for this analysis from a theoretical perspective. 'Relevant' implies that variables selected were theoretically expected to be correlated with treatment status and treatment effects, thereby introducing selection bias in a simple comparison of treatment outcomes between control and treatment groups.

To be 'valid', variables had to be unaffected by the programme. In a repeated cross-sectional setting, PSM is implemented at both baseline and midline, using a consistent selection model. However, outcome variables at midline are influenced by the programme, and therefore we restricted the list of valid variables to those unaffected by the programme.

2. Increase the set of valid covariates based on algorithmic approaches

In addition, we employed forwards and backwards stepwise regressions to rationalise the number of covariates. The underlying idea behind both approaches is to check each covariate, step-by-step, for significant correlation with the outcome and treatment assignment variable separately. We set the level of significance at 5%, and only those variables that showed significance in either of two sets of regressions were retained for further consideration.

3. Increasing the set of covariates with polynomial and interaction terms using algorithmic selection

In a third step, we employed the same method of stepwise regressions (backwards and forwards) to augment the set of covariates by quadratic terms or interactions of variables that had already been selected in steps one and two. The rationale behind

this is the fact that balance might only be achieved if the propensity score is estimated using non-linear transformations of the variables selected in the first two steps (Imbens and Rubin, 2015, p. 287).

4. Assessing whether the covariates are compatible with the midline data

In the case of panel data, matching variables are selected using baseline data. However, model selection in a repeated cross-sectional setting is less straightforward; an additional step involved ensuring that the selection model was consistent across both datasets. Consequently, any variables in the baseline selection model that displayed multi-collinearity in the midline data were dropped from the selection model. The result of this process was the identification of an optimal selection model comprising a set of covariates that were included in the first stage estimation of the propensity score.

4.1.2 Second stage algorithm selection

There are a variety of algorithms available to implement the second stage of PSM, i.e. to match control and treatment units to each other based on the propensity score estimated in the first stage. For all approaches, the goal is to find appropriate (i.e. sufficiently similar) control group members for treatment group members. We follow Caliendo and Kopeinig (2005) in determining the appropriate algorithm for this study.

Selecting the appropriate matching algorithm for a PSM exercise is not straightforward and requires careful analysis of how well balanced samples are after employing algorithms with certain sub-specifications. In general, however, the selection of models in this study was based on the fact that discriminating between models poses a bias/variance trade-off in the estimated treatment effect.

We selected kernel matching with appropriate trimming and enforcement of common support as the main algorithm as it is a good compromise between these different approaches. In order to find the optimal estimation model this study used different kernel matching algorithms with different bandwidths and trimming levels. These different results were then compared with respect to the best balancing properties, with the best performing approach being selected as the optimal one. This was again conducted for each estimation strategy for each of the outcome variables and for both rounds of data separately.

4.1.3 Key PSM assumptions: common support and conditional independence

There are two key assumptions that need to hold for PSM to be a valid approach to estimating treatment effects: the common support assumption and the conditional independence assumption.

The **common support assumption** states that the estimated propensity score for all individuals in the treatment and control groups must lie within 0 and 1. Expressed differently, individuals in both groups must have a positive non-zero probability of belonging to either the treatment or control group and the distribution of those

probabilities across the two groups must be such that comparable individuals across the groups can be found. This can easily be enforced by only comparing observations with appropriate propensity scores.

The second key assumption is the **conditional independence assumption**, which posits that, once observable characteristics have been accounted for, the outcome measure is not related to the treatment status anymore, other than via the effect of the programme. This means that any bias that arises due to participation in the programme has been dealt with. Note that this includes biases that arise due to unobservable factors – PSM cannot control for these and the assumption is that once observable characteristics have been dealt with no unobservable bias will remain.

The validity of any PSM approach therefore crucially depends on how well the approach reduces any imbalance between treatment and control groups. If the groups show good balancing properties on observables, then it is reasonable to assume that there is no imbalance on unobservables. Therefore, assessing the balance of covariates after matching is a key step for any PSM analysis. The more balanced samples are after matching, the more plausible is it that the conditional independence assumption holds. The following paragraphs explain how balance assessments were implemented in the current study.

4.1.4 Assessing balance

To select between different matching algorithms and to assess covariate balance after matching, we compared matching models along a variety of dimensions. First, individual covariate balance was assessed across samples by looking at the standardised difference in means across treatment and control groups both before and after matching. This standardised difference is the difference in group averages over the square root of the average of the sample variances. If samples are balanced, this difference should be small and matching should reduce this standardised difference in comparison to the unmatched samples.

In addition, we performed t-tests to assess whether differences across treatment and control groups were statistically significant. If balance is achieved with PSM, differences between treatment and control groups should be negligible and therefore should not be significantly different from zero.

In this context, the covariates' variance ratios of the treated over the control measures was also assessed. If there is perfect balance across samples, then covariates should be distributed equally and hence this ratio should be equal to one.

All these measures give an indication of whether specific individual covariates are balanced across treatment and control groups. To assess overall variance, this study used two statistics that summarise covariate balance in the sample at hand: Rubin's B and Rubin's R. Rubin's B reflects the absolute standardised difference of the means of the propensity score in the treated and control groups (unmatched and matched). Rubin's R is the ratio of the treated to control variances of the propensity scores. Rubin (2001) suggests that the value of B should lie below 25 and that R should lie between .5 and 2 for overall balance to be sufficient. Together, Rubin's B and Rubin's R provide an informative indication of the trade-off between bias and variance across the

treatment and control groups, as both test results change before and after the matching procedure.

Matching procedures were implemented using the `psmatch2` package in Stata (14.1) and balancing tests were carried out using the `pstest` package, which provides the results for all of the statistics mentioned above.¹

Finally, the distribution of propensity scores was also analysed graphically. Ideally, propensity scores should be distributed equally across treatment and control groups. Very skewed/diverging distributions could be an indication that balance has not been achieved successfully.

PSM was used as the core strategy to answer questions of programme impact. However, some outcome indicators showed significant difference between treatment and control groups, despite showing appropriate covariate balance, at baseline.

In order to address this issue, the current study combined PSM with a DID approach. In the following sections, we present the general theoretical principles underlying DID, and then discuss how the two methods have been combined in this study.

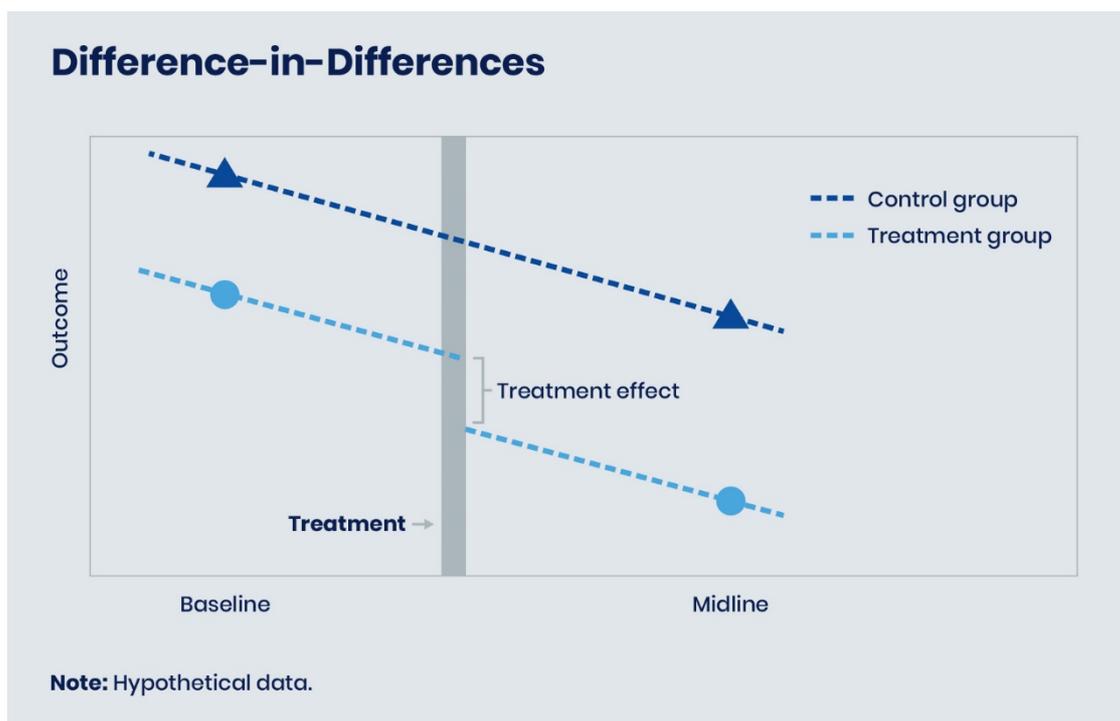
4.2 DID

DID is an approach that exploits the fact that data from the same treatment and control schools were collected at two points in time, i.e. at baseline and at midline. The idea behind this approach is quite straightforward: it compares data from treatment and control schools both at baseline and midline. First, this happens separately. Then, in a second step, these baseline and midline comparisons are compared to each other. If, for example, the difference at baseline between treatment and control was smaller than at midline, this would indicate that the treatment has had an effect on treatment observations. Figure 2 below exemplifies this logic.

In the present case, the comparisons at baseline and midline in the first step are not simple comparisons of descriptive statistics but rather PSM estimations of any statistical significant differences between treatment and control groups. Estimates from these are then, in a second step, compared to each other across time. The key impact estimates presented in this paper are the results of this double difference of PSM estimates.

¹ See <http://fmwww.bc.edu/repec/bocode/p/pstest.html> for details.

Figure 2: Visual representation of DID analysis



The key assumption that needs to hold for DID to identify programme effects is that, as can be seen in Figure 2 above, without the treatment (i.e. the EQUIP-T intervention) the difference between control and treatment groups at the second time point (i.e. the midline of the EQUIP-T evaluation) would have been the same as in the first time point (i.e. the baseline of the EQUIP-T evaluation). This is referred to as the parallel trend assumption.

In the present case, this means that, without the treatment, imbalances remaining after PSM would be the same at baseline and at midline. Note that this means that such imbalances must be assumed to be constant across time. Taking the second difference across time removes such baseline imbalances from the estimation, which hence allows programme impact to be isolated and robustly inferred.

Importantly, for panelled observations, this also includes time-invariant unobservable characteristics that might be correlated to the outcome measure and the treatment status. In the present case, this means that any such school-level characteristics are also controlled for. This increases the robustness of findings because PSM alone cannot control for unobservable characteristics driving selection bias.

Therefore, combining DID with PSM helps to control for remaining imbalances that may exist between treatment and control groups after matching. Taking the difference between matched comparisons at baseline and at midline allows researchers to isolate with confidence the programme impact on beneficiaries (i.e. teachers and pupils in this case).

4.3 Combining DID and PSM

In this study, two different approaches have been used to combine PSM with DID:

1. **Strategy A:** Directly comparing ATT estimates at midline and baseline across time.
2. **Strategy B:** Matching treatment observations across time to construct a pseudo panel of treatment observations and to construct an overall ATT estimate using this pseudo panel only.

In Strategy A, the impact estimate is derived as the direct difference of baseline and midline estimations of ATTs derived from PSM at baseline and midline. Essentially, this amounts to comparing two estimated treatment coefficients with each other. In theory, ATT estimates at baseline should be close to zero because EQUIP-T had not started at that time yet. However, as described above and as can be seen in Section 4, this was not always the case, despite good balancing performance of models at baseline. Taking into account the ATT estimate at midline therefore means that the overall impact of EQUIP-T is defined as the difference that EQUIP-T made in the estimated ATT at midline, compared to the baseline estimate:

$$ATT_{overall} = ATT_{midline} - ATT_{baseline} . (2)$$

Of course, the main goal is to conduct inference on this estimate, i.e. to see whether the overall ATT estimate is different from zero or not. Test statistics for the estimate defined in Equation 2 are calculated using the formula for comparing coefficient estimates as presented in Paternoster *et al.* (1998). Using this test statistic, this study then calculates whether the estimated ATT is significantly different from zero or not from a statistical point of view. Note that all standard errors for the midline and baseline ATT used are based on bootstrapping procedures for PSM estimates.

Strategy B is the main innovation of this paper. There, additional matching is used to create a ‘pseudo panel’ of treatment observations (i.e. teachers and pupils in EQUIP-T schools) across time, given that these have not been panelled and were surveyed as repeated cross-sections. Figure 3 depicts this process graphically.

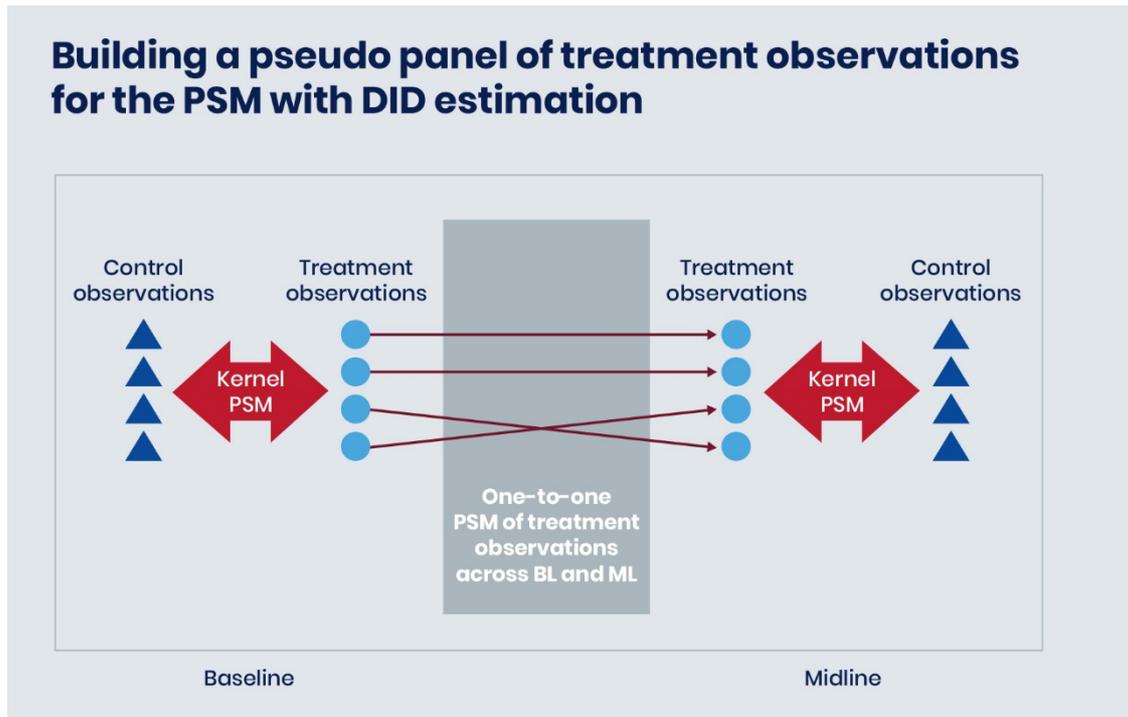
In a first step, treatment observations from teacher and pupil samples are uniquely matched across the two time periods. This is done using a Nearest Neighbour PSM approach without replacement. This means that for each treatment observation at baseline a unique comparator is found at midline.

For this ‘pseudo panel’ of treatment observations, values obtained for their respective matched comparisons at baseline and midline are then used to calculate differences between estimated control group and treatment group individuals at baseline and at midline separately, using the same PSM models as in the main estimations. Note that kernel matching at baseline and midline provides, for each treatment observation, an appropriate estimated counterfactual value based on the PSM estimation. This value is used to calculate the first difference between treatment observations and counterfactuals, as part of the double differencing approach underpinning the DID analysis. In a final step, those differences are then compared across baseline and

midline for the ‘pseudo panel’. The average of this double difference for the pseudo panel is the estimated overall ATT. Note that, in implementing this approach, the current study is one of only a handful to follow a suggestion by Blundell and Costa Dias (2000, p. 451).

The key difference between strategies A and B is that this double differencing in the latter is implemented only across treatment observations that are similar to each other, as they have been matched one-to-one in the first step.

Figure 3: Visual representation of Strategy B for PSM with DID



5 Results

We illustrate the estimation methodology outlined in the previous section by applying it to estimate the impact of EQUIP-T on four pupil learning outcomes. Pupil learning is assessed based on early grade reading and Mathematics tests administered to Grade 3 pupils. The raw pupil scores are analysed using the Rasch model of item response to produce estimates of pupil performance and item difficulty on a common interval scale. The pupils are then classified into curriculum-linked performance bands based on their performance, which constitute the final impact indicators:

- Proportion of pupils in the bottom performance band of the interval scale for Mathematics;
- Proportion of pupils in the top performance band of the interval scale for Mathematics;
- Proportion of pupils in the bottom performance band of the interval scale for Kiswahili; and
- Proportion of pupils in the top performance band of the interval scale for Kiswahili.

5.1 Presentation of results

For each outcome variable, three sets of results are presented in this paper: (a) the second stage results; (b) the propensity score matched outcomes at baseline and midline; and (c) the PSM-DID estimates. The following paragraphs use the example of Figure 4 to explain the interpretation of results in detail.

First, the second stage results for Strategy A are presented, as illustrated in Figure 4 for the indicator on the top performance band for Mathematics. The figure is divided into two panels: the top panel shows the baseline results and the bottom panel the midline results. The format for each panel is as follows:

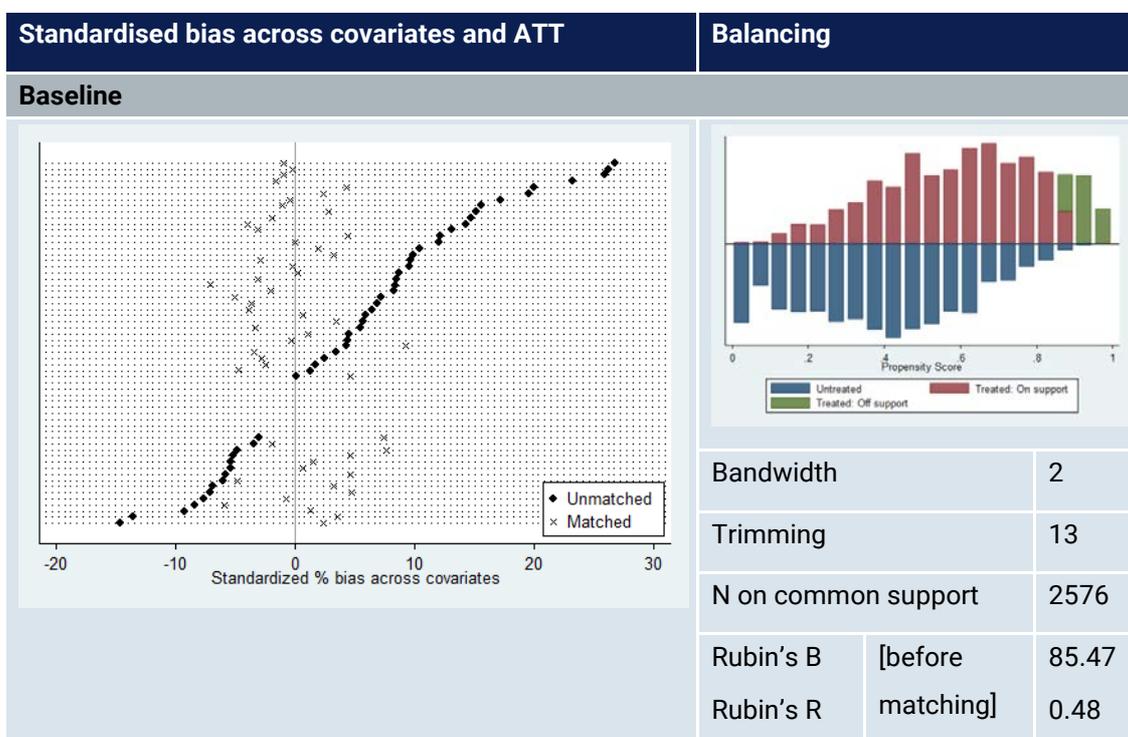
- The first graph on the left-hand side indicates how individual variables balance before and after matching. The x-axis displays the standardised bias, which is the percentage difference of the sample means in the treated and non-treated (unmatched or matched) subsamples as a percentage of the square root of the average of the sample variances in the treated and non-treated groups (Rosenbaum and Rubin, 1985). In Figure 4 below, for example, the unmatched samples display large imbalances with standardised bias being present across many of the covariates of interest. However, once matching takes place, the standardised imbalances are diminished. We present distinct graphs for each outcome as the selection models differ across outcomes due to the data-driven selection process.
- The second graph, on the right-hand side, shows the distribution of propensity scores across treatment and control groups. This graph visually confirms that, after dropping observations that are off common support, both treatment and control groups contain observations with propensity scores across the full range of the

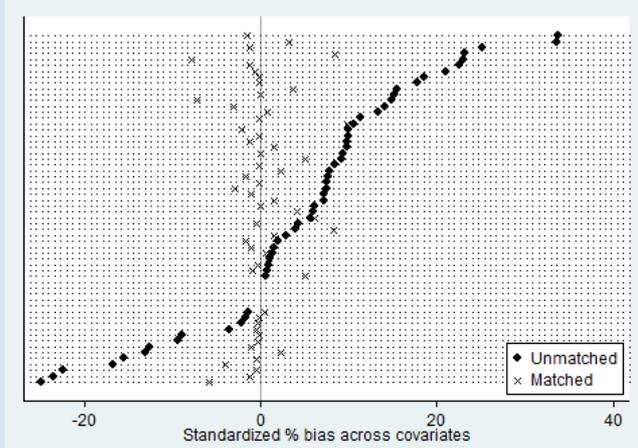
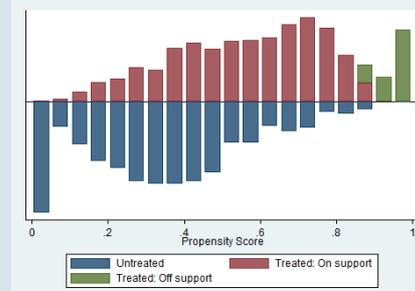
distribution. This is an indication of overall balance. Although the distributions of propensity scores across treatment and control groups would ideally be symmetric, the presence of some level of skewness does not put at risk the estimation procedure, as indicated by the balance achieved for each covariate and the overall values of Rubin’s R and B after matching.

- The remaining rows on the right-hand side display information related to the PSM model. The bandwidth and level of trimming for the optimal PSM model can be found in the first two rows. For example, the optimal model has a bandwidth of 2 and a trimming value of 13 for the baseline sample in Figure 4. This is then followed by the number of observations on common support in the next row, and then the Rubin’s R and Rubin’s B values both before and after matching. Generally, a Rubin’s B score under 25 after matching is desirable, while a Rubin’s R score between 1 and 1.25 is the preferred range after matching (Rubin, 2001). The unmatched samples are particularly unbalanced; for instance, the Rubin’s B for the baseline sample and the midline sample is 85.47 and 70.87 respectively. However, the Rubin’s B scores after matching, which are all below 25, show how matching removes the previous imbalances.
- Finally, the remaining rows on the right-hand side indicate the ATT for each corresponding survey wave and the associated standard errors. Both bootstrapped and non-bootstrapped standard errors are presented for robustness purposes.

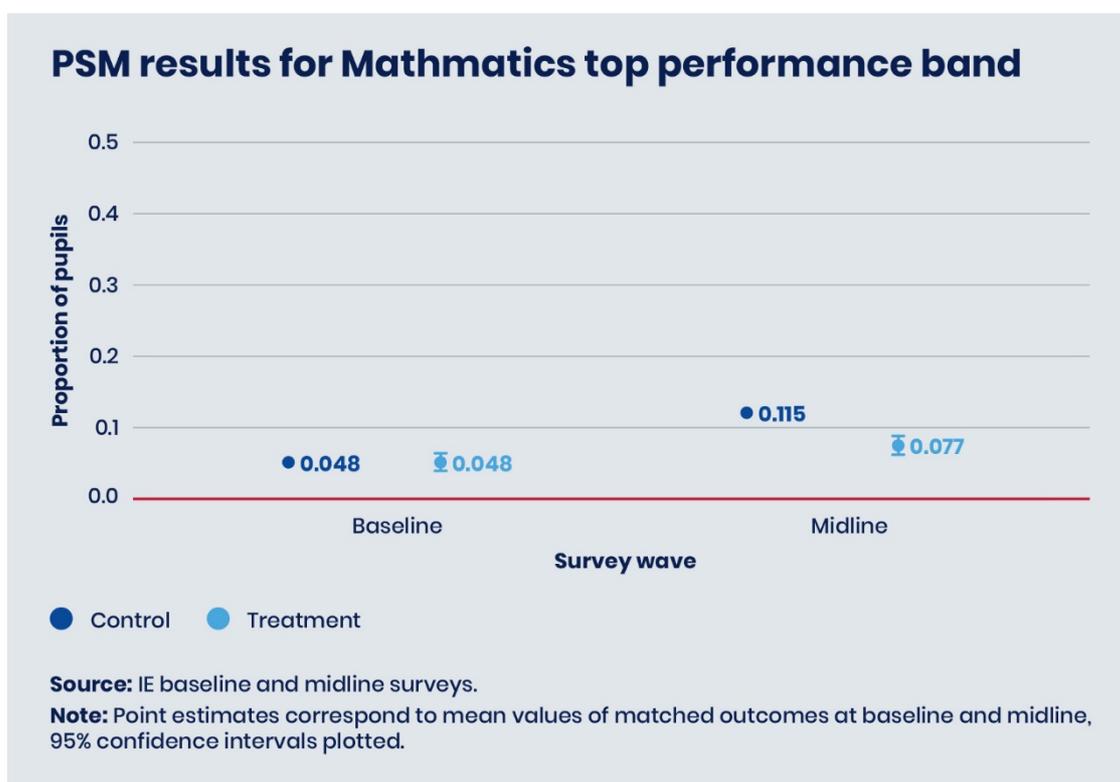
Proportion of pupils in the top performance band for Mathematics

Figure 4: Mathematics top band: Second stage results (Strategy A)



ATT	0.00	Rubin's B	[after	25.51
SE (bootstrapping)	(0.011)	Rubin's R	matching]	1.09
SE (no bootstrapping)	(.01)			
Midline				
				
		Bandwidth		4
		Trimming		13
		N on common support		2505
		Rubin's B	[before	70.87
		Rubin's R	matching]	1.09
ATT	-0.03	Rubin's B	[after	22.31
SE (bootstrapping)	(0.018)	Rubin's R	matching]	0.94
SE (no bootstrapping)	(0.014)			

Second, the mean values of the matched outcome and associated confidence intervals at baseline and midline for the treatment group and the control group are plotted. An example can be seen in Figure 5 for the top performance band in Mathematics. For the treatment group, the mean of the outcome variable is plotted for observations on common support. For the control group, the mean of the counterfactual outcome estimated by the matching algorithm is plotted here.

Figure 5: Mathematics top band: Matched outcome at baseline and midline

Finally, the PSM-DID estimate for both Strategy A and Strategy B are presented, along with the associated bootstrapped and non-bootstrapped p-values. Table 1 provides an example of how the overall impact result should be interpreted across the two strategies. In that table, the PSM-DID estimate from Strategy B shows a statistically significant negative trend in EQUIP-T schools, although this finding is not confirmed by Strategy A, which fails to detect a similarly significant negative trend.

Table 1: Mathematics top band: PSM-DID estimate

	Strategy A	Strategy B
PSM-DID estimate	-0.03	-0.04
P-value (bootstrapping)	(0.13)	(0.003)
P-value (no bootstrapping)	(0.07)	(0.003)

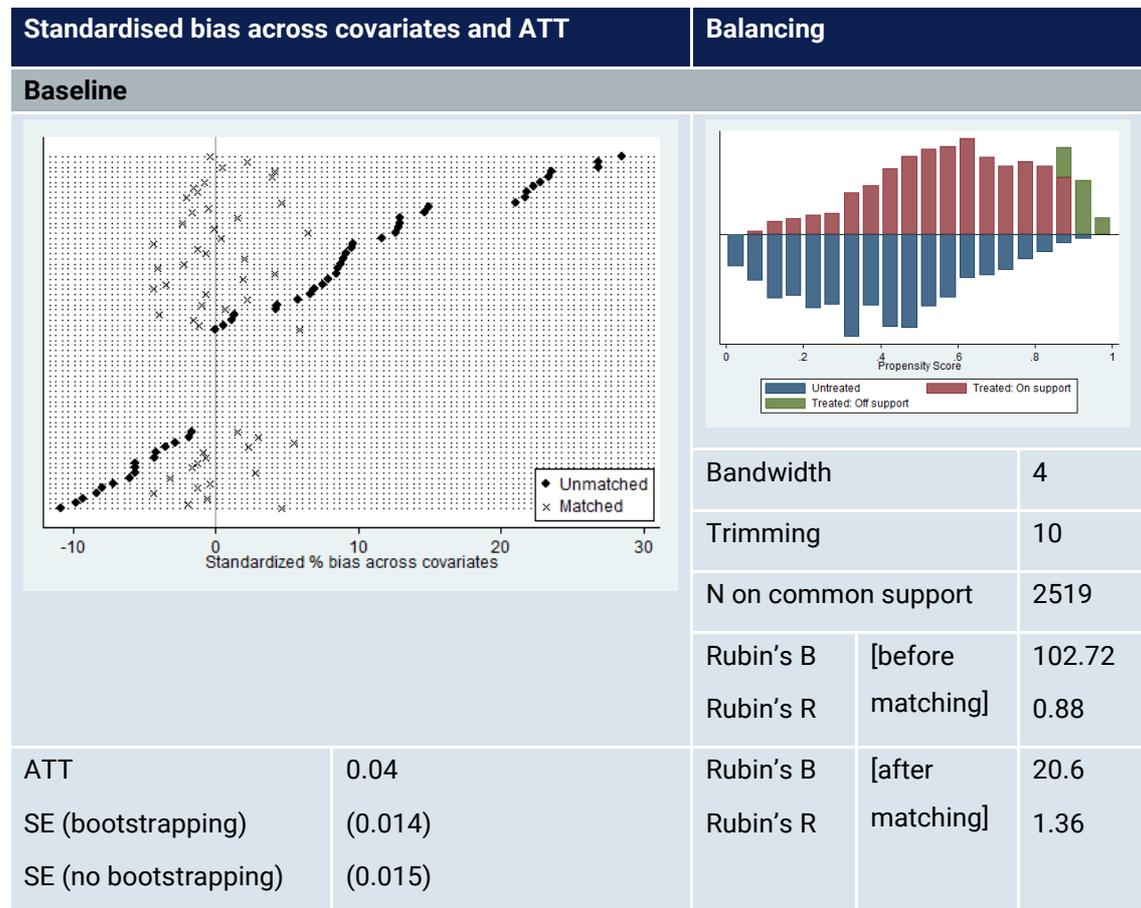
The balancing results for Strategy B – where treatment observations across the two survey waves are matched – are also summarised at the end for each outcome indicator, as illustrated in Table 2. This table shows that the balancing properties for this matching process concerning this particular indicator were not ideal – note that Rubin’s R is above 25. Although this strategy does not confirm the finding from Strategy A, this cannot lead us to change our overall conclusion that EQUIP-T did not have a significant impact on this outcome indicator.

Table 2: Mathematics top band: Balancing results (Strategy B)

Balancing results from matching treatment observations across baseline and midline			
Caliper			.4
N for common support			1586
Rubin's B	[before		89.31
Rubin's R	matching]		1.27
Rubin's B	[after matching]		28.4
Rubin's R			0.9

Proportion of pupils in the bottom performance band for Mathematics

Figure 6: Mathematics bottom band: Second stage results (Strategy A)



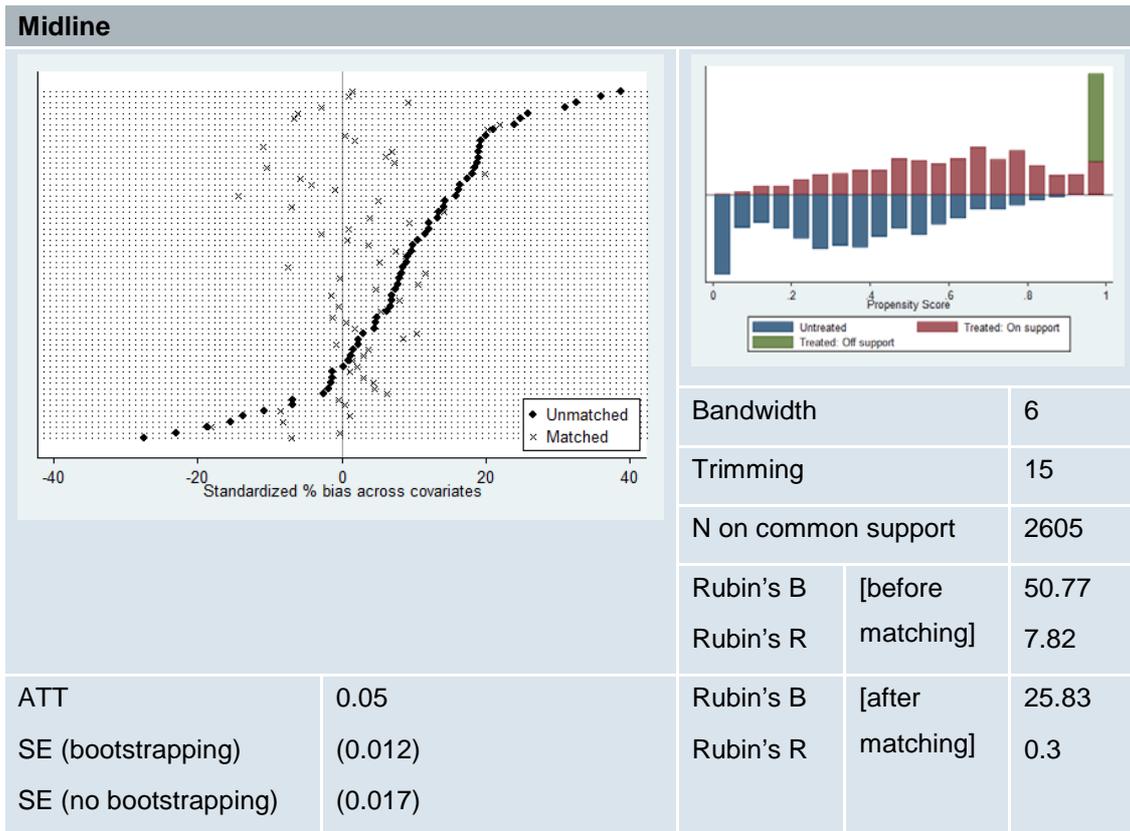


Figure 7: Mathematics bottom band: Matched outcome at baseline and midline

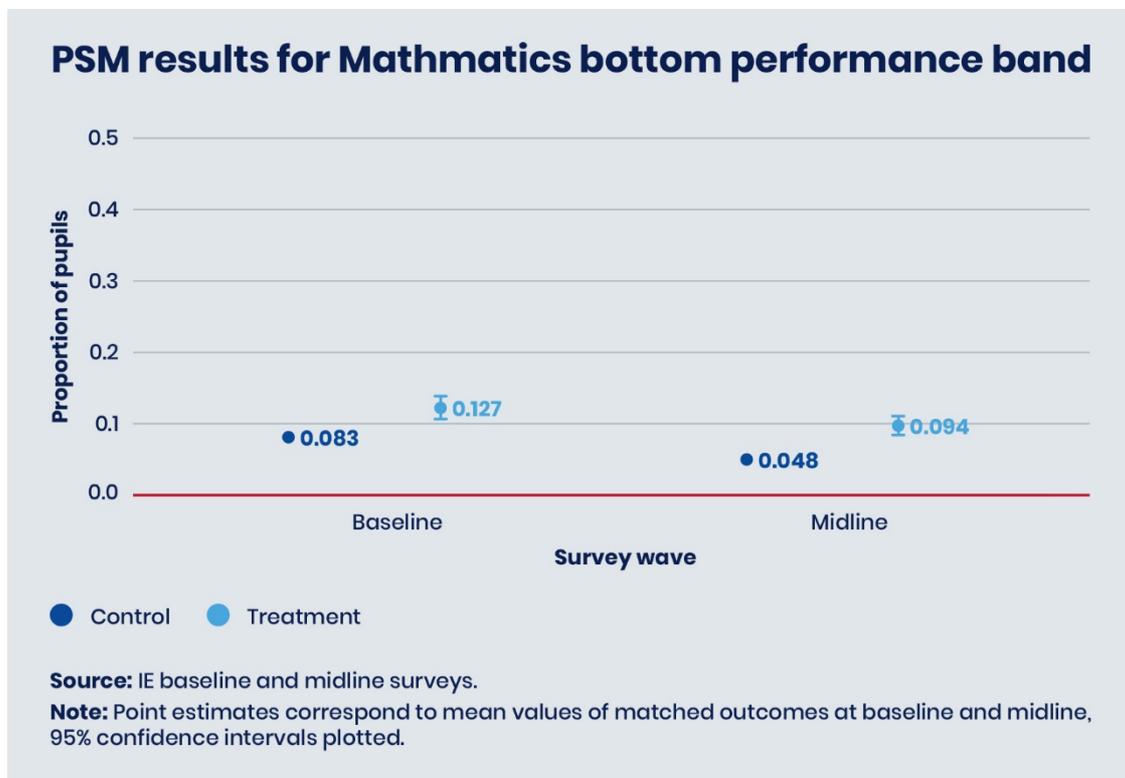


Figure 7 above shows that the PSM estimates point to an overall decrease in the proportion of pupils in the bottom performance band for Mathematics, but without much difference in this trend across treatment and comparison schools. As can be seen in Table 3 below, this means that the study does not find any evidence of a statistically significant impact of EQUIP-T on the proportion of pupils in the bottom performance band for Mathematics. The two strategies are consistent with each other in regard to this assessment.

Table 3: Mathematics bottom band: PSM-DID estimate

	Strategy A	Strategy B
PSM-DID estimate	0.002	-0.001
P-value (bootstrapping)	(0.92)	(0.50)
P-value (no bootstrapping)	(0.93)	(0.50)

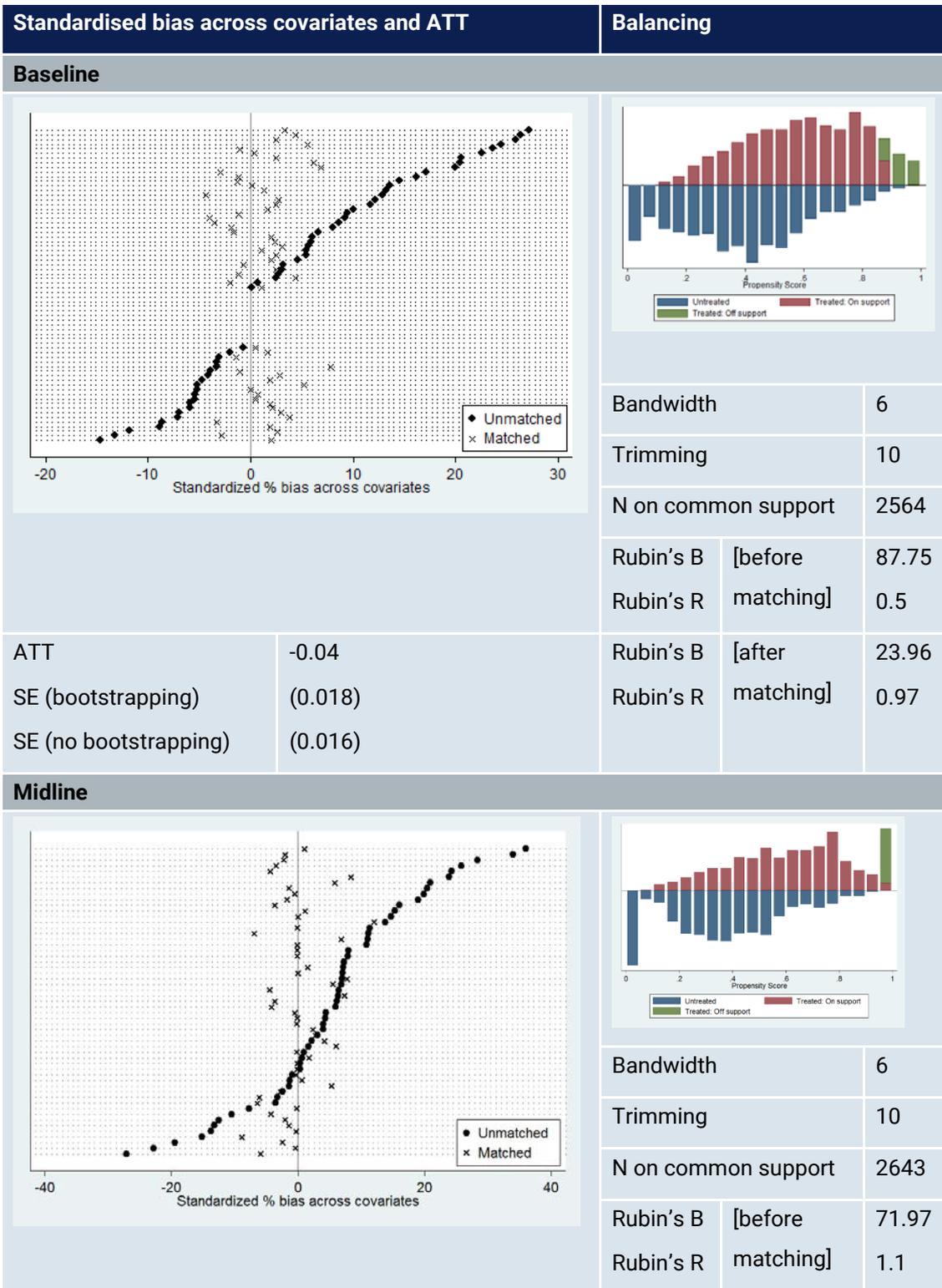
Table 4 below presents results on the balancing properties of Strategy B. As can be seen in the 'after matching' row, balancing is not ideal for treatment observations across time.

Table 4: Mathematics bottom band: Balancing results (Strategy B)

Balancing results from matching treatment observations across baseline and midline		
	Caliper	0.4
	N for common support	1844
Rubin's B	[before	85.89
Rubin's R	matching]	1.04
Rubin's B	[after matching]	28.09
Rubin's R		0.81

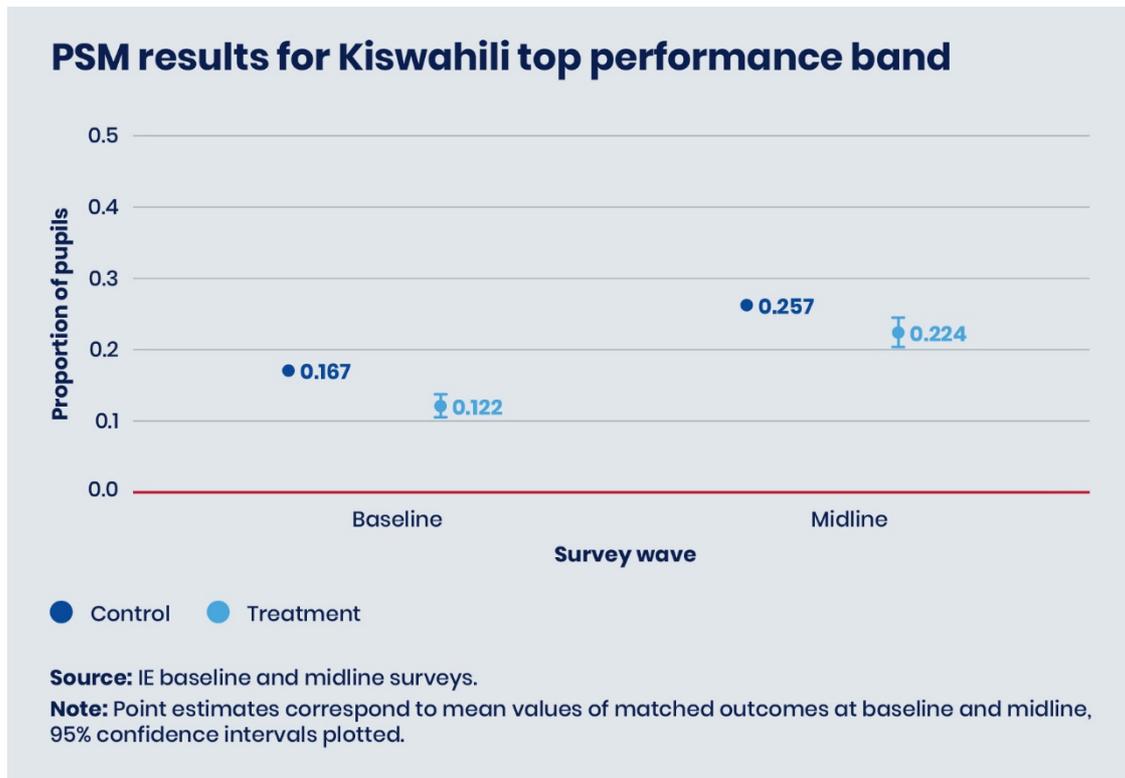
Proportion of pupils in the top performance band for Kiswahili

Figure 8: Kiswahili top band: Second stage results (Strategy A)



ATT	-0.03	Rubin's B	[after	28.83
SE (bootstrapping)	(0.023)	Rubin's R	matching]	1.47
SE (no bootstrapping)	(0.021)			

Figure 9: Kiswahili top band: Matched outcomes at baseline and midline



While both strategies show a positive change in the proportion of pupils in the top performance band for Kiswahili, this result is not statistically significant and, therefore, the analysis is unable to provide a positive assessment on the impact of EQUIP-T on this indicator.

Table 5: Kiswahili top band: PSM-DID estimate

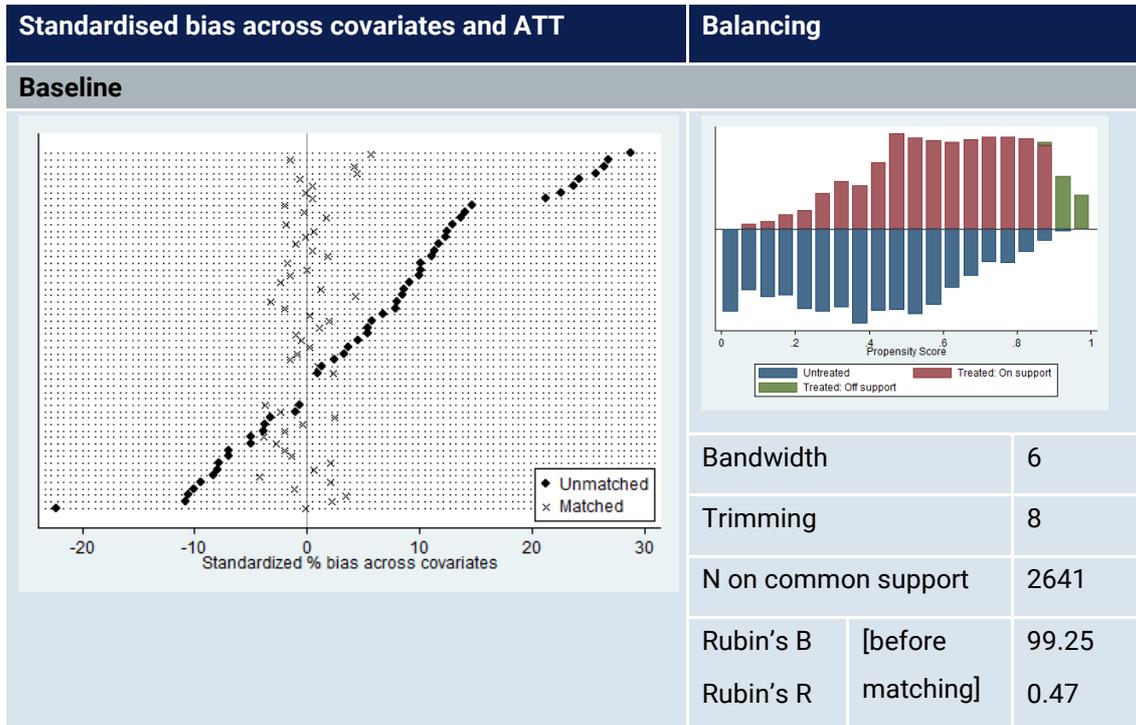
	Strategy A	Strategy B
PSM-DID estimate	0.012	0.02
P-value (bootstrapping)	(0.67)	(0.28)
P-value (no bootstrapping)	(0.68)	(0.28)

Table 6: Kiswahili top band: Balancing results (Strategy B)

Balancing results from matching treatment observations across baseline and midline			
Caliper			0.3
N for common support			1630
Rubin's B	[before		88.5
Rubin's R	matching]		1.18
Rubin's B	[after matching]		24.63
Rubin's R			0.43

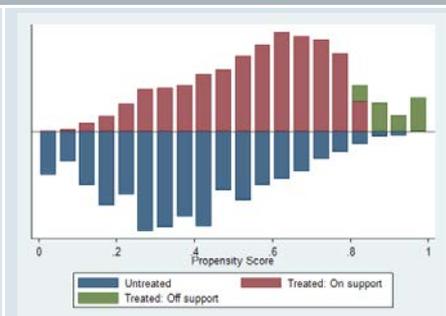
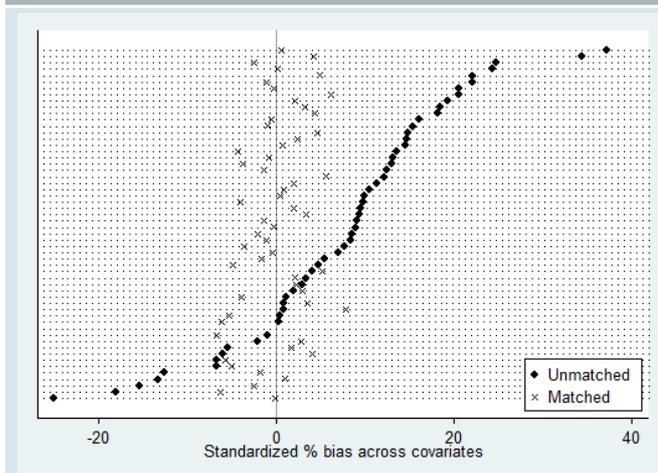
Proportion of pupils in the bottom performance band for Kiswahili

Figure 10: Kiswahili bottom band: Second stage results (Strategy A)



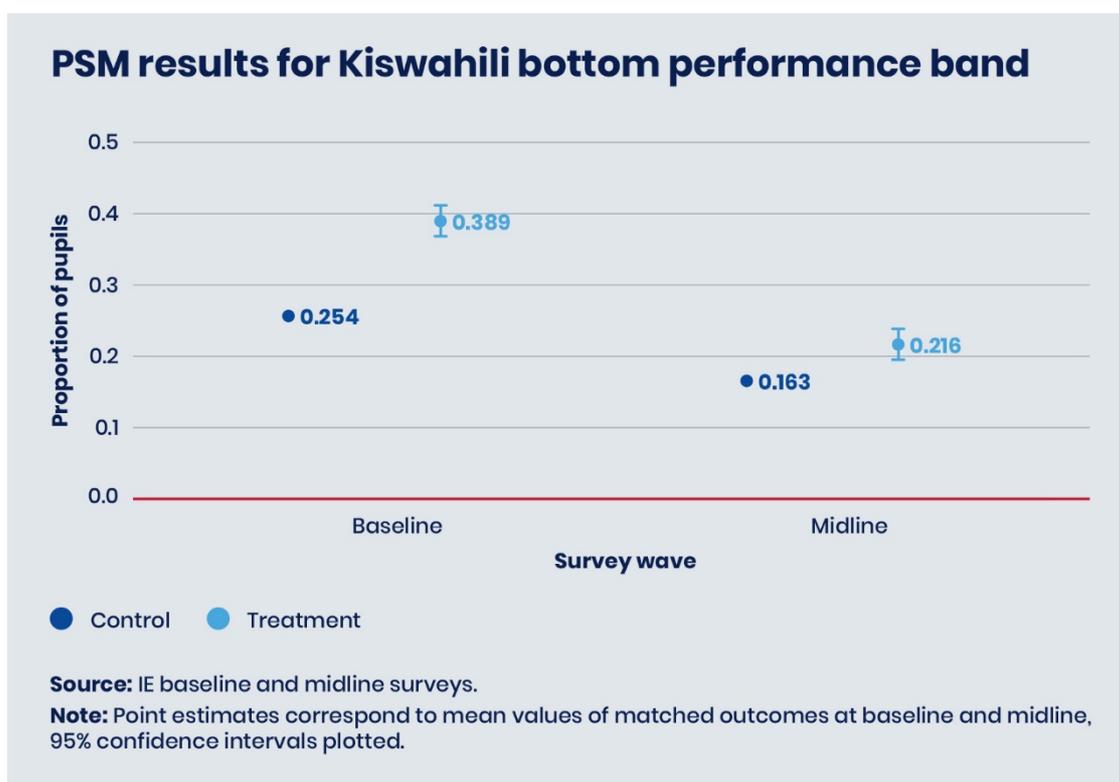
ATT	0.13	Rubin's B	[after	18.98
SE (bootstrapping)	(0.02)	Rubin's R	matching]	1.5
SE (no bootstrapping)	(0.02)			

Midline



ATT	0.05	Rubin's B	[after	26.83
SE (bootstrapping)	(0.016)	Rubin's R	matching]	1.33
SE (no bootstrapping)	(0.018)			

Bandwidth	6	
Trimming	10	
N on common support	2577	
Rubin's B	[before	68.57
Rubin's R	matching]	0.81

Figure 11: Kiswahili bottom band: Matched outcomes at baseline and midline

As seen in Figure 11, PSM analyses at baseline and midline point to a decreasing gap between treatment and comparison schools in terms of pupils who are in the bottom performance band for Kiswahili. This means that the overall PSM-DID analysis finds strong evidence that EQUIP-T has reduced the proportion of pupils in the bottom performance band for Kiswahili in programme schools (see Table 7 below). These results remain strong and highly significant across both Strategy A and Strategy B.

Table 7: Kiswahili bottom band: PSM-DID estimate

	Strategy A	Strategy B
PSM-DID estimate	-0.08	-0.07
P-value (bootstrapping)	(0.00)	(.00)
P-value (no bootstrapping)	(0.00)	(0.001)

The balancing results for Strategy B across time for treatment observations, presented below, show that for this outcome indicator the balancing after matching is within acceptable ranges. This further strengthens the findings presented above suggesting that EQUIP-T has significantly reduced the proportion of children in the bottom performance band for Kiswahili in treatment schools, compared to a counterfactual situation without EQUIP-T.

Table 8: Kiswahili bottom band: Balancing results (Strategy B)

Balancing results from matching treatment observations across baseline and midline		
Caliper		0.4
N for common support		1798
Rubin's B	[before matching]	91.49
Rubin's R		1.18
Rubin's B	[after matching]	24.99
Rubin's R		1.01

6 Limitations

Four key caveats related to the present estimation strategy need to be mentioned here. First, PSM only controls for observable characteristics that cause selection bias. This is a problem for any impact identification strategy that relies on controlling only for factors (variables) that can be observed in the data – not only PSM. PSM helps address this by allowing for extensive balancing checks after matching, which can provide substantial evidence for the fact that balance is achieved across a large variety of characteristics and – by implication – is also likely to extend to unobservables. In this study, such extensive balancing checks were implemented. In addition, as explained above, the DID strategy implemented in the present case helps to control for remaining imbalances that may be due to time-invariant unobservable variables.

Second, DID helps to deal with time-invariant imbalances but not time-variant ones. This means that only time-invariant imbalances that remain after PSM would be controlled for, in contrast to imbalances that vary over time. In the present case, this is addressed by extensive balancing tests, which show little remaining covariate imbalance in general after PSM, by showing that results are robust to a variety of different PSM specifications, and by showing that results are robust to two separate DID strategies. Together, this evidence suggests the results are robust, remaining imbalances are small, and results are unlikely to be sensitive to or to be driven by such imbalances – even if they were time variant.

Finally, calculating standard errors of estimated treatment effects using PSM methods is not straightforward. As Caliendo and Kopeinig (2005, p. 18) put it, ‘The problem is that the estimated variance of the treatment effect should also include the variance due to the estimation of the propensity score, the imputation of the common support, and possibly also the order in which treated individuals are matched’. These estimations increase the variation of the treatment effect estimates over and above normal sampling variation. There is no consensus in the literature on how to take this into account.

A popular approach to solve this problem is to bootstrap standard errors for the estimated treatment effect. Each bootstrap draw re-estimates both the first and second stages of the estimation. This produces N bootstrap samples for which the ATT is estimated. The distribution of these means approximates the true sampling distribution, and therefore the standard errors of the population mean (Caliendo and Kopeinig, 2005, p.18). Following this approach, we implemented bootstrapping using 200 repetitions to estimate the standard errors of the estimated treatment effects.

7 Conclusion

In contexts where an RCT is not possible or is not appropriate, alternative approaches are necessary to identify impact. PSM tackles the problem of selection bias by using data from a control group to construct appropriate comparisons to pupils or teachers in the treatment group, thus building a valid counterfactual. However, even after implementing a matching procedure, some imbalances across treatment and control groups can remain, which potentially could invalidate an impact identification strategy unless further analysis is implemented. In this paper we have demonstrated an innovative approach that builds a more efficient and unbiased PSM model and then combines PSM with DID analysis through two different techniques to control for time-invariant imbalances by comparing data from treatment and control schools at both baseline and midline.

In the first approach, the ATT was compared across time, between baseline and midline. In the second, PSM was used to match treatment units over time to construct a pseudo panel from repeated cross-sections to estimate overall ATT. In the absence of panel data, the conventional PSM approach of matching individuals at baseline and then calculating impact at endline is not possible. The innovative pseudo panel approach addresses this, following a suggestion by Blundell and Costa Dias (2000, p. 451).

The PSM approach augmented by DID was applied to the evaluation of an education programme in Tanzania. This study is the first practical application of this PSM with DID procedure for a repeated cross-section in an education-related evaluation of teachers and pupils. It is also one of the few studies that presents in detail the model implementation, including an innovative approach to mixing PSM with DID.

References

- Aerts, K. and Schmidt, T. (2008). 'Two for the price of one?' *Research Policy* 37, 806–822.
- Blundell, R. and Costa Dias, M. (2000) 'Evaluation Methods for Non-Experimental Data'. *Fiscal Studies* 21, no. 4: 427–68.
- Blundell, R. and Costa Dias, M. (2009). 'Alternative Approaches to Evaluation in Empirical Microeconomics'. *Journal of Human Resources*. 44, 565–640.
- Bönke, T., Schröder, C. and Jochimsen, B. (2013). Fiscal Federalism and Tax Administration – Evidence from Germany. DIW Berl.
- Cambridge Education (2014) Final EQUIP-T Inception Report, 5 February.
- Caliendo, M. and Kopeinig, S. (2008). 'Some practical guidance for the implementation of propensity score matching'. *Journal of Economic Surveys*, 22(1), 31-72.
- Imbens, G. and Rubin, D. (2015) Causal Inference for Statistics, Social, and Biomedical Sciences, Cambridge University Press.
- Hashim, N. and Strong, N. (2015). Do formal risk assessments improve analysts' target price accuracy?
- Heckman, J.J., Ichimura, H., and Todd, P.E. (1997). 'Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme'. *Review of Economic Studies* 64, 605–654.
- Hong, S.-H. (2013). 'Measuring the Effect of Napster on Recorded Music Sales: Difference-in-Differences Estimates Under Compositional Changes'. *Journal of Applied Economics*. 28, 297–324.
- OPM (2015a) *EQUIP-Tanzania Impact Evaluation Final Baseline Technical Report, Volume I: Results and Discussion*. Oxford Policy Management. Available at: www.opml.co.uk/sites/default/files/OPM%20IE%20Final%20Baseline%20Report%20Volume%20I.pdf
- OPM (2015b) *EQUIP-Tanzania Impact Evaluation Final Baseline Technical Report, Volume II: Methods and Technical Annexes*. Oxford Policy Management. Available at: www.opml.co.uk/sites/default/files/OPM%20IE%20Final%20Baseline%20Report%20Volume%20II.pdf
- OPM (2016) *EQUIP-Tanzania Impact Evaluation: Midline Planning Report*. Oxford Policy Management.
- Ordine, P. and Rose, G. (2016). Two-tier labor market reform and entry wage of protected workers: evidence from Italy | SpringerLink. *Empir. Econ.*
- Paternoster, Raymond, et al. "Using the correct statistical test for the equality of regression coefficients." *Criminology* 36.4 (1998): 859-866.
- Rosenbaum, Paul R., and Donald B. Rubin. "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score." *The American Statistician* 39.1 (1985): 33-38.

Rubin, D. (2001) 'Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation'. *Health Services & Outcomes Research Methodology* 2, 169–188.

Smith, J. and Todd, P. (2005). 'Does matching overcome LaLonde's critique of nonexperimental estimators?' *Journal of Economics*. 125, 305–353.

About Oxford Policy Management

Oxford Policy Management is committed to helping low- and middle- income countries achieve growth and reduce poverty and disadvantage through public policy reform.

We seek to bring about lasting positive change using analytical and practical policy expertise. Through our global network of offices, we work in partnership with national decision makers to research, design, implement, and evaluate impactful public policy.

We work in all areas of social and economic policy and governance, including health, finance, education, climate change, and public sector management. We draw on our local and international sector experts to provide the very best evidence-based support.

Find out more

For further information

visit: www.opml.co.uk

Or email: admin@opml.co.uk

    OPMglobal



Oxford Policy Management Limited

Registered in England: 3122495

Registered office: Clarendon House,

Level 3, 52 Cornmarket Street,

Oxford, OX1 3HJ, United Kingdom