EVALUATION PROTOCOL – APPENDIX A

Evaluation of the 'SAFE' (Support, Attend, Fulfil, Exceed) Taskforces

RAND Europe, University of Westminster, FFT Education Datalab

Principal investigator: Dr Ana FitzSimons

YOUTH ENDOWMENT FUND

Evaluation of the 'SAFE' (Support, Attend, Fulfil, Exceed) Taskforces

Appendix A: Preliminary analysis to inform the Study Protocol

Evaluating institution: RAND Europe, University of Westminster, FFT Education Datalab

Principal investigator: Dr Emma Disley

Table of contents

Table of contents 1

Preamble 2

Introduction 2

Potential comparison groups 3

Descriptive Statistics 5

Difference-in-differences 6

Synthetic control 10

Regression Discontinuity 13

Conclusions 16

Implications 18

References 18

Preamble

FFT Education Datalab and University of Westminster undertook this preliminary analysis to inform the study protocol of a quasi-experimental impact evaluation of the SAFE Taskforces programme. The original plan had been to evaluate the impact of the SAFE Taskforces programme as a whole, with a quasi-experimental design using LA-level data for serious violence outcomes (serious violence offences and hospital admissions for serious violence) and educational outcomes (school attendance, suspensions and exclusions). The findings of the preliminary analysis, set out below, led to a reconsideration of the impact evaluation design, which is set out in full in the study protocol.

Introduction

Briefly, we examine and compare pre-existing trends in participating Taskforce areas and local authorities (LAs) not participating in the intervention using Open Police Data[1] and administrative data from the National Pupil Database (NPD) for the years 2014 to 2021, some data limitations (described below) notwithstanding.[2]

The purpose of the analysis is to:

• Determine suitable outcome measures for the evaluation.

• Set up data processing steps to create analytical datasets.

• Test different methodological options.

• Assess pre-existing trends in outcomes.

• Calculate approximate minimum detectable effect sizes.

• Help make decisions about the methodology we will pursue when undertaking the impact evaluation of the programme following the 2022/23 academic year.

We test three different approaches for analysing the impact of the programme at Taskforce level:

- Difference in differences

[1] https://data.police.uk/

[2] NHS Digital Data on hospital admissions for assault with sharp object and A&E attendances, which we planned to analyse in the evaluation report, were not made available in the SRS by the NHS by the time we drafted this document.

- Synthetic control
- Regression Discontinuity

Data tables supporting this Appendix A may be found in the Addendum (attached Excel file). These tables comprise:

- Appendix A, Table 1 Pre-treatment trends in serious violence.
- Appendix A, Table 2 Pre-treatment trends in educational outcomes (permanent exclusions, suspensions, absence) for 10-13 year olds.
- Appendix A, Table 3 Pre-treatment trends in post-16 participation (in school at age 16, in school or employment at age 16) for 13-year-olds.
- Appendix A, Figures 1-8 Pre-treatment trends in outcomes.
- Appendix A, Table 4 Pre-treatment trends in outcomes for Taskforces and different control groups, with and without controls.
- Appendix A, Table 5 Estimated Treatment Effects (Placebo tests) for primary and secondary outcomes.
- Appendix A, Table 6 Selection of synthetic control group by Root Mean Squared Prediction Error.
- Appendix A, Table 7 Balance of predictors between Taskforce areas and synthetic control group.
- Appendix A, Table 8 Weights used for synthetic control groups.
- Appendix A, Figures 9-14 Pre-treatment trends in outcomes for Taskforce and synthetic control group.
- Appendix A, Figure 15 Distribution of the serious violence and hospital admission score.
- Appendix A, Figures 16-23 Outcomes and mean, linear and quadratic fits above and below cut-off, by bandwidth and cohort.
- Appendix A, Figure 24 Covariates and mean, linear and quadratic fits above and below cut-off, by cohort.

Potential comparison groups

The methodology for choosing SAFE areas used data on two metrics for serious violence:

a) Annual hospital admissions data for assault with a sharp object.

b) Monthly average offence data for offences that fall under the serious violence definition.

Each LA was ranked separately according to each metric, over a period of 5 years. These rankings were then combined into an overall ranking. The top 10 locations were chosen from this list.

To select a suitable comparison group, we can exploit the delivery of the APST (Alternative Provision Specialist Taskforces) intervention, piloted by the DfE between November 2021 and March 2025. The programme aimed to embed teams of specialists in 22 alternative provision (AP) schools in serious violence (SV) hotspots across England. Local authorities participating in APST and SAFE taskforces were selected on the basis of the same metrics as SAFE, but rankings were calculated over a different time period: a 1-year period for APST (hospital admissions between April-September 2020 and serious violence offences monthly average for 2019), rather than a 5-year period for SAFE (from 2016/17 – 2020/21).

Table 1 below sets out a list of the LAs in which the two programmes are implemented. APST areas in which SAFE is also implemented are in bold.

Table 1: SAFE and APST Local Authorities

SAFE Taskforce Areas	APST Areas
1. Birmingham	**1. Birmingham**
2. Manchester	**2. Manchester**
3. Leeds	**3. Leeds**
4. Sheffield	**4. Sheffield**
5. Liverpool	**5. Liverpool**
6. Newham	**6. Newham**
7. Lambeth	**7. Southwark**
8. Southwark	8. Bristol
9. Bradford	9. Brent
10. Haringey	10. Leicester
	11. Bradford
	12. Salford
	13. Lambeth
	14. Hackney
	15. Croydon
	16. Enfield
	17. Tower Hamlets
	18. Haringey

	19. Doncaster 20. Nottingham 21. Sandwell 22. Ealing

Three different comparison groups are considered:

1. All other local authorities non-participating in SAFE (142 LAs)[3]. This comparison group includes LAs involved in neither SAFE nor APST and LAs involved in APST only.
2. The 12 local authorities of the 22 local authority areas with the highest levels of serious violence participating in APST but not in SAFE.
3. All other local authorities, excluding the 12 LAs participating in APST only (130 LAs).

The estimated impacts will have a different interpretation in each case:

With comparison group 1, the estimate will be an effect that is partly relative to APST, though the extent of this is hard to interpret as the APST impact cannot be singled out. With comparison group 2, the estimate will be an effect relative to APST. With comparison group 3, the estimate will not be a relative effect.

Descriptive Statistics

Summary data for each outcome for participating SAFE Local Authorities and comparison groups of LAs non-participating in SAFE can be found in Appendix A, Tables 1, 2 and 3 for the three different comparison groups (all non-treated LAs, APST LAs, and non-APST LAs).

The outcomes of interest are:

- Serious violence offences, measured in levels and yearly change.
- Permanent exclusions and suspensions
- Attendance, measured as the fraction of sessions missed due to overall absence, due to unauthorised absence and due to persistent absences.

[3] In practice, a small number of non-participating LAs are dropped from the analysis due to missing/ incomplete data.

- Post-16 outcomes, measured as being enrolled at a state-funded school or on a learning aim in a college; and being enrolled at a state-funded school, on a learning aim in a college or employed.

Serious violence relates to all ages for both the perpetrator and the victim as the data are not disaggregated by age group. Educational outcomes (permanent exclusions, suspensions, absence) refer to 10–13-year-olds,[4] which is the target population for the intervention; post-16 participation outcomes (in school at age 16; in school or employment at age 16) relates to 13 year olds (but are measured when they are 16). The tables also include the numbers of LAs in each group and a Figure showing the pre-trends across the four different groups: treated and the three comparison groups. These can be found in Appendix A, Tables 1, 2 and 3 and Figures 1-8 respectively. For serious violence, measured in level (Table 1), rate of suspension (Table 2), % of 13-year-olds in sustained education at age 16 and in sustained education or employment at age 16 (Table 3), the comparison group composed of APST LAs approximates the SAFE group most closely, as the pre-intervention trends in some cases look parallel between treated and this comparison group. In the other cases, no parallel pattern emerges. These are just visual representations rather than a formal assessment of whether trends are parallel.

Difference-in-differences

For each outcome, we fit DiD models using pre-treatment data, to understand whether DiD would provide an appropriate statistical approach to estimating the impact of the SAFE programme. This has a threefold purpose:

- To test whether pre-treatment trends in SAFE LAs and the comparison LAs are parallel.
- To run placebo tests (to confirm our expectation that we should not find any significant difference for a fake treatment year).
- To understand the likely size(s) of confidence intervals with which treatment effects will be estimated ("power calculations").

For each outcome, we estimate a model comparing treated areas with the three alternative comparison groups. For each outcome, we calculate models with and without controls. This means there are six specifications for each outcome. The models are fitted in Stata. Note that

[4] At the time this preliminary analysis was undertaken it was assumed that some SAFE Taskforce's 'transition' work with young people entering into Yr7 was done with Yr 6 young people (age 10). However, subsequent discussions clarified that such transition work was only done during the summer holidays prior to entering Yr7. In the Study Protocol the age group of Yr7-9 is therefore used throughout as they will only be captured as Yr7s in the evaluation.

the preliminary regressions for the Taskforce level analyses use local authorities as units of analysis, not pupils, hence residuals are not clustered. All models are estimated through OLS.

The number of LAs included in the SAFE Taskforces treatment group, and in the three non-SAFE comparison groups, are reported in Tables 1, 2 and 3.

The controls for the Taskforce level analysis are the following:

- Fraction of pupils eligible for free school meals.
- % of pupils in state-funded education aged 10-13 with SEN met by a statement or EHC plan.
- % of pupils in state-funded education aged 10-13 with SEN not met by a statement or EHC plan.
- Local Authority mean IDACI score, 2019.[5]
- Local Authority total population.

We used information from the placebo tests to calculate the minimum detectable effect size (MDES) for each outcome we plan to evaluate on the assumption that the post-treatment data behave in a similar fashion. We calculate the minimum detectable effect (MDE) as a multiple, M, of the robust standard error, s. Here, M is calculated as the 0.025 upper tail critical value of the t distribution (reflecting the 95% two-tail significance level) plus the 0.8 upper tail critical value of the t distribution (reflecting 80% power). Although the precise value of M depends on the degrees of freedom, $M \approx 2.8$. The MDES is obtained by dividing the MDE by the standard deviation:

$$\text{MDES} = \frac{s*M}{\sigma}$$

Parallel trends

To ensure that the comparison group provides a plausible counterfactual to the treatment group, we explore whether there are similar pre-intervention trends in outcomes between the treatment and the comparison groups. We thus test differences in outcomes between treatment and comparison groups in all available pre-intervention years.

[5] 2019 is the latest available data.

The results contained in Table 4 of Appendix A show the parameter estimates of treated_area*year for each outcome each year (relative to the base year of 2014). The table shows 258 estimates:

- 3 outcomes x 6 specifications (i.e. treatment versus 3 comparison groups, each with and without controls) x 7 years (126 total)
- 3 outcomes x 6 specifications x 6 years: there are no data for the percentage of absences, unauthorized and persistent absences in 2020 (108 total)
- 2 outcomes x 6 specifications x 2 years: for post-16 outcomes, only data relative to 2018 and 2019 can be observed (24)

Of these 258 estimates, 23 reach statistical significance at the 5 % level. 13 of these statistically significant estimates occur for the serious violence outcome when measured as a continuous outcome. The remaining 10 are distributed across the other outcomes, suggesting that the proportion of significant estimates is about 4 per cent, roughly the percentage that would be expected based on chance alone. We note that they are concentrated in the year 2021 and in absences.

For serious violence, 13 estimates out of 42 are statistically significant when serious violence is measured as a continuous outcome, only 2 when it is measured as a change with respect to the previous year. This indicates that pre-trends were not parallel across any of the six specifications for the first measure, but roughly so for the second. The results are better (see Table 4), i.e. there are no significant coefficients for serious violence in level, for the comparison group composed of LAs with a similar level of serious violence as the SAFE Taskforces., i.e. the APST Taskforces areas, but only in the specification without controls. For serious violence measured in change, non-SAFE APST LAs deliver non-significant coefficients with and without controls. This suggests that pre-intervention trends in serious violence in non-SAFE APST LAs were similar to those in SAFE LAs.

Table 4 shows no particular concerns about the prior trends for the educational outcomes (permanent exclusions, suspensions and absences), suggesting that pre-intervention trends in educational outcomes in comparison LAs were similar to those in SAFE LAs. However, the degree of variation in the coefficients gives some grounds to worry about the power of this test. Only 8 of the 108 values of treated_school * year are significantly different from zero for the specifications using the three different comparison groups. All of the absences related results were significantly different in correspondence of the year 2021, suggesting that trends in absences in comparison LAs were different to those in SAFE LAs in 2021, possibly because Covid-related LA responses affected them.

For post-16 Participation (in school at age 16 and in school or employment at age 16) there are no concerns about the prior trends. None of the 24 values of treated_school * year

significantly differ from zero for the specifications using the three different comparison groups, indicating that pre-intervention trends in post-16 participation in comparison LAs were most similar to those in SAFE LAs.

As the results for the non-APST comparison group are second best to the one using non-SAFE APST LAs, we suggest performing a second DiD estimator using the non-APST comparison group. This would allow the impact of SAFE relative to no intervention to be estimated.

Placebo Tests

Results of placebo tests setting 2019 as a fake treatment year are shown in Appendix A, Table 5. We avoid using the years 2020 and 2021 so results are affected by COVID-related distortions (including missing outcomes). The regressions are all run at the local authority level, and they include all years for which data are available over the period 2014-2019. As the unit of analysis is the LA, standard errors are not clustered. Estimates for all nine outcomes fitted under OLS are not significant for any of the comparison groups.

The estimated effect sizes for the placebo tests suggest that pre-existing differences between treatment and comparison groups were substantial. The Minimum Detectable Effect Size (MDES) for the primary outcome, serious violence, is 0.28 at the level and 0.84 when measured in change. A smaller MDES indicates a higher ability to identify small effects. While a threshold of 0.2 is often employed in trial design, adjusting the MDES by involving more LAs is not feasible in this evaluation. Instead, the MDES serves as an indicator of result sensitivity and aids in selecting the most suitable analytical approaches. For all secondary outcomes, the MDES exceeds 0.2, although this estimate is conservative. The impact analysis will consider data from three post-intervention years, diverging from the assumed single post-intervention year in this preliminary analysis.

It has to be noted that there are limitations to our ability to test parallel pre-trends, because these tests may have low power in detecting statistically significant pre-trends, and sample bias in the treatment group can create selection bias from only analysing cases with insignificant pre-trends. Rambachan and Roth, (2023) propose a sensitivity test to present robust inference in settings where the parallel trends assumption may not hold. We suggest performing these tests if the difference in difference method is chosen for the Taskforce level impact evaluation.

In summary:

- For serious violence, the comparison group with the most similar pre-intervention trends to the treatment group is represented by non-SAFE APST LAs.

- For educational outcomes and post-16 outcomes, there were similar pre-intervention trends in the treatment group and the three comparison groups.
- Non-APST LAs can provide a second comparison group to test the robustness of the results and the impact of SAFE relative to no intervention.
- A placebo test using a DiD approach with all three comparison groups did not suggest a significant placebo effect.
- All the three comparison groups generate MDES larger than 0.2.

Synthetic control

The synthetic control method, pioneered to look at the effects of treatments or policies affecting large geographic areas (Abadie et al., 2003; Abadie et al., 2010; Abadie et al., 2015), seeks to construct a control group by weighting other aggregate units (here, local authorities) that best approximate the outcome for the treated group (here, the local authorities that comprise the SAFE Taskforce areas) over the pre-treatment years. The difference between the synthetic control and the treatment group in the post-treatment phase represents the estimated impacts. The limit of our analysis is that the number of pre-periods is quite short compared to standard synthetic control studies. This might make that approach less useful and appropriate.

As the synthetic control analysis uses aggregate LA-level units, there is little sample variation across the different units. This means that inference cannot be conducted in the usual fashion by calculating standard errors and confidence intervals. Hence, we cannot perform the placebo test we performed for the difference-in-differences model.

To test whether the synthetic control represents a valid counterfactual, placebo permutation tests can be performed, but only once observations in the post-treatment period are available. In addition, the method relies on a sufficiently long run of pre-intervention data points. This limits our possibility of assessing the validity of the synthetic control group as a valid counterfactual to the treatment group in the preliminary phase. However, it seemed a valid option to explore without an experimental design.

From a 'donor pool' of local authorities not affected by the programme, a synthetic control group is constructed as the weighted average of those units that best resemble the SAFE Taskforce areas in terms of pre-treatment characteristics and outcomes for a pre-treatment time period as long as possible. The limitation in our setting is that we can only explore this from 2014 and not for a longer time period before the intervention, as usually performed in this type of analysis.

We effectively treat the ten Taskforces as a single local authority. Thus, outcomes for the treatment group represent the average outcome across all ten SAFE Taskforce local authorities. Acemoglu et al., (2016) and Krief et al., (2016) use an alternative approach for cases with multiple treated units. From the donor pool of control units, they randomly simulate regions or groups of a similar size to the treatment group and then apply synthetic control methods to these simulated groups. They repeat this approach up to 5,000 times to obtain standard errors and confidence intervals. We did not follow this approach for several reasons. First, we are looking at the SAFE Taskforce programme as a whole, rather than at individual SAFE LAs, since the programme has natural policy relevance. Second, treating the Taskforces as a single treated unit maps onto the approach taken with DiD, where the impact is captured by one single treatment dummy for all Taskforce areas. Finally, the amount of repetition required by the alternative approach would have been highly computationally intensive and require a lengthy investment of time. We take all local authorities as our potential donor pool and find the set of weights across local authorities that best allows us to match pre-treatment trends in outcomes and other characteristics in the ten Taskforces areas.

We excluded the following local authorities. The Isles of Scilly and City of London are excluded as they are both very small and unusual relative to the rest of England. City of London is also excluded due to very different time trends. There are three LAs (Dorset, Poole and Bournemouth[6]) in the comparison sample for which data are missing that are dropped from the analysis.

The resultant weights may and indeed do differ across our main outcomes. The fact that the weights are different across different outcomes is not unexpected as it is likely that different underlying factors drive different outcomes and, as a result, different sets of local authorities are likely to serve as suitable controls for different outcomes.

We seek a balance on the observable characteristics that predict levels and trends in outcomes. We select these from the same list of control variables listed above (fraction of pupils eligible for free school meals; % of pupils in state-funded education aged 10-13 with SEN met by a statement or EHC plan; % of pupils in state-funded education aged 10-13 with SEN not met by a statement or EHC plan; local Authority mean IDACI score, 2019; local authority total population). We select lagged outcomes for selected years across different specifications.

[6] These LAs underwent re-organisation becoming LAs Dorset and "Bournemouth, Christchurch, and Poole". Although it would be feasible to recalculate measures derived from NPD, IDACI would remain missing.

Our main measure to assess the quality of the match is the root mean squared prediction error (RMSPE) between the Taskforces and the synthetic controls in the lagged outcome over the pre-treatment period.[7] The RMSPE is effectively a measure of how close our predictions come to capturing the true changes in the outcomes before the treatment.

Table 6 shows how we select our preferred donor pool and specification by calculating the RMSPE across a range of potential specifications, which differ in the set of lagged outcomes included.

For serious violence, we use the outcome as the change between two consecutive years, as the variable in level was more difficult to match to a comparison group. For serious violence and permanent exclusions, using specification (5) leads to the lowest RMSPE. For suspensions, overall absences and persistent absences, the lowest RMSPE comes from using specification (4). For unauthorized absences, the lowest RMSPE comes from specification (6). Figures 9-14 show the pre-trends in the outcomes between SAFE Taskforce areas (the treated group) and the synthetic control group, for the specification with the lowest RMSPE. The absences and suspension lines move in parallel until 2021, while the serious violence and exclusions lines look different from 2020 as the Covid-19 pandemic may have induced different responses across different LAs.

In Table 7, we show the balance of predictors of the specification with the lowest RMSPE. As a summary statement, the synthetic controls look very similar to the SAFE areas in respect of these variables. In Table 8, we show local authorities and weights that form the synthetic controls for each outcome. About nine to twelve local authorities receive positive weight. The list of local authorities is different across the outcomes. For serious violence, Sandwell and Islington are weighted highly with weights of around one-third and one-fifth. For exclusion, Islington is weighted one-fourth, followed by Newcastle upon Tyne (one-fifth), Kent, Nottinghamshire, Middlesbrough and St Helens (all about one-tenth). For suspension, Lancashire receives a weight of one-fourth, followed by Newcastle upon Tyne, City of Kingston upon Hull, St Helen and Nottingham, all with about one-tenth of weight. For overall absence, Nottingham and Lancashire receive a weight of about one-fourth and one-fifth, with other LAs receiving smaller weights. For unauthorized absences, Nottingham receives a weight of one-fourth, with other LAs receiving smaller weights. For persistent absences, Nottingham and St Helen are weighted one-fourth, and the other LAs have smaller weights.

[7] The prediction error is the difference between the actual value and the predicted value. This is then squared in order to ensure negative and positive errors are treated in the same way. We take the mean average of these squared errors. Then we take the square root of this mean to ensure the final value captures the difference between the actual and predicted value in the original units of measurement.

In summary:

- Synthetic control relies on a sufficiently long pre-intervention being observed. In our case, the number of pre-periods available is quite short compared to other implementations of the approach. Consequently, it is unlikely to be a suitable approach in this case.
- The balance of predictors looks very similar between the chosen synthetic control group and the SAFE areas.

Regression Discontinuity

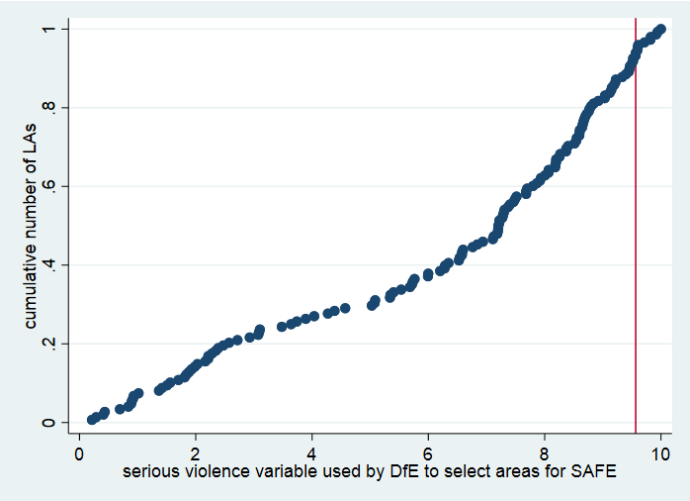
As described at the beginning of the Appendix, SAFE Taskforce areas were chosen as the top 10 locations for serious violence based on the ranking of LA-level metrics on hospital admissions for assault with a sharp object and the monthly average offences for serious violence, over a five year period. The cut-off level of the metric is 9.5 and only LAs with values of the metric above 9.5 were selected for the SAFE Taskforce programme.

This discontinuity can be exploited for estimation purposes. The intuition is that areas just above the cut-off of 9.5 are likely to be similar to those just below the cut-off, such that we can view treatment status among those close to the cutoff as being as good as randomly allocated. Treatment-control comparisons of outcomes among those close to the cut-off can provide an estimate of treatment impact.

For an RD approach to be valid, we require the percentile score to be continuous in the region of the cut-off and also that the outcome that would prevail in the absence of the treatment be continuous around the cut-off.[8] The chart below (Figure 15) demonstrates this, showing the cumulative number of SAFE Taskforces areas in 2022 as the sum of the percentile scores increases. The cut-off is at 9.5 and the maximum value of the metrics is 10.

[8] We also require that the percentile score cannot be manipulated by areas; this is guaranteed in our case.

Figure 1: Distribution of serious violence and hospital admission scores



Subsequent graphs in the Appendix **Error! Reference source not found.**concentrate on areas close to the cut-off and examine whether outcomes in previous years show a change at the cutoff. We define this closeness in two ways: i) the difference between the threshold of the metrics (9.5) and the maximum value of the metric (10) provides a bandwidth of 0.5, which encompasses all treated Taskforces, and ii) half of this value, 0.25. Looking at earlier years shows whether and how outcomes differed between areas above and below the cutoff before the intervention. The rationale for examining this is that it provides a clue as to whether untreated potential outcomes (the unobserved outcomes that would have prevailed without the SAFE intervention) would be expected to be continuous around the cut-off after SAFE is introduced.

The first set of graphs in the Appendix (within Figure 16) focuses on the change in serious violence outcome (variable name sv_pchange). We perform this for the years 2014 to 2019. We avoid the years 2020 and 2021 as results in these years are affected by COVID-related distortions (including missing outcomes). The results on the left-hand side impose a bandwidth of 0.25 and the results on the right-hand side impose a bandwidth of 0.5. Each chart plots the mean outcome for each outcome against that area's percentile sum, with each circle reflecting a single LA. The y-axis denotes the change in serious violence over the years.

Three lines are shown on either side of the cutoff. The red line marks the mean outcome among LAs on each side but within the bandwidth (the local mean). The green line shows an estimated linear relationship between the points on each side (the local linear regression). The blue line shows an estimated quadratic relationship between the points on each side (the local quadratic regression). The RD estimator is essentially the vertical difference between same-colour lines at the cut-off, where red, green and blue lines correspond to progressively more flexible ways of modelling the relationship between the outcome and the score variable.

We hope to see in these graphs that there is smooth continuous progression around the cut-off point such that the RD estimate is close to zero.

Summarising the results for serious violence, under the larger bandwidth (0.5), local mean and quadratic lines (blue lines) perform best. The estimated difference is close to zero only in 2015 and 2018.

For exclusion (Figure 17, variable name perm_rate), we see that under both bandwidths (0.25 and 0.5), local means (red lines) perform best. However, the difference is close to zero only in 2018 and 2019. For other years, the estimated difference is greater than zero, showing worse outcomes in SAFE Taskforce areas, to the left of the threshold.

Figure 18 indicates the set of graphs for the suspension outcome (variable name susp_rate) over the pre-intervention years. All lines perform poorly under the narrower bandwidth (0.25) and provide a greater than zero estimate. The bandwidth of 0.5 provides consistently better estimates (closer to zero) than the bandwidth of 0.25. The local means consistently outperform the more flexibile specifications.

Figure 19 shows the set of graphs for the fraction of absences (variable name pct_abs) over the pre-intervention years. As for suspension, all lines perform poorly under the narrower (0.25) bandwidth and better under the larger (0.5) bandwidth. Among the three lines, local means perform best, except in 2019, where the quadratic line works best.

Figure 20 shows the set of graphs for the fraction of unauthorized absences (variable name pct_unauth). All lines perform poorly, except for the quadratic lines using the bandwidth of 0.5 and only for the years after 2016.

Figure 21 shows the set of graphs for the fraction of persistent absences (variable name pct_persabs). All lines perform poorly, except for the linear lines, that show consistently close to zero estimates.

Figure 22 shows the set of graphs for the fraction of 10-13 year-olds in sustained education at age 16 (variable name eet_sustained). The pre-intervention years cover only 2018 and 2019. Linear and local means perform better than the quadratic line when using the 0.5 bandwidth.

Figure 23 shows the set of graphs for the fraction of 10-13 year-olds in sustained education or continuous employment at age 16 (variable name eet_emp_sustained). In most cases, the local mean estimator consistently out-performs the more flexible specifications.

To sum up, outcome balance is varied, both by outcome and by cohort. However, the balance using the 0.5 bandwidth results performs better than the 0.25 bandwidth and the local mean tends to out-perform local linear or local quadratic.

Another check on the validity of the RD design is to inspect whether covariates change on either side of the cut-off. The intuition here is that if there is a difference in the value of covariates on either side of the cut-off, we have less confidence that the RD estimate captures the effect of the treatment alone, since it may also be capturing the effect of compositional differences.

The charts in Figure 24 consider the set of covariates (fraction of pupils eligible for free school meals; % of pupils in state-funded education aged 10-13 with SEN met by a statement or EHC plan; % of pupils in state-funded education aged 10-13 with SEN not met by a statement or EHC plan; local Authority mean IDACI score, 2019; local Authority total population). For compactness (and in light of its superior performance shown already), only results for the wider bandwidth are shown. Furthermore, we only consider the 2018 and 2019 cohorts. The format of the charts is as before.

The covariates balance is good: FSM percentage, the SEN percentage (whether with a statement or EHC plan or not) and IDACI are very similar on either side of the cut-off, while the total population shows substantial differences. To address this difference, regression should control for the influence of this variable.

The difference in regression discontinuity estimator using the local mean specification and the 0.5 bandwidth, which is wide enough to include all SAFE Taskforces, could be used as a good robustness check for DiD results. Imposing the 0.5 bandwidth is roughly like restricting the comparison to the other APST areas not selected for SAFE, which would replicate the results presented using DiD and the non-SAFE APST comparison group in Section 2. Hence, we did not replicate the estimates.

In summary:

- The outcome balance varies, both by outcome and by cohort, but using the 0.5 bandwidth outperforms using the 0.25 bandwidth and the local mean performs better than the local linear or local quadratic.
- The covariates balance is good across observable characteristics.
- Given the similarity of the regression discontinuity comparison group to the non-SAFE APST group, there is no value added in using this specification as an additional check.

Conclusions

We undertook the preliminary analysis of pre-treatment data for participating and non-participating LA areas to inform the statistical plan.

Firstly, we examined pre-existing differences in outcomes between Taskforces and three potential comparison groups of non-participating areas: all non-treated LAs, APST LAs, and non-APST LAs.

Secondly, we tested three different methodological options for estimating treatment effects: difference-in-differences, synthetic control and regression discontinuity.

Finally, we wrote the code to estimate treatment effects and used this to calculate them for a "placebo" year to ensure that non-significant pre-treatment differences were recovered. This is done for the difference-in-differences model only, as the synthetic control does not allow to conduct inference by calculating standard errors and confidence intervals, and the regression discontinuity did not provide consistent results across different models over the years and the bandwidths in the preliminary analysis. We caveat that the analysis will be unable to detect small significant impacts given the high MDES estimated, but may inform the likelihood of impact of the programme.

Our conclusion is that the strongest evaluation approach, of those considered here, is to use difference-in-differences with the APST LAs as comparison group. This outperforms DiD using alternative comparison groups and also outperforms regression discontinuity. The synthetic control approach is likely to be hampered by the short run of pre-intervention data available. DiD with the non-SAFE APST LAs as a comparison group (correctly) did not suggest a significant placebo effect and also gave the fewest pre-intervention significant differences in the parallel trend analysis.

We make two final observations. First, because the SAFE areas and APST areas were selected in ways that, while different, are closely related, non-SAFE APST LAs are more similar to SAFE LAs in respect of the serious violence measure used to identify SAFE areas than is true of non-SAFE LAs as a whole. This means that using non-SAFE APST LAs as a comparison group is similar to selecting the comparison group on the basis of the serious violence measure, as an RD estimator would. Hence, DiD using non-SAFE APST LAs as the comparison group has close similarities to the difference-in-regression discontinuities (RDR) estimator being used as a sensitivity check in the APST evaluation and it is not implemented for the SAFE evaluation.

Second, we highlight that using DiD with non-SAFE APST LAs as a comparison group provides an estimate of the impact of SAFE relative to APST. It needs to be noted that this estimate, especially for the serious violence outcome, is confounded by the APST programme, and the extent of it depends on the timing of the APST intervention to become effective. However, for outcomes other than serious violence, APST should not be relevant. In addition, the APST intervention was introduced in both SAFE and comparison non-SAFE APST areas, so if the impact was of similar magnitude in the two areas, it should not affect outcomes differently in our context.

For these reasons, and because the non-APST comparison group performance as a comparison group is second best to non-SAFE APST LAs, we could run a robustness check using non-APST LAs as comparison group to assess the impact of SAFE relative to no intervention.

Of the approaches considered here, the best approach would therefore be to:

- Use the difference-in-difference estimator with non-SAFE APST LAs as the comparison group, with a caveat on the low probability of being able to detect significant impact.
- Perform the Rambachan and Roth sensitivity test to present robust inference.
- Perform the difference-in-difference estimator using non-APST LAs as a comparison as a robustness check on our DiD results.

Implications

Given the high minimum detectable effect sizes estimated in this preliminary analysis, which mean the approach above would be unable to detect small significant impacts, the study team has agreed with the Youth Endowment Fund to explore new designs for the evaluation approach that may have a better chance of detecting any impact the programme may have.

References

Abadie, A. and Gardeazabal, J. (2003) 'The Economic Costs of Conflict: A Case Study of the Basque Country', *American Economic Review*, American Economic Association, 93 (1, March), pp. 113–132.

Abadie, A., Diamond, A., Hainmueller, J. (2010) 'Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program', *Journal of the American Statistical Association*, 105 (490), pp. 493–505.

Abadie, A., Diamond, A. and Hainmueller, J., 2015. Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), pp.495-510.

Acemoglu, D., Johnson, S., Kermani A., Kwak, J. and Mitton T. (2016) 'The Value of Connections in Turbulent Times: Evidence from the United States', *Journal of Financial Economics,* 121, pp. 368–391.

Krief, N., Grieve, R., Hangartner, D., Turner, A., Nikolova, S. and Sutton, M. (2016) 'Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units', *Health Economics,* 25 (12).

YOUTH ENDOWMENT FUND

Impetus

Home Office

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413