what works centre for local economic growth

Lessons and recommendations from What Works Growth's evaluations of local growth policies

1. Introduction

Rigorous evaluation plays an important role in understanding the effectiveness of local growth policies. What Works Growth aims to make impact evaluation easier for policymakers, by providing advice and support. This briefing summarises the lessons learned from our 'demonstrator evaluations'. These projects involve working with partners to demonstrate that evaluation of local growth policies is possible and to assess the applicability of different evaluation methodologies to common evaluation questions.

The focus on the issues encountered while working on our demonstrators means that the briefing does not cover every issue that will arise in developing and implementing an impact evaluation. The lessons relate to three broad topics – understanding the intervention, analysis, and data. Understanding the intervention includes understanding what it aims to do and how it was implemented, specifically how it was targeted ('selection into treatment'). Analysis focuses on identifying a comparison group and selecting appropriate methods that allow for the possibility of unintended policy effects (for example, displacement). Once methods are identified, whether they can be implemented depends on what data is available, at what spatial level, and on what outcomes.

We illustrate the lessons using examples from four demonstrator evaluations of the Eat Out to Help Out (EOTHO) scheme, Enterprise Zones (EZs), the Growth Vouchers Programme (GVP), and Local Major transport schemes.

The EOTHO scheme was introduced by HM Treasury to support economic recovery after the first Covid-19 lockdown, aimed to boost demand and protect jobs in the hospitality sector. Our evaluation focused on the impact on footfall and job posts and aimed to demonstrate the ways in which timely data from non-traditional sources could be used for evaluation.[1]

The EZ demonstrator looks at the impact of these area-based policy interventions on employment, turnover, and the numbers of establishments. The evaluation aims to assess the applicability of existing distance-based methods, with a focus on combining these with other dimensions of the EZ programme (in particular, the staggered rollout and the existence of runners-up).

GVP was a large-scale randomised controlled trial (RCT) of a programme supporting small businesses. It was set-up to assess the effects on firm performance from tackling three main barriers that businesses face when accessing advice – the perception that strategic advice is not needed, uncertainty about where to find good advice, and the cost of purchasing advice (BIS, 2014).[2,3] Our demonstrator focused on the effects on turnover and employment and aimed to address some of the methodological issues present in the original analysis of the GVP trial data.[4] GVP is the only demonstrator to look at outcomes at the business level.

The Local Majors schemes evaluation considered 94 local transport schemes funded by the Department for Transport (DfT). The demonstrator focused on the impacts on employment, turnover, and number of establishments.[5] It aimed to assess the applicability of existing distance-based and accessibility methods for evaluating the impact of small additions to the transport network.

We consider each of the lessons in turn. A final section provides recommendations for improving the robustness and feasibility of future evaluations.

2. Lessons

2.1 Understanding the intervention

Understanding the intervention's aims and objectives

Understanding the intervention's aims and objectives is a crucial first step in establishing a clear causal question that an impact evaluation can try to answer.[6] But this can be challenging, particularly if they are too vague, broad or unrealistic.

1 The details and findings of the EOTHO demonstrator are published as: González-Pampillón, N., Nunez-Chaim, G., and Overman, H.G. (2024). The economic impacts of the UK's eat out to help out scheme. *Journal of Urban Economics,* 143, 103682.

2 GVP was designed and implemented by the Department for Business, Innovation and Skills (BIS) and the Behavioural Insights Team (BIT) in 2014 and 2015.

3 Department for Business Innovation and Skills (BIS) (2014). *Growth Vouchers Programme: Formative Evaluation Report.* BIS Research Paper Number 219.

4 The details and findings of the GVP demonstrator are published as: Nunez-Chaim, G., Overman, H.G., and Riom, C. (2024). *Does subsidising business advice improve firm performance? Evidence from a large RCT* (No. dp1977). Centre for Economic Performance, LSE.

5 The details and findings of the Local Majors demonstrator are published as: Donnat, H., Gómez-Hernández, L.Y., González-Pampillón, N., Nunez-Chaim, G. and Overman, H.G. (2025). *Evaluating the local economic impacts of transport projects and programmes (with an application to UK Local Major schemes)* (No. dp2081). Centre for Economic Performance, LSE.

6 Causal questions are more specific than policy questions and they require more information. A policy question – what is the effect of the EZs on employment – and a causal question – what is the average treatment effect of the EZs on the level of employment in the selected LEPs and surrounding areas (displacement) – can provide different information. The causal question aims to isolate the specific impact attributable to the policy, which is the focus of impact evaluation.

Lesson 1: If the policy aims and objectives are too vague it can be difficult to figure out which outcomes should be the focus of evaluation. For example, in the EOTHO demonstrator, the aim of helping the economy recover from the impacts of Covid-19 (EOTHO) is not specific. Our evaluation focused on footfall and job postings to translate vague objectives into concrete outcomes.

Lesson 2: Broad aims and objectives complicate the analysis because of the need for data on lots of outcomes and the additional difficulties of assessing multiple hypotheses about potential policy impacts.[7] For example, in the GVP demonstrator, the main objective was to improve firm performance across multiple aspects, and diverse capabilities and practices. While our analysis focused on analysing the impact on turnover and employment using secondary data, if all outcomes for which data was gathered through custom surveys had been analysed this would have involved multiple hypothesis tests for 56 different outcomes. At 10 percent significance levels, roughly six of these would be expected to show some effect of GVP (10 percent of 56) even if GVP had no effect.

Lesson 3: Even if aims and objectives are clearly specified, evaluation will be difficult if these are unrealistic – either because the policy does not affect the outcome of interest, or because effects are likely to be swamped by other factors. Impact evaluation aims to detect statistically significant changes in outcomes – that is, changes that could not have happened by chance. This requires there to be some way for the policy to realistically achieve the objectives (i.e. there is a 'treatment effect' to detect). In our Local Majors demonstrator, improved accessibility from a road scheme could, in principle, impact local employment so we might expect there to be a treatment effect. But if the impact of the scheme is overwhelmed by the impact of other transport investments, the arrival of a large employer in the area or another significant shock to the local economy, it may be hard to detect.

Understanding selection into treatment and intensity

Impact evaluation uses comparison groups to establish causality. The standard approach is to create a group of individuals, businesses or areas that are similar to those receiving the treatment but did not receive it. This group is known as the comparison group, while the group that receives the treatment is called the treatment group. By comparing changes in outcomes between these two groups, we can assess the impact of the treatment. The main challenge is finding a comparison group that is similar to the treatment group. Proper identification of treatment effects requires sufficient information on why some individuals, businesses or areas are treated and not others ('selection into treatment') – so that the evaluator can choose a suitable comparison group.

Lesson 4: Understanding the eligibility criteria and policy selection rules is crucial for choosing the right comparison group. Without this information, it is possible that the treated group were chosen for some reason that might affect the outcome of interest. When we compare the treated and comparison groups, we may wrongly attribute any differences in outcomes to the policy, when in fact they reflect the way in which the treatment group was chosen. For example, Local Enterprise Partnerships (LEPs) competitively bid for funding for EZs, identifying locations and potential sector focuses. However, there is no clear information on how this was done, meaning that the EZs could support either fast growing or struggling areas. If EZs mostly support promising areas, then we may over-estimate the effect of EZ status because it seems like EZ status caused faster growth, but in fact faster growth led to EZ status (and vice versa for struggling areas).[8]

7 Multiple hypothesis testing increases the risk of false positives, where a hypothesis is incorrectly accepted as true. The more tests are performed, the higher the chance that at least one will show a significant result. This means that we might think the policy is working on some outcomes when, in fact, the association only happens by chance.

8 Having incomplete data or knowledge about the variables or relationships within an evaluation is often referred to as partial information. This can happen for various reasons, such as missing data, unobserved variables, or limitations in data collection methods. Unobservables (or unobserved variables) are factors that can influence the outcome of interest but for which data is not available.

Lesson 5: Information on the policy can be used to construct comparison groups. For example, one strategy to construct comparison groups is to use those eligible for the policy but not treated. Another is to use groups that are 'just ineligible'. For example, if GVP had not been implemented as an RCT it might have been possible to construct a comparison group from businesses that were just ineligible to apply for the programme (because they had just over 50 employees).[9] Comparison groups can also be constructed using information on selection amongst those eligible. For example, if potential EZs submitted to central government were initially scored by civil servants and the scores had been used to decide which EZs were chosen, this information could be used to construct a comparison group.

Lesson 6: Variation in how the policy is applied influences the interpretation of findings. If it is unclear what 'being treated' means, then we can only interpret the treatment effects as the 'average' across different treatments. For example, businesses located in EZs were granted various benefits.[10] However, it is unclear whether all areas took advantage of every benefit or just a sub-set. This uncertainty limits our ability to determine which elements are driving impacts, making it difficult to discuss the causal effects of the intervention precisely. Similarly, for the GVP programme we know that treated businesses used their vouchers to buy one of five different types of advice from consultants, but we do not know what advice the consultants provided to the business or how long they spent providing the advice. If treatment has more than one element, we may want to separately assess the impact of each element. For example, in the GVP programme all businesses that apply to the programme are referred to a directory of approved consultants who are able to offer advice, but only some businesses get a voucher to help subsidise the cost of that advice. We may be interested in evaluating the effect of being referred to the directory or the joint effect of being referred and getting the voucher.

Lesson 7: Sometimes there can be variation in the intensity of treatment. This variation can be used in the evaluation, enabling a shift from a simple treated versus untreated approach to understanding how effects vary with treatment intensity. Less intensively treated areas can serve as the comparison group for more intensively treated areas, and vice versa. For example, for Local Majors, the intensity of treatment – how much a transport scheme improves journey times, reliability, etc. – varies across schemes and areas. Our demonstrator measures the intensity of treatment for road schemes using changes in a market accessibility index, derived from changes in travel times across areas combined with a measure of the economic 'weight' of different areas.

2.2 Analysis

Once we have information about the policy, the next step involves selecting appropriate methods to estimate the causal effect. If the chosen method is poorly matched to the context or available data, the evaluation may lead to biased conclusions and policy decisions. Choosing the appropriate methods is challenging because each comes with specific assumptions, data requirements, and limitations that can affect the reliability and accuracy of results.

Lesson 8: Evaluators must identify and apply the appropriate methods. Our guide to scoring the evidence provides more information on each method. The most robust way of assessing the causal

9 Businesses were eligible if they had been trading for at least one year, had less than 50 employees and had not paid for advice in the previous three years.

10 EZ designation offers businesses located in EZ areas a business rates discount of up to £275,000 per business in total over a 5-year period or a 100 percent enhanced capital allowance for plant and machinery investment within the zone for EZs with a manufacturing focus. Selected areas also benefit from simplified local planning rules, with Local Development Orders granting automatic planning permission for some types of development, such as new industrial buildings or changes to existing buildings, as well as investments in infrastructure and superfast broadband rollout. The LEP or local authority in charge of the EZ is also entitled to keep all business rate growth in the EZs for 25 years, which it can use for local reinvestment. Finally, areas may receive additional *ad hoc* support, for example international trade support, or help addressing local transport issues.

impact of policies for different types of individuals, businesses, and areas is by conducting an RCT. However, policy interventions often take place in complex, real-world settings where experimental conditions required by RCTs are difficult to achieve.

A quasi-experimental design (QED) may be possible when an RCT cannot be carried out. Methods include instrumental variables (IV), which use a variable related to the treatment but not directly to the outcome to estimate causal effects, regression discontinuity design (RDD), which compares outcomes for units of observation just above and below a cut-off point where the treatment is assigned, difference-in-differences (DiD), which compares changes in outcomes over time between a treatment group and a comparison group, and propensity score matching (PSM), where the comparison group is chosen using information that the evaluator has on characteristics that might affect outcomes.[11] Each QED method has different strengths.

Choosing the right QED method is challenging as the methods rely on different assumptions. For example, IV requires a valid instrument that affects the outcome only through the treatment, while RDD is appropriate when there is a clear cut-off in policy eligibility. DiD relies on similar ('parallel') trends in the treatment and comparison group in the pre-treatment period, and PSM is most effective when observable characteristics can predict selection into treatment.

Lesson 9: As outlined in Lesson 7, in some cases, the analysis focuses on the intensity rather than the occurrence of treatment. This means considering how strongly the policy affects different groups which may have implications for the appropriate QED for the impact evaluation. In our Local Majors demonstrator, we examined two methods for evaluating the local effects of transport schemes, starting with the identification of treated and comparison groups and, where relevant, a measure of treatment intensity (the degree of improvement in accessibility). Selecting an appropriate evaluation methodology depended on assumptions about the potential reach of transport scheme impacts, how treatment intensity varies across areas, and the identification of unaffected areas for the comparison group. Suitable assumptions vary with each scheme (or type of scheme), so it is important to understand the schemes to assess the suitability of different methodologies.

For schemes where treatment impacts are expected to be concentrated in a limited number of areas near the scheme and if these areas can be identified from scheme details (for example, around stations on a new tram line), treatment and comparison groups can be constructed based on distance. This distance-based approach assumes that nearby areas are generally similar, so areas close to a scheme but not directly affected may serve as a comparison group. For schemes where treatment and intensity depend on factors other than distance, accessibility-based approaches are more suitable. For example, improvements to a road junction may benefit far away areas if many journeys to or from those areas pass through the junction. In these cases, evaluation focuses on how the impact of schemes varies with the intensity of treatment measured by changes in 'accessibility' along the transport network, rather than using a treatment and comparison group approach. This means impacts can differ across areas that are the same distance from a scheme.

Lesson 10: Methods need to be carefully adapted to ensure that evaluations are tailored to the specific context. Replicating the same analysis as in previous impact evaluations can be misleading without carefully understanding the policy context and assessing whether methods are valid in the new policy setting. For example, Gibbons *et al.* (2019) estimate the effects of transport schemes in the UK using year-on-year changes in accessibility, which captures the joint effect of all road schemes.[12] Replicating this analysis to look at the effect of Local Majors would result in inaccurate estimation, as these are smaller local schemes that represent additions to the overall network.

11 Unit of observation refers to the individuals, businesses, or areas for which data on outcomes is collected.

12 Gibbons, S., Lyytikäinen, T., Overman, H. G., and Sanchis-Guarner, R. (2019). New road infrastructure: the effects on firms. *Journal of Urban Economics*, 110, 35-50.

Lesson 11: Less robust methods can be misleading in real world settings because they may fail to account for all of the factors influencing the results of the policy. For example, PSM assumes that we can predict who gets the treatment based on observable factors for which we have data. However, if there are factors affecting who receives the treatment for which we do not have data, we might not be able to choose the right comparison group for the treated group, making the results unreliable. Our GVP demonstrator illustrates the limitations of PSM in business support interventions. Despite using comprehensive administrative data and implementing the method carefully, PSM results show positive, lasting impacts on turnover and employment over time, which are inconsistent with the RCT findings.[13] The lesson from GVP is that unobserved factors that affect selection in to treatment (in this case, businesses that are already likely to grow tend to apply to GVP) can introduce bias in PSM that is both substantial and cumulative over time, even with high-quality data. These results highlight the importance of selecting the most robust QED method feasible.

Lesson 12: Methods are developing constantly, and impact evaluation can take advantage of these methodological improvements. An example of this is the evaluation of policies where treatment occurs at different points in time rather than all at once. Recently, previously unrecognised issues in these estimations have been identified and new tools have been developed to better identify effects when a policy has different treatment timings. A staggered treatment adds complexity, as treatment effects may vary across time and among treated groups. This issue is evident in our EZ demonstrator, where we lack information on the policy's design and eligibility criteria and cannot confirm whether the incentives are standardised across cohorts. Additionally, unobservable characteristics may cause treatment effects to vary depending on whether the EZ was designated in the 2011 or 2015 rounds with short-run effects potentially differing from long-run effects. Even assuming that incentives were standardised within cohorts, the effects may still vary according to the area or the specific group characteristics. Given the multiple groups and time periods of the treatment, our demonstrator uses a recently developed staggered difference-in-differences approach to improve estimates.[14]

Lesson 13: Including deadweight and displacement effects in the analysis is important. To achieve their objective, interventions must generate additional economic activity in the targeted area. Deadweight may be present if the policy incentives do not change behaviour – for example, in our EZs demonstrator, if businesses in the targeted area benefit from the tax breaks without doing anything or if changes such as expansion were already planned by businesses before the area was designated. The extent to which policies generate benefits for the wider local economy also depends on local displacement. Displacement occurs when economic activity shifts from one area to another as a result of the policy. For example, businesses may relocate in response to EZ designation, benefiting the target area but negatively impacting nearby areas. In this case, increasing local growth in one area comes at a cost of lower (or negative) growth for neighbouring areas. The spatial extent of displacement is important in assessing whether the policy supports local growth or only benefits the target zone. Evaluating local growth policies involves estimating the effect of the policy on economic outcomes while separating this effect from displacement.

Displacement makes the selection of a comparison group more complex. Nearby areas may be similar to the treated area but are also likely to experience displacement effects. In our Local Majors and EZ demonstrators, we address displacement by using rings around the intervention at increasing distances.[15] This way, rather than assume effects are a linear function of distance, the definition of rings around the treated group allows for non-linearity in how distance affects intensity. For example,

13 RCT is a more robust method than PSM, as set out in Lesson 8 and our guide to scoring the evidence.

14 For a synthesis of estimators robust to treatment effect heterogeneity, see Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2), 2218-2244.

15 In both demonstrators, we implement spatial difference-in-differences using rings of comparison areas around the EZ at increasing distance.

in Local Majors, creating rings of one kilometre (km) around a public transport station, we can compare rings of treated areas within 0 to 1km, 1 to 2km and 2 to 3km of the station to comparison areas 3 to 4km away. If new businesses observed within 0 to 1km of a scheme are moving from 1 to 2km ring (displacement) then we should expect to see negative effects in the 1 to 2km ring. We might expect the intensity of treatment and the impact of the scheme to be greatest in the closest ring, but the use of rings does not impose this assumption (whereas modelling the intensity of treatment as a linear function of distance would). This flexibility comes at a cost because it reduces the number of areas in each ring and hence the precision of the estimates. Having more observations provides more information which makes it easier to estimate the impact of the policy. Choosing an appropriate distance threshold is not a simple task. Ideally, the threshold should allow for all affected areas to be within the rings of treated areas while excluding unaffected areas, enabling a clearer analysis of displacement effects without interference from the treatment. However, there is no clear way to define this.

2.3 Data

The availability of suitable, timely data on the outcomes of interest (such as employment, productivity or wages) and for area-level evaluations, at the appropriate spatial level, ensures that impact evaluation can draw valid conclusions. Without the right data, it becomes challenging to measure the true impact of interventions, identify causal relationships, and address potential biases. This often requires merging data. Confidential data – such as tax records or employment details – may provide crucial demographic or socioeconomic details for tracking participants over time. By combining with open data, evaluators can better estimate treatment effects. Although merging confidential and open data can improve the accuracy of findings, accessing confidential data can be difficult, and for area-level evaluations, it is essential to ensure that both datasets can be reconciled to the same level of spatial granularity and detail, particularly for outcomes. Making a choice between secondary and primary data, or a combination of both, is also important.

Lesson 14: Outcome data that is not available or measured with errors can cause problems when evaluating local growth policies and result in over- or under-estimating the policy's benefits. For example, in our EOTHO demonstrator, we focused on job posts data in the absence of employment data, but this means it is not possible to determine if job posts led to hires or if any changes in employment were lasting. We also lack data on whether EOTHO increased turnover or improved business survival probabilities. Access to data on these outcomes could have improved the evaluation but would also require the government to be willing to identify subsidised businesses. Furthermore, limitations in outcome data sources – such as footfall data (skewed towards younger individuals) and online job posts (biased towards expanding businesses) – may overstate EOTHO's overall impact.

Lesson 15: Impact evaluation of policies whose outcomes are measured at the area-level requires that data on relevant characteristics and on outcomes is at the appropriate spatial level. Data with low spatial granularity can lead to bias in the estimated impacts and limit evaluators' ability to detect smaller, more localised effects.[16] For example, if a strong positive impact from a policy exists but data is only available for a larger area that includes untreated areas which are not expected to change, the effect could be diluted. If there are displacement effects around the treated areas, if both the treated area and neighbouring area experiencing displacement are included in the geography available, they will offset each other fully or partly. In both cases, using larger spatial units for the analysis makes it

16 Data granularity refers to the level of detail or precision contained within a dataset. In the context of spatial data, it refers to the level of detail regarding geographic locations within a dataset. High spatial granularity captures fine geographic details, such as specific addresses, GPS coordinates or small administrative zones, allowing for precise location. Low spatial granularity represents broader geographic areas, such as regions, cities or countries, where details within those areas are aggregated or averaged.

more difficult to identify the precise effect and detect displacement, particularly if the areas nearest to the treated area are the most likely to be affected.

In the UK, full postcodes allow better geo-location of schemes and improved definition of which areas or businesses are treated, which should give more precise estimates for smaller schemes where the economic impact is expected to be small and help distinguish growth from displacement. Unfortunately, postcode information has been removed from the Business Structure Database that provides data on numbers of establishments, employment, and turnover, making it more difficult to evaluate area-based policies.

For Local Majors, evaluating small additions to an existing network requires granular and comprehensive data, as the expected benefits from minor additions may be positive but relatively small. As the most granular data available is ward-level, this affects our ability to identify effects of road improvements. In the case of EZs, there was an average overlap of just 16 percent between the geography our demonstrator could use for the identification of 'treated' areas (Lower Layer Super Output Areas, LSOAs) and the designated EZs. Since the same strategy was used to establish comparison groups, the comparison areas are similarly imprecise.

Lesson 16: Choosing between primary and secondary data, or a combination of both, involves important trade-offs.[17] While primary data collection may allow for the collection of information that aligns closely with that needed by the evaluation, it can be much more costly. Secondary data is often more accessible and affordable, making it attractive for evaluations with limited budgets. But secondary data may not align as closely with the evaluation's specific requirements, leading to gaps in crucial information or reducing the precision of the findings. Both types of data can vary in quality, which may impact the robustness of the results if not carefully managed.

Many evaluations use a combination of primary and secondary data. For example, in the GVP demonstrator, we assess impacts using administrative data from the Inter-Departmental Business Register (IDBR) and complement this with information from baseline and endline survey data collected 12 and 24 months after businesses received diagnostic support.

In earlier impact evaluations of GVP, only the surveys were used for the analysis. This is problematic for two main reasons. First, the approach to collecting the information was different across cohorts of surveys. The baseline and 12-month surveys used a census approach, while the 24-month survey only contacted participants who responded to the 12-month survey. This inconsistency can affect the comparability of the data across different time points because there is a significant drop-off in the number of businesses completing each subsequent survey (from a little over 28,000 at baseline to a little under 2,000 at 24 months). There can be significant differences between those who stay and those who leave. For this reason, the final group might not represent the original group accurately, and the findings can be misleading. Moreover, this can lead to a comparison group not being similar to the treatment group, the main challenge in impact evaluation, as discussed in Lessons 4 to 7. Second, self-reported performance by businesses can lead to inconsistencies such as over-reporting success or under-reporting issues. By combining survey data with administrative data from IDBR, we can achieve a more robust impact evaluation.

Another challenge in using secondary data is that it can have a publication lag. For example, there is a two-year lag in turnover data in the IDBR.

17 Primary data refers to information collected by the researcher for a specific purpose, including surveys, interviews or experiments. Secondary data is information that has already been collected and published by others, such as census data, tax records, and statistical databases.

3. Recommendations

For many of the individual lessons, the discussion should point to ways that evaluation of local growth policies can be improved. Here we provide some cross cutting recommendations.

Recommendation 1: Collect detailed information on the policy

In an ideal scenario, impact evaluation should be integrated into the project development process. This maximises the evaluation options available and ensures that understanding the intervention, conducting the analysis, and gathering all necessary data can be done at a lower cost. If this is not feasible, it will be crucial to allocate time to gather information on the timing, objectives, and target population of the intervention during the evaluation planning phase.

Impact evaluation requires understanding selection into treatment. To serve as an accurate counterfactual, the comparison group must closely resemble the treated group in key characteristics. Collecting detailed information on treatment assignment can also help identify runners-up. In some cases, runners-up can serve as a comparison group because they closely resemble the treatment group, having undergone the same selection process but narrowly missing out on receiving the intervention. This similarity helps to minimise selection bias, as differences between the groups are less likely to affect the outcomes being measured. Collecting information on the policy can also help identify variations in the intensity of treatment. This way, the effect of the treatment can be determined by shifting from a binary treatment and comparison definition to a continuous treatment analysis.

Recommendation 2: Choose your method carefully

If RCTs are not feasible, a deep understanding of the policy, along with clear definitions of its aims, target population, and expected outcomes can help selecting the right QED. In choosing the method, consider additional information or assumptions about participant behaviour. If possible, apply two or more methods and compare both the magnitude and direction of results. Even methods regarded as reliable in past impact evaluations need thorough scrutiny each time.

When pre-treatment periods are unavailable, or unobservable differences between treatment and comparison groups may affect outcomes, be cautious of using methods that rely only on observable data to construct the comparison group. If available data covers a wide range of observable characteristics and unobservables are unlikely to have significant influence, then it may be reasonable to rely on QED methods that use observable data to construct the comparison group. For example, PSM may be used to match units of observation relying on a propensity score created using a substantial set of observable variables. This may be less problematic in evaluations involving individuals, as data on personal characteristics is often more comprehensive and accessible. As our GVP demonstrator shows, the gap between RCT and PSM findings can be large and economically and statistically significant when unobservable characteristics are more important. In the case of GVP, bias also increases over time, suggesting it may be important to consider the evaluation's time horizon.

If available data does not provide information on a wide range of observable characteristics or if unobservables are likely to have significant influence on both outcomes and whether or not a unit of observation is treated, relying on methods (such as PSM) that construct comparison groups on the basis of observables may be very misleading and careful consideration should be given to whether it is feasible to implement another QED that better allows for unobservable factors. Our guide to scoring the evidence provides more information on each method.

Recommendation 3: Consider displacement effects when needed and possible

When considering displacement effects for outcomes at the area level, using rings around intervention areas is often an easy-to-use approach to test for displacement. To guide decisions on ring size and the boundaries of the treated area, evaluations can incorporate additional data or assumptions on how users benefit from the intervention, for example, estimating how far people are willing to walk to access public transport in Local Majors schemes.

The choice of threshold also influences whether nearby areas are similar to the treated area. One approach is to compare pre-trends in economic outcomes for both groups before the policy, which can offer reassurance that they are comparable, though it does not guarantee that the rings are unaffected by the intervention. To account for this, as with the threshold for treated groups, robustness checks can use different thresholds to see if results change. Ideally, the threshold distance for the treated group should include all affected areas while excluding unaffected ones, although as outlined earlier, data granularity can affect whether this is feasible in practice.

Recommendation 4: Define appropriate outcomes and the relevant unit of observation for analysis

Impact evaluations rely on data on the outcomes of interest (for example, employment or wages). These should be realistic (i.e. possibly affected by the intervention) and measurable. It is also essential to establish the unit of observation – i.e. will outcomes be measured at the individual, business or area level. If outcomes are measured at the area level, this involves defining the spatial level (for example, local authorities, wards or LSOAs) that will used. All data – including on covariates – will need to be available at the same spatial level.[18] While granular data is better for precision, aggregate data can be useful for a more macro picture. For example, when evaluating transport schemes, larger geographies help account for displacement and the overall response of employment when road capacity or coverage changes. Using larger geographies, however, has its limitations, as displacement can lead to zero effects for larger units which would be revealed by more granular data.

Recommendation 5: Collect information on treatment and outcomes using geospatial tools

Combining geospatial data with socio-economic information can allow for a more precise and contextualised definition of treatment group (and intensity) and comparison groups. The combination of survey and administrative data with geospatial data may help ensure comparison areas are similar to treatment areas, resulting in more accurate estimates of the treatment effect. This approach not only enhances the analysis of effectiveness but also makes it easier to identify how localised policy impacts are and assess their spatial distribution, determining whether policies are targeting the appropriate areas.

Geospatial data is necessary to correctly identify the effects of spatial spillovers and study displacement. The rings method for causal effect estimation allows us to assess to what extent the effects observed at the intervention site result from the displacement of economic activity from nearby areas (see Lesson 13 and Recommendation 3). Proper definition of the ring size is crucial, as it determines the potential bias in our estimates, with the bias reflecting the difference in treatment

18 Covariates are variables that are related to both the treatment and the outcome in an evaluation. For example, for a policy that supports individuals, they could include characteristics such as age, income level, education or location. Including covariates in an analysis is important because they help control for other factors, ensuring that the observed effects are more likely due to the intervention itself rather than underlying differences among participants. By accounting for covariates, the evaluation can produce more accurate, unbiased estimates of the policy's true impact on the target outcomes.

effects experienced by the 'treated' ring and the 'comparison' ring.[19] If the treated ring is defined too narrowly, comparison units may be influenced by the intervention, leading to biased estimates. However, if the ring is too large, then a positive effect on the treated area can be diluted by a negative effect in nearby areas where the economic activity was displaced. The feasibility of using the appropriately sized rings will depend on the spatial granularity of the data.

Recommendation 6: Consider time and cost of the data collection

Carefully consider the time and cost of data collection. This helps to maintain data quality while preventing delays that could render the findings less relevant for policymaking. High-quality, timely data (as described in Section 2.3) is crucial for producing reliable results, but extended timelines or excessive costs may compromise evaluation scope or lead to unreliable conclusions.

Evaluations that are too expensive or time-consuming may be difficult to replicate or scale in other contexts or adapt to other growth policies. Balancing these factors allows for a sustainable, cost-effective evaluation process, with results that are relevant to the policy questions.

4. What Works Growth resources

What Works Growth have a range of resources on 'how to evaluate' including:

- An eight-step guide to better evaluation.
- Our guide to scoring the evidence providing an overview of the main evaluation approaches.
- Proformas to help identify which evaluation methodologies might be suitable for an intervention (for example, on transport).
- Evaluation case studies providing examples of how different methodologies have been applied.
- Detailed guidance on specific evaluation topics, such as programme-level evaluation and evaluating devolution.

We also offer training and can provide support for some projects in UK local and central government. If you are planning an impact evaluation of a local growth policy and would like to discuss how we could support, please get in touch.

19 Gerardi et al. (2015) and Sullivan (2017) discuss this problem.

- Gerardi, K., Rosenblatt, E., Willen, P.S., and Yao, V. (2015). Foreclosure externalities: New evidence. *Journal of Urban Economics* 87 42–56.
- Sullivan, D.M. (2017). The true cost of air pollution: Evidence from the housing market. *Working Paper.*

what works centre for
local economic growth

This work is published by the What Works Centre for Local Economic Growth, which is funded by a grant from the Economic and Social Research Council, the Department for Business and Trade, the Ministry of Housing, Communities and Local Government, and the Department for Transport. The support of the Funders is acknowledged. The views expressed are those of What Works Growth and do not represent the views of the Funders.

March 2025
What Works Centre for Local Economic Growth

info@whatworksgrowth.org
@whatworksgrowth

www.whatworksgrowth.org

HM Government

UKRI Economic and Social Research Council