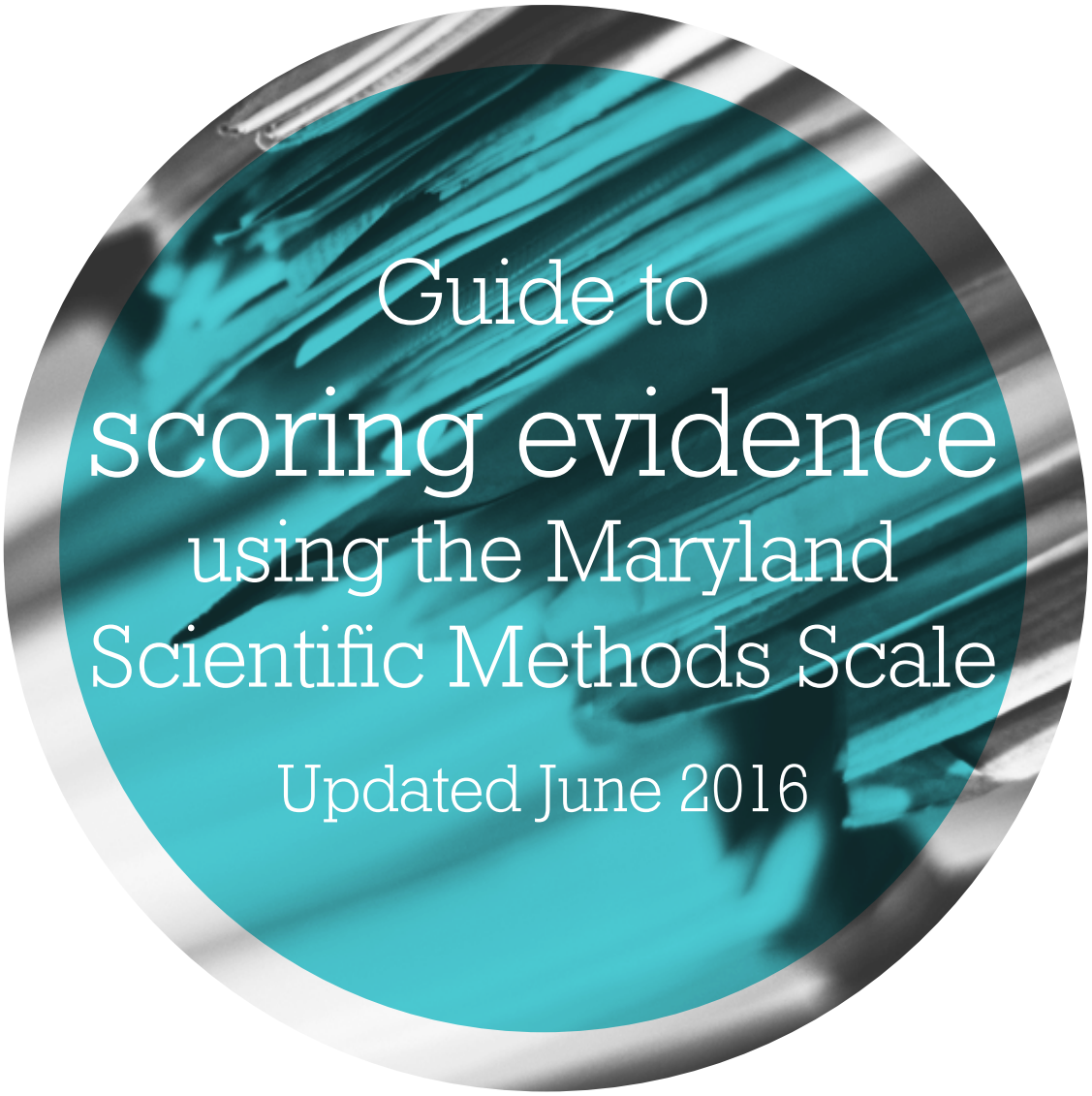

Guide to scoring evidence using the Maryland Scientific Methods Scale

Updated June 2016

what works centre for
local economic growth

00

Contents

1. Introduction 1
2. Impact evaluation 3
    Using Counterfactuals 3
3. SMS 5 methods 6
    Randomised Control Trial (RCT) 6
4. SMS 4 methods 10
    Instrumental variables (IV) 10
    Regression Discontinuity Design (RDD) 13
5. SMS 3/4 methods' 17
6. SMS 3 methods 18
    Difference-in-differences (DiD) 19
    Panel data methods 20
    Propensity Score Matching (PSM) 23
7. SMS 2 methods and below 26
    Cross-sectional regression 26
    Before-and-after 27
    Additionality (not SMS scoreable) 28
    Impact modelling (not SMS scoreable) 29
Appendix 1: Quick-scoring scoring guide for the Maryland Scientific Method Scale 30
Appendix 2: Mixed methods scoring 4 or 3. 33
    Hazard Regressions 33
    Heckman two-stage correction (H2S)/ Control function (CF) 36
Appendix 3: Arellano-Bond method 39

01

Introduction

The What Works Centre for Local Economic Growth (WWG) produces systematic reviews of the evidence base on a broad range of policies in the area of local economic growth. An important step in the review process is the assessment of whether an evaluation provides convincing evidence on likely policy impacts. Our assessment is based on the scoring of papers on the Maryland Scientific Methods Scale (SMS), which ranks policy evaluations from 1 (least robust) to 5 (most robust) according to the robustness of the method used and the quality of its implementation. Robustness, as judged by the Maryland SMS, is the extent to which the method deals with the selection biases inherent to policy evaluations.

This document examines a wide variety of commonly employed methods and explains how we place them on the Maryland SMS. It then looks at a number of examples of policy evaluations for each method, scoring them on the quality of their implementation.

This scoring guide plays several important roles. Firstly, consistent with our commitment to openness, it serves as documentation as to how we rank studies by robustness for our series of systematic reviews and enables us to be confident that we are achieving a level of consistency in our ranking across the team. Secondly, it acts as a scoring handbook for anyone wanting to assess the robustness of a particular policy evaluation. This provides useful guidance on how much weight to put on a particular piece of evidence. Finally it can help organisations undertaking evaluations either in assessing them after completion or in helping choose between methodologies beforehand.

This document is no substitute for better technical training and expert advice. But it should help those with some knowledge of evaluation techniques to better understand recent advances and the way that we treat these in our systematic reviews. It's also important to note that the ranking of individual studies is not an exact science and often involves a degree of judgement. Statistical testing can only take us so far in assessing the suitability of a given method and the quality of its application in practice. When assessing an individual study (including the specific examples that we discuss here) there is always scope for some disagreement on the exact ranking. Indeed, anyone who has attended an academic seminar will know the extent to which such issues can be hotly disputed. That said, on average our scoring will tend to produce rankings on which many evaluation experts would broadly

agree. Of course, when pulling together our evidence reviews we look at the overall balance of the evidence and it's therefore highly unlikely that a specific ranking on any given study will influence the conclusions that we reach on policy effectiveness.

The examples that we use are drawn from a wide range of studies. Not all of them are specifically focussed on local economic growth but in all cases they demonstrate approaches that could feasibly be taken in future evaluations.

02

Impact evaluation

Governments around the world increasingly have strong systems to monitor policy inputs (such as the amount of loans guaranteed to SMEs, which policymakers hope will create new jobs) and outputs (such as the number of firms that have received loan guarantees). However, they are less good at identifying policy outcomes (such as the effect of providing a loan guarantee on firm employment). In particular, many government-sponsored evaluations that look at outcomes do not use credible strategies to assess the causal impact of policy interventions.

By causal impact, the evaluation literature means an estimate of the difference that can be expected between the outcome for (in this example) firms 'treated' in a programme, and the average outcome they would have experienced without it. Pinning down causality is a crucially important part of impact evaluation. Estimates of the benefits of a project are of limited use to policy makers unless those benefits can be attributed, with a reasonable degree of certainty, to that project.

The credibility with which evaluations establish causality is the criterion on which this review assesses the literature.

Using Counterfactuals

Establishing causality requires the construction of a valid counterfactual – i.e. what would have happened to programme participants had they not been treated under the programme. That outcome is fundamentally unobservable, so researchers spend a great deal of time trying to rebuild it. The way in which this counterfactual is (re)constructed is the key element of impact evaluation design.

A standard approach is to create a counterfactual group of similar places, people or businesses not undertaking the kind of project being evaluated. Changes in outcomes can then be compared between the 'treatment group' (those affected by the policy) and the 'control group' (similar places, people or businesses not exposed to the policy).

A key issue in creating the counterfactual group is dealing with the 'selection into treatment' problem. Selection into treatment occurs when participants in the programme differ from those who do not participate in the programme.

An example of this problem for access to finance programmes, as in our example above, would be when more ambitious firms apply for support. If this happens, estimates of policy impact may be biased upwards because we incorrectly attribute better firm outcomes to the policy, rather than to the fact that the more ambitious participants would have done better even without the programme.

Selection problems may also lead to downward bias. For example, firms that apply for support might be experiencing problems and such firms may be less likely to grow or succeed independent of any advice they receive. These factors are often unobservable to researchers.

So the challenge for good programme evaluation is to deal with these issues, and to demonstrate that the control group is plausible. If the construction of plausible counterfactuals is central to good policy evaluation, then the crucial question becomes: **how do we design counterfactuals?** This scoring guide explains the ways in which we assess how well researchers have done in answering this question. In particular it explains how we rank impact evaluations based on an adjusted version of the Maryland Scientific Methods Scale (SMS) to do this.[1] The SMS is a five-point scale ranging from 1, for evaluations based on simple cross sectional correlations, to 5 for randomised control trials. Box 1 provides an overview of the scale and highlights some of the approaches that we discuss in more detail in the remainder of this guide.

1 Sherman, Gottfredson, MacKenzie, Eck, Reuter, and Bushway (1998).

Box 1: Our robustness scores (based on an adjusted Maryland Scientific Methods Scale)

Level 1: Either (a) a cross-sectional comparison of treated groups with untreated groups, or (b) a before-and-after comparison of treated group, without an untreated comparison group. No use of control variables in statistical analysis to adjust for differences between treated and untreated groups or periods.

Level 2: Use of adequate control variables and either (a) a cross-sectional comparison of treated groups with untreated groups, or (b) a before-and-after comparison of treated group, without an untreated comparison group. In (a), control variables or matching techniques used to account for cross-sectional differences between treated and controls groups. In (b), control variables are used to account for before-and-after changes in macro level factors.

Level 3: Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual (e.g. difference in difference). Justification given to choice of comparator group that is argued to be similar to the treatment group. Evidence presented on comparability of treatment and control groups. Techniques such as regression and (propensity score) matching may be used to adjust for difference between treated and untreated groups, but there are likely to be important unobserved differences remaining.

Level 4: Quasi-randomness in treatment is exploited, so that it can be credibly held that treatment and control groups differ only in their exposure to the random allocation of treatment. This often entails the use of an instrument or discontinuity in treatment, the suitability of which should be adequately demonstrated and defended.

Level 5: Reserved for research designs that involve explicit randomisation into treatment and control groups, with Randomised Control Trials (RCTs) providing the definitive example. Extensive evidence provided on comparability of treatment and control groups, showing no significant differences in terms of levels or trends. Control variables may be used to adjust for treatment and control group differences, but this adjustment should not have a large impact on the main results. Attention paid to problems of selective attrition from randomly assigned groups, which is shown to be of negligible importance. There should be limited or, ideally, no occurrence of 'contamination' of the control group with the treatment.

Note: These levels are based on but not identical to the original Maryland SMS. The levels here are generally a little stricter than the original scale to help to clearly separate levels 3, 4 and 5 which form the basis for our evidence reviews.

03

SMS 5 methods

SMS 5 methods require full randomisation of programme participation. This can only be done by a randomised control trial. As such there is only one method in this category on the SMS. Randomisation, properly applied, means there is no selection into the treatment. This ensures that there are no differences between the treatment group and control group either on observable (e.g. age) or unobservable (e.g. ability) characteristics. Any difference post-treatment must therefore be an effect of the treatment. Evidence of this type is the highest quality of evidence and is considered the 'gold standard' for policy evaluations.

Randomised Control Trial (RCT)

An RCT is defined by random assignment to either the treatment or control group. A typical RCT will involve the following steps. First, the original programme applicants may be pre-screened on eligibility requirements. Second, a lottery (computer randomisation) assigns a percentage of the eligible applicants (usually 50%) to the control group and the remainder to the treatment group. Third, baseline data is collected (either from an existing data source or from a bespoke 'baseline' survey). Fourth, the treatment is applied. Fifth and finally, data is collected some time after the treatment (again, either from an existing data source or a bespoke 'follow–up' survey). Because individuals are randomly assigned to the treatment and control groups, there is no reason that they should differ on either observable or unobservable characteristics in the baseline data. This means that differences in the outcome variables in the post-treatment data will be entirely attributable to the treatment i.e. are free from selection bias. For this reason an RCT scores the maximum score of 5 on the Maryland scale.

In order for an RCT to achieve the maximum SMS 5 score on implementation, three criteria must be met. Firstly, the randomisation must be successful. Whether or not this is the case, is often tested using "balancing tests" that compare treated and control individuals in the baseline data on a range of characteristics. If randomisation has been successful then the balancing tests should show no significant differences between the two groups. Secondly, attrition must be carefully addressed. Attrition happens whenever individuals drop out from the study e.g. do not complete treatment or do not provide follow-up data (when this is collected using a follow-up survey). The availability of follow-up data is often a particular problem for individuals in the control group since they have less incentive

to stay in the study. If drop out is on a random basis, then attrition is not an issue. However, there may be elements of selection bias for dropping out. For instance, it may be the case the less skilled individuals in a control group choose to leave an unemployment training study. This would make the remaining control group more skilled on average increasing the likelihood that they find employment. In turn, this would lead to a downwardly biased treatment effect. For this reason, policy evaluators should pay careful attention to the issue of attrition. If attrition is liable to selection bias, then evaluators should satisfactorily address the issue.

For instance, in a study that looks at the impact of neighbourhood crime on the propensity of refugees to commit crime themselves, if certain individuals drop out of the study because they move abroad, then there is a problem of potentially biased attrition. To overcome this, the authors could look into what factors lead refugees to go abroad, and include these controls in their study (we discuss an example of this below). Thirdly, the experiment should be designed such that the treatment does not spill over to the control group i.e. contamination must not be issue. In order for the control group to be a suitable comparator, it must not in any way be exposed to the treatment. If this condition is not respected, then the treatment effect may be downwardly biased because the control group partly benefits from the treatment. Contamination may happen if, for example, individuals in a given neighbourhood are given financial management advice, and they share their knowledge with people in other neighbourhoods who are supposed to go untreated.

Method	Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Randomised Control Trial (RCT) *a.k.a.* Field Experiment	5, 5 if • Randomisation is successful • Attrition carefully addressed or not an issue • Contamination not an issue	5, 4 if • One of the criteria is severely violated • 5,3 if • Two or more of the criteria are severely violated

RCT Example 1 (SMS 5, 5)

A 2013 paper by Jens et al. published in the American Economic Review evaluates the impact of neighbourhood quality on life outcomes.[2] In this case, the evaluation is difficult because it is possible that the residents of disadvantaged neighbourhoods have worse outcomes because of individual or family characteristics. To an extent, certain types of families "self-select" into bad neighbourhoods, meaning that there is a problem of selection bias. In this instance, the authors surpass this issue by evaluating the impact of Moving to Opportunity (MTO), a programme that gave families within disadvantaged neighbourhoods the opportunity to move to more affluent areas on a random basis. In order for the study to achieve the maximum SMS score of 5, the three criteria must be met.

1. Randomisation is successful

Pass

The authors make a convincing argument in favour of the truly random nature of the MTO programme. In a preliminary stage, interested residents of poor neighbourhoods (4,604 low-income public housing families) enrolled in MTO. These families were randomly assigned to one of three groups: the Experimental group (which received housing vouchers that subsidised private-market rents in low-

2 Jens, L., Duncan, G., Gennetian, L., Katz, L., Kessler, R., Kling, J., Sanbonmatsu, L. (2013). Long-Term Neighborhood Effects on Low-Income Families: Evidence from Moving to Opportunity. American Economic Review, 226-231.

poverty communities), the Section 8 group (which received unconstrained housing vouchers), and the control group. Balancing tests show that treatment and control groups are similar according to a range of observable characteristics, meaning that the randomisation was successful.

2. Attrition carefully addressed

Pass

Around 90 per cent of the individuals that were originally enrolled in the program were interviewed 10 to 15 years after the baseline year (from 1994 to 1998). This is a fairly high rate, so that sample attrition is not a problem in this instance.

3. Contamination not an issue

Pass

The vouchers were in no way transferable, meaning that contamination is not an issue in this case.

Given that all three criteria are satisfied, this study achieves SMS 5 for implementation.

RCT Example 2 (SMS 5, 5)

A 2014 paper by Damm and Dustmann published in the American Economic Review evaluates the impact of early exposure to neighbourhood crime on subsequent criminal behaviour.[3] Isolating how neighbourhoods affect young people's propensity to commit crimes is difficult, as it is possible that children growing up in crime-ridden communities come from family environments that are themselves conducive to criminal behaviour. To overcome this problem, the authors exploit an event in which asylum-seeking families were randomly assigned to communities with differing levels of crime by the Danish government.

1. Randomisation is successful

Pass

Upon being approved for refugee status, asylum-seekers were assigned temporary housing in one of Denmark's 15 counties. Within the county, the council's local office assigned each family to a municipality. This assignment was random, although councils were aware of birth dates, marital statuses, number of children, and nationalities. The authors highlight the fact that the council in no way based its decision on the family's educational attainment, criminal record, or family income, simply because it was not privy to this information. It also did not know anything about the family besides what was written in the questionnaire. Furthermore, the family's personal preferences were not taken into account. The baseline balancing test confirms that the treatment and control groups were indeed observably similar before the treatment.

2. Attrition carefully addressed

Pass

Of the sample of 5,615 refugee children, 975 left Denmark before the age of 21. Furthermore, an additional 215 children were not observed in every year between arrival and age 21. Nonetheless, the authors find no significant relationship between leaving Denmark or not being observed all years and the outcome variables. This indicates that there is no significant selection bias inherent to dropping out of the study.

3 Damm, A. & Dustmann, C. (2014). Does Growing Up in a High Crime Neighborhood Affect Youth Criminal Behavior?. American Economic Review, 1806-1832.

3. Contamination not an issue

Pass

Although families were randomly assigned to a municipality and were encouraged to stay there for at least the duration of the 18-month introductory period, they were no strict relocation restrictions. This means that it is possible that families that received the treatment (being in a neighbourhood with more crime) did not necessarily adhere to the treatment, and families that were controls (lived in safer neighbourhoods) may have been exposed to the treatment. The authors find that around 80 per cent of the families stayed in their assigned area after the first year, and half of all families stayed in the assigned area after eight years.

Given that all three criteria are satisfied, this study achieves SMS 5 for implementation.

RCT Example 3 (SMS 5, 3)

A 2012 report by Doyle evaluates the impact of Preparing for Life (PFL), a programme that provides support to families from pregnancy until the child is old enough to start school.[4] It is likely that there is a significant selection into treatment – more concerned parents may choose to join the programme. Therefore, in order to accurately understand the effect of the treatment, the author employs a randomised control trial.

1. Randomisation is successful

Fail

PFL is specific to certain deprived catchment areas in Ireland. Within these catchment areas, 52 per cent of all pregnant women participated in the programme, with the remaining 26 per cent rejecting the offer, and another 22 per cent not having been identified. In a subsequent stage, PFL recipients were randomly assigned to one of two groups: high treatment and low treatment. An observationally similar community was later used as a control, which implies that treatment was not, in fact, randomly assigned.

2. Attrition carefully addressed

Fail

The authors do not discuss the extent to which treated women dropped out of the programme.

3. Contamination not an issue

Fail

Given that a large part of PFL entails the provision of information (i.e. mentoring and development information packs) within a community, it is possible that neighbours share the contents of their treatment. This is particularly true for the high and low treatment groups, as they live in the same community.

Given that all three criteria are violated, this study achieves SMS 3 for implementation.

4 Doyle, O., & UCD Geary Institute PFL Evaluation Team. (2012). Preparing for Life Early Childhood Intervention Assessing the Early Impact of Preparing for Life at Six Months.

04

SMS 4 methods

SMS 4 methods are characterised by the exploitation of some source of 'quasi-randomness'. That is, randomness that has not been deliberately imposed but arises because of some other reason. Usually, this means identifying historical, social or natural factors that result in policy being implemented in a way that is to some extent random. By identifying cases where a policy was implemented to some extent randomly, the SMS 4 methods try to ensure that the treatment and control groups are similar on observable and unobservable characteristics. However, unlike controlled randomisation, they need to make the case that the resulting variation in treatment is truly random. If this argument fails in practice then treatment and control groups differ and this can lead us to incorrect conclusions about treatment effects. It is for this reason that these methods score 4, rather than a maximum 5, on the SMS scale.

Instrumental variables (IV)

To solve the problem of selection bias, a policy evaluation can use the instrumental variables (IV) method. This approach entails finding something that explains treatment but has no direct effect on the outcome of interest (and is not related to any other factors that might determine that outcome). This factor, the 'instrument', substitutes for the treatment variable that may itself be correlated with other characteristics that affect outcomes (and thus cause selection bias). For instance, when looking at the impact of highways on population density within cities, it is possible that not only might highways impact population density, but that the construction of highways might respond to changes in population density. A good way to circumvent this problem, then, is to employ an IV. For example (discussed further below) Baum-Snow (2007) uses a planned US highway grid that was planned in 1947 (but not necessarily built) as an instrument for actual US highways. The logic behind this instrument is that it is essentially random in its assignment (as far as population density goes), because the 1947 grid was planned for military and trade purposes. This element of randomness is essential for a successful IV. In this sense, it is a way of simulating the random assignment to treatment that is done with randomised control trials. A successful instrument has the potential to eliminate differences between treatment and control group on observable or unobservable characteristics and thus score a maximum of 4 on the SMS scale. In order to achieve this score, the

instrument must satisfy three main criteria. Firstly, the instrument must be exogenous, or random in assignment. Secondly, it must be relevant, or credibly related to the variable that it replaces. Thirdly, it must be excludable, meaning that it does not directly impact the outcome.

Method	Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Instrumental Variable (IV) *a.k.a.* Two-Stage Least Squares (2SLS)	4, 4 if instrument is: • Relevant (explains treatment) • Exogenous (not explained by outcome) • Excludable (does not directly affect outcome)	Scored as per underlying method if: • Instrument is invalid • e.g. cross section with invalid IV scores 2; difference-in-difference with invalid IV scores 3 (see below)

IV Example 1 (SMS 4, 4)

A 2007 paper by Nathaniel Baum-Snow published in the Quarterly Journal of Economics evaluates the impact of highways on suburbanisation.[5] This is a difficult question to evaluate because of the potential presence of selection bias: highways that were built may have been constructed to accommodate suburbanisation. In order to overcome this problem, the author uses the IV method, with highways that were originally planned (but not built) in 1947 as an instrument for actual highways. In order to achieve the maximum SMS score of 4, the instrument must satisfy the three criteria.

1. Instrument is relevant

Pass

There is no doubt that the original 1947 plan is a relevant instrument because the current highway system is partly based on it.

2. Instrument is exogenous

Pass

The 1947 plan is an exogenous instrument because it was not planned with suburbanisation in mind but to connect cities for trade purposes. One threat to exogeneity is that it may have been the case that larger cities were given denser highway systems in the 1947 plan. Therefore the author controls for the 1947 population of each city. Conditional on this important control, the instrument is convincingly exogenous.

3. Instrument is excludable

Pass

By virtue of the 1947 grid being simply a plan, there is no way it could have directly affected suburbanization, hence the exclusion restriction holds.

Given that all three criteria are satisfied, this study achieves SMS 4 for implementation.

5 Baum-Snow, N. (2007). Did Highways Cause Suburbanization?. Quarterly Journal of Economics,775-805.

IV Example 2 (SMS 4, 4)

A 2013 paper by Collins and Shester evaluates the impact of urban renewal programmes on economic outcomes in various U.S. cities.[6] These programmes are particularly problematic in terms of evaluation because it may be that poorer cities self-select into treatment because they are more in need of urban renewal schemes, or conversely, that richer cities self-select into treatment because they can afford to undertake more urban renewal. To overcome these issues, the authors use variation in the timing of when states approved the urban renewal laws as an instrument.

1. Instrument is relevant

Pass

During the roll-out of the policy, the instrument of state-level delays strongly predicted whether or not urban renewal programmes were in place.

2. Instrument is exogenous

Pass

The authors argue that the states' decision to allow urban renewal laws is somewhat random in its precise timing. They concede, however, that states' receptiveness to federal intervention may be an indicator of liberalness, which may in turn be positively related to economic progress. Thus, controls for state conservatism are included. Conditional on this control, the argument appears sound.

3. Instrument is excludable

Pass

They argue that if timing of state approval of urban renewal laws directly affected urban economic outcomes, then it would also affect the economy of the rest of the state. Finding that the rural areas of states that passed the laws later were no economically different from the rural areas of states that passed the laws earlier, the authors convincingly demonstrate that the exclusion restriction holds.

Given that all three criteria are satisfied, this study achieves SMS 4 for implementation.

IV Example 3 (SMS 4, 2)

A 2011 paper by Nishimura and Okamuro evaluates the impact of industrial clusters on the number of patents a firm applies for. This evaluation is problematic because firms that choose to locate within an industrial cluster are probably inherently different from those that do not. To overcome this, the authors use firm age as an instrument for cluster participation. The logic behind this instrument is that smaller firms are attracted to industrial clusters. Firm size is, in turn, related to firm age. Furthermore, the authors argue that the age of a firm is somewhat random given that the number of patents that a firm applies for does not affect its age.

1. Instrument is relevant

Pass

The authors argue that only small and medium firms gravitate towards clusters because large, international firms, are less interested in regional collaborations. They defend that the size of a firm is, in turn, significantly related to its age. This appears a reasonable argument.

6 Collins, W. & Shester, K. (2013). Slum Clearance and Urban Renewal in the United States. American Economic Journal: Applied Economics, 239-273.

2. Instrument is exogenous

Fail

This instrument is not exogenous because firm age is not randomly assigned; only relatively successful firms survive beyond a certain age.

3. Instrument is excludable

Fail

The authors claim that firm age does not directly influence the number of patents a firm applies for. However, young firms (e.g. start-ups) may be more innovative and apply for more patents. Conversely, old firms are typically larger and may have a greater budget for R&D that could lead to patents. If either is true then the exclusion restriction is violated. The authors do not address these concerns.

Given that only one of the three criteria is satisfied, this study is scored according to its underlying method (cross sectional with controls) and achieves SMS 2 for implementation.

Regression Discontinuity Design (RDD)

This method can be applied in cases where there are cut-offs for treatment eligibility. These discontinuities in treatment (whereby units are treated as long as they are above/below a certain observable threshold) can be exploited to generate quasi-random assignment in to treatment. Essentially, this involves comparing units who are just above the threshold (and hence treated) to those that are just below the threshold (and hence untreated). Whilst in general the units that are treated are different to those that are not, the units that are just either side of the cut-off are likely to be similar (both in terms of observable and unobservable characteristics). This makes treatment around the cut-off almost random. Given that this type of study exploits a certain randomness in treatment (as with the IV method), it ensures that the treatment group and control group are similar on observable and unobservable characteristics, therefore warranting a maximum of SMS 4.

Method	Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Regression Discontinuity Design (RDD)	4, 4 if: • Discontinuity in treatment is sharp (e.g. strict eligibility requirement) or fuzzy discontinuity method used • Only treatment changes at boundary • Behaviour is not manipulated to make the cut-off	Scored as per underlying method if • Discontinuity conditions severely violated • e.g cross section with invalid IV scores 2; difference-in-difference with invalid IV scores 3 (see below)

However, in order for the study to achieve this maximum score, three assumptions must be satisfied. Firstly, the discontinuity in treatment must be sufficiently sharp. This means that the threshold for treatment must be clear and enforced in practice, so that when individuals above and below the threshold are compared the researcher is in fact comparing treated to untreated individuals. Secondly, only the treatment should change at the boundary. This is important to ensure comparability between treatment and control groups. This assumption would be violated if, for example, firms that had less than ten employees had to pay lower payroll taxes in addition to being eligible for the grant. Finally, behaviour mustn't be manipulated around the cut-off, so as to guarantee that individuals around

the threshold are in fact comparable in every way except for their exposure to treatment. This is particularly relevant if the threshold itself induces units to change their behaviour – for example, firms could artificially ensure that they have less than ten employees even though they would otherwise hire more, just because they want to be eligible for the grant. If this happens, then it cannot be credibly held that treatment is random around the boundary.

RDD Example 1 (SMS 4, 4)

A 2013 paper by Pop-Eleches and Urquiola published in the American Economic Review evaluates the impact of school quality on economic outcomes.[7] This is a particularly difficult question to address, as pupils of better schools are academically more capable in the first place, and therefore more likely to get better results regardless of school quality. To overcome this selection bias, the authors employ the regression discontinuity method. They exploit a discontinuity in treatment, whereby students that achieve above a certain grade point average are admitted into a higher quality high school, while those that achieve just below the required grade average attend a lower quality high school. In order for the study to achieve the maximum SMS score of 4, the three criteria must be met.

1. Discontinuity in treatment is sufficiently sharp

Pass

In Romania, high school admissions are determined solely by school averages and examination results. The standards apply to all students in the same manner. Therefore, it can be argued with certainty that the discontinuity is sharp, in that those that score below the threshold definitely do not get a place in the best school, and those that score above definitely do.

2. Only treatment changes at the boundary

Pass

Because this study looks at school quality, it can be plausibly held that the only thing that changes at the boundary is the quality of school, since students that surpass the cut-off do not get further benefits (such as merit awards for example).

3. Behaviour is not manipulated to make the cut-off

Pass

In order for this condition to hold, it cannot be, for instance, that students just above the threshold implored their teachers for slightly better grades, as that would make them inherently different to those that did not. For this case, it is unlikely that this happened given that the threshold is only known once exam results are out.

Given that all three criteria are satisfied, this study achieve SMS 4 for implementation.

RDD Example 2 (SMS 4, 4)

A 2014 study by Anderson published in the American Economic Review evaluates the impact of public transport transit on highway congestion.[8] In this case, it is difficult to isolate the direction of causality, as the number of people who use public transport can affect highway congestion, but at

7 Pop-Eleches, C. & Urquiola, M. (2013). Going to a Better School: Effects and Behavioral Responses. American Economic Review, 1289-1384.

8 Anderson, M. (2014). Subways, Strikes, and Slowdowns: The Impacts of Public Transit on Traffic Congestion. American Economic Review, 2763-2796.

the same time, highway congestion can also affect the number of people that use public transport. To overcome this issue, the authors exploit a natural experiment whereby Los Angeles public transport workers suddenly went on strike for 35 days. More specifically, they use the RDD method to compare highway congestion just before the strike started and highway congestion just after the strike started.

1. Discontinuity in treatment is sufficiently sharp

Pass

Here, the discontinuity is very sharp. From one day to another, transit workers went on strike, meaning that they worked normally one day, and did not work at all the next day, and the days thereafter. This led to the almost complete shutdown of public transport services, with minor exceptions of some contract-operated bus services. The authors argue that these emergency services were only useful for a very small fraction of public transport users.

2. Only treatment changes at boundary

Pass

The strike was spurred by a disagreement between the transport mechanics and the employers over contributions to a health care fund. The authors argue that the timing of the strike is exogenous because it started just after the expiration of a 60-day court-ordered injunction that prohibited striking.

3. Behaviour is not manipulated to make the cut-off

Pass

It is possible that if people were expecting the strike, they changed their work arrangements to adjust to the interruptions. For instance, it could be that people started working from home when they otherwise wouldn't have. If this were the case, the effects of public transport on highway congestion would be downwardly biased. However, the authors argue that the strike led to an abrupt and unexpected halt in service, meaning that it is unlikely that people anticipated it.

Given that all three criteria are satisfied, this study achieves SMS 4 for implementation.

RDD Example 3 (SMS 4, 2)

A 2007 report by Haegland and Moen evaluates the impact of an R&D tax credit on actual firm R&D investment.[9] In principle, this is a tough policy to evaluate as firms opt into receiving the tax credit. To overcome this problem, the authors use the RDD method and exploit the fact that firms that invested up to 4 million NOK (about £342,000) received an 18 per cent tax credit on every NOK spent, while firms that invested above this amount received the same credit up to the 4 million threshold, but no credit thereafter.

1. Discontinuity in treatment is sufficiently sharp

Fail

The discontinuity seems extremely fuzzy, as firms above the threshold still get the treatment, but to a lesser extent. This is because firms that spend more than the threshold still get a tax credit for the first 4 million NOK spent, but do not receive a credit for any amount invested beyond that. For example, a firm that invests 4 million (below threshold) gets an 18 per cent tax credit (i.e. 720,000, about

9 Haegland, T., & Moen, J. (2007). Input Additionality in the Norwegian R&D Tax Credit Scheme. Discussion Paper, Statistics Norway.

£61,500) but a firm that invests 4.1 million (above threshold) also gets 720,000 which is 17.5 per cent. Thus there is virtually no difference in treatment on either side of the boundary.

2. Only treatment changes at boundary

Pass

There does not seem to any sort of other element of variation at the 4 million NOK cut-off.

3. Behaviour is not manipulated to make the cut-off

Fail

In this case, firms have a clear incentive to spend less than 4 million NOK in R&D, as every NOK spent beyond that threshold is not subsidised. Therefore, it is possible that only much larger and more successful firms spend beyond the threshold, while smaller, less successful firms deliberately spend less than they otherwise would have in the absence of the 4 million cut-off.

Given that only one of the criteria is satisfied, this study achieves SMS 2 for implementation.

05

SMS 3/4 methods

There are a number of methods which may score a 4 or a 3 depending on the particular way in which they are specified. This is a particular issue for hazard regressions and Heckman selection (and other 'control function'-type). Because these approaches are not frequently encountered in existing studies of local economic growth, and because aspects of these approaches are particularly technical, we've relegated consideration of them to Appendix 2. These approaches may see increased application in future, but in the main text we focus on the more frequently encountered approaches.

06

SMS 3 methods

We define SMS 3 methods to be the minimum standard for our reviews. These methods must involve a comparison of treated units against a control group before-and-after the treatment date. These methods control for selection on observable characteristics, and through a before and after comparison, eliminate any fixed unobservable difference between treatment and control group. However, they do not inherently control for unobservable differences that do vary with time. For this reason, they score a maximum of 3 on the SMS. In order to achieve such a score, however, they must attempt to somehow control for time-varying unobservable selection bias. This is often done through the construction of a valid control group. Furthermore, they should control for observable differences between the treatment and control groups using some form of regression, matching, control for selection, or control variables.

Difference-in-differences (DiD)

The DID method entails comparing a treatment and control group before and after treatment. More specifically, the treatment effect is calculated by first evaluating the change in the outcome variable for the treated group, and then subtracting the change in in the control group over the same period. Here, the control group provides the counterfactual growth path i.e. what would have happened in the treatment group had it not been treated. This is much better than a simple before and after treatment comparison, because it accounts for the fact that changes in outcome can be due to many different factors and not just the treatment effect. Moreover, because DID subtracts the differences between treatment and control both before and after the treatment, it effectively controls for any unobserved, time-invariant differences between the two groups. However, it does not account for unobservable differences between the two that vary with time. For this reason, it achieves a maximum of SMS 3. In order to achieve this score, two main criteria must be satisfied. Firstly, it must be credibly argued that the treatment group would have followed the same trend as the control group. Secondly, there must also be a known time period for treatment so that the groups can be compared before and after the treatment.

Method	Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Difference in differences (DID) *a.k.a.* Diff in diff	3,3 if • Control group would have followed same trend and treatment group • Known time period for treatment	3, 2 if • Either of the criteria is not satisfied

DiD Example 1 (SMS 3, 3)

A 2011 study by Currie, Greenstone, and Moretti evaluates the impact of hazardous waste site clean-ups on the birth outcomes of mothers living within 2,000 metres of the site.[10] They use births to mothers who live between 2,000 and 5,000 metres away from the site as a control group. In order for the study to achieve the maximum SMS score of 3, it must satisfy the two criteria.

1. Treatment group would have followed same trend as control group

Pass

Firstly, the control group is chosen because it is sufficiently far away so that it is unaffected by treatment, because hazardous waste sites impact people through direct contact, fumes, and water supply, and only those living close enough to it (i.e. within 2,000 metres) are affected. Secondly, the control group was chosen because it is similar to the treatment group, as the people living within 2,000 metres of a waste site are comparable to those living slightly farther away. Even so, the authors recognise that mothers living extremely close to the site may be slightly different to mothers living a bit farther away. For this reason, they control for maternal education, race, ethnicity, birth order, and smoking, as well as other neighbourhood characteristics.

2. Known time period for treatment

Pass

Because the authors use data from 154 sites that were cleaned up between 1989 and 2003, it is hard to pinpoint precise dates for each clean-up. However, the authors use data from the Environmental Protection Agency, which includes the date when the site was added to the National Priority List, when the clean-up was initiated, and when it was completed.

Given that the two criteria are satisfied, this study achieves SMS 3 for implementation.

DiD Example 2 (SMS 3, 3)

A 1993 study by Card and Krueger evaluates the impact of a New Jersey minimum wage increase on the employment levels of fast food restaurants.[11] It is hard to accurately gauge the effects of such an increase, as it is likely that changes in employment are caused by other factors. To overcome this problem, the authors employ a difference-in-differences approach whereby they compare the employment in New Jersey with the employment in Pennsylvania, a neighbouring state that did not increase its minimum wage.

1. Treatment group would have followed same trend as control group

10 Currie, J., Greenstone, M., Moretti, E. (2011). Superfund Cleanups and Infant Health. MIT CEEPR Working Papers.
11 Card, D., Krueger, A. (1993). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. American Economic Review, Vol. 84, No. 4., pp. 772-793

Pass

The authors make a convincing argument that New Jersey and Pennsylvania are, in fact, comparable states. They analyse the restaurant composition of each and find that there are no significant differences between the two. Furthermore, because they are neighbouring states, it can be credibly held that they are exposed to the same macroeconomic shocks.

2. Known time period for treatment

Pass

The minimum wage increased from $4.25 to $5.05 per hour on the 1st of April 1992. The authors show that the new minimum wages law was, in fact, binding in New Jersey, which means that we can reasonably assert that the treatment was also significant.

Given that the two criteria are satisfied, this study achieves SMS 3 for implementation.

DiD Example 3 (SMS 3, 2)

A 2005 study by Valentin and Lund Jensen evaluates the impact of a law that transferred patent rights from individual researchers to universities on collaborative drug discovery research in Denmark.[12] Because the change in research following the law could be attributed to many different causes, the authors employ a DiD approach using Sweden as the control group.

1. Treatment group would have followed same trend as control group

Fail

The authors argue that the Danish and Swedish biotech industries are similar in size, history, response to the 2001 high-tech bubble, number of inventions, and amount of inventors mobilised to bring inventions about. They argue that the only difference between them is that Danish scientists tend to work in firms, while Swedish scientists tend to be academics. To control for this, they employ a variable for the ratio of biopharmaceutical to small molecule patents (as biopharmaceutical research tends to be done in universities, while small molecule research tends to be done in firms). However, the authors do not adequately address the fact that Sweden and Denmark are subjected to different macroeconomic climates, and have different sets of federal laws. Given that they did not use any controls to tackle these issues, it cannot be credibly held that their treatment and control groups changed in similar ways, meaning that they are not directly comparable.

2. Known time period for treatment

Pass

The law that transferred the ownership of patents from individual researchers to universities was enacted in Denmark in 2000.

Given that only one of the criteria is satisfied, this study achieves SMS 2 for implementation.

Panel data methods

These methods use data that follows the same individuals over time allowing the researcher to control for things that remain constant for each individual across time. This can be done in a variety of ways. The most common is the fixed effects (FE) approach which entails including dummy variables (the

12 Valentin, F. & Jensen, R. (2005). Effects on Academia-Industry Collaboration of Extending University Property Rights. Working Paper.

"fixed effects") for both individual characteristics and anything that happened in a given time period (time dummies). Alternatively, the first differences (FD) approach can be used, whereby the outcome of a previous period is subtracted in the final regression. This essentially means that everything that happened in the previous period is removed from the final regression, therefore cancelling out anything that remains constant across time.[13] In this case, because everything that remains constant across time is differenced out, it is unnecessary to include fixed effects.

Panel data methods successfully control for observable and unobservable characteristics that remain constant throughout time. However, they do not account for unobservable characteristics that change with time. For this reason, they can achieve a maximum SMS 3. In order to achieve the maximum score, two main criteria must be met. Firstly, yearly fixed effects must be accounted for, so that the treatment is not tainted by the context of a given time period. Secondly, when using the fixed effects method, the fixed effects must be at the unit of analysis. This means that if the sample is of people, then the fixed effects must refer to them. Lastly, although panel data methods control for characteristics that do not change with time, they do not control for characteristics that do not change with time. For this reason, it is important to include other time-varying controls that may also affect the outcome.

Method		Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Panel methods	**Panel Fixed Effects (FE)**	3, 3 if • Fixed effect is at the unit of observation • Year effects are included • Appropriate time-varying controls are used	3, 2 if • One or more of the three criteria is not satisfied
	First Differences (FD)	3, 3 if • Year effects are included • Appropriate time-varying controls are used	3, 2 if • Either of the two criteria is not satisfied

FE Example 1 (SMS 3, 3)

A 2011 study by Dinkelman published in the American Economic Review evaluates the impact of electricity roll-out on the employment outcomes of rural communities.[14] In this case, the evaluation may be tainted by selection bias, as it is possible that politically important or growing areas are favoured over others (the reverse could also be true). To overcome this issue, the author employs a fixed effects strategy. In order for this study to achieve the maximum SMS score of 3, it must satisfy the three criteria.

1. Fixed effect is at the unit of analysis

Pass

The author includes a variable for community fixed effects. Given that the unit of observation is the community, this criterion is met.

13 Sometimes researchers think that outcomes in previous periods may directly affect outcomes today which cause complications in panel data settings that are sometimes addressed using the Arellano-Bond method (see Appendix 3).

14 Dinkelman, T. (2011). The Effects of Rural Electrification on Employment: New Evidence from South Africa. American Economic Review, 3078-3108.

2. Year effects are included

Pass

Controls for time trends within the district are included.

3. Appropriate time-varying controls used

Pass

In this particular example, it is likely that there are many key variables that vary over time, and are therefore not captured by the fixed effects. For this reason, the author includes a vector of community covariates which include household density, fraction of households living below the poverty line, fraction of white adults, educational attainment of adults, share of female-headed households, and female to male ratio.

Given that all three criteria are satisfied, this study achieves SMS 3 for implementation.

FE Example 2 (SMS 3, 2)

A 2013 study by Atasoy analyses the impact of broadband internet expansion on employment.[15] This evaluation poses the same issues of selectivity as the electricity roll-out example. To overcome this, the author employs the fixed effects strategy.

1. Fixed effect is at the unit of analysis

Pass

The author derives county-level broadband adoption rates using zip-code data. He subsequently uses county fixed effects, which are the relevant unit of observation.

2. Year effects are included

Pass

Year dummy variables are included.

3. Appropriate time-varying controls used

Fail

The author includes controls for time-variant population characteristics such as income, age, gender, race, and density. However, he neglects certain county-level characteristics that affect employment, and are not captured by the county fixed effects because they change with time. These controls can include transportation infrastructure, investment in education and job training, and level of funding granted by the state.

Given that only two of the criteria are satisfied, this study achieves SMS 2 for implementation.

15 Atasoy, H. (2013). The Effects of Broadband Internet Expansion on Labour Market Outcomes. ILR Review.

FE Example 3 (SMS 3, 2)

A 1997 study by Guellec and de la Potterie evaluates the impact of government subsidies on R&D expenditure.[16] In this case there is a clear potential problem of selection bias, as firms that choose to use the subsidies are different from firms that do not. To overcome this, they use the first differences method. In order for this study to achieve the maximum SMS score of 3, the two criteria must be satisfied.

1. Year effects are included

Pass

The authors include a year dummy variable in the final regression.

2. Appropriate time-varying controls used

Fail

The authors do not include any time-varying controls. This means that what is thought to be the impact of the government subsidies may in fact be due to things like growth in the number of firm employees.

Given that only one of the criteria is satisfied, this study achieves SMS 2 for implementation.

> Box 2: Fixed effects in the cross-sectional data context
>
> Sometimes, cross-sectional studies will use the term "fixed effects". However, it is important to distinguish fixed effects in this context, and fixed effects in the panel data context. While panel data considers a number of individuals across time, cross-sectional data only considers individuals in a certain time period. For this reason, panel fixed effects refer to what remains constant across time for a particular individual, and cross-sectional fixed effects simply refer to controls. For instance, in a 2014 study by Gibbons et al, the authors look at the impact of natural amenities on property prices. Although their data is cross-sectional, they include "Travel To Work Area (TTWA) level fixed effects", which in this case means controls for the TTWA the property is located in.

It is worth noting that even cross-sectional studies that include relevant controls can only get a maximum SMS of 2.

Propensity Score Matching (PSM)

This method entails matching every treated unit with an observationally similar untreated unit, and then looking at the differences between them to gauge the treatment effect. The treated and untreated units are not matched according to whether they have similar characteristics per se, but according to a propensity score. This score is essentially the probability that a unit is treated, according to a set of characteristics. The propensity score is typically calculated using a regression that deals with a 0 to 1 probability as the outcome variable (e.g. a logistic or probit model). However, even so, this method does not entirely correct for selection bias because it only accounts for observable characteristics. For this reason, if used in a purely cross-sectional context, it can achieve a maximum SMS 2. In order to achieve this score, it must be credibly held that the matching criteria are relevant to selection into treatment. Furthermore, a key issue with PSM is that units that are actually treated are more likely to have higher propensity scores, while units that are actually untreated are more likely to have lower

16 Guellec, D. & de la Potterie, B. (1997) Does government support stimulate private R&D? OECD Economic Studies

propensity scores. Therefore, having a sufficiently large area of overlap (whereby treated units are matched with similar untreated units) is essential for the method to work. In order for the method to be of SMS 3 quality, it must be used with other methods that attempt to control for unobservable characteristics. For instance, it can be combined with methods that exploit time variation in treatment (e.g. DID). In this case, in order for the method to achieve SMS 3, the criteria for DID must be respected.

Method		Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Propensity Score Matching (PSM) *a.k.a.* Matching	**With DID or panel method**	3, 3 if • Matching criteria satisfied (see below) • DID or panel criteria satisfied	3, 2 if • One of the criteria is violated
	Cross-sectional	2, 2 if • Good matching variables (i.e. relevant to selection) • Significant common support	2, 1 if • One of the criteria is violated

PSM Example 1 (SMS 3, 3)

A 2012 study by Buscha, Maurel, Page, and Speckesser evaluates the impact of high school employment on educational outcomes.[17] This question is difficult to evaluate because high school students that have a part-time job may not be directly comparable to their non-working peers; it could be that they are more hard-working and driven and therefore have better grades anyway, or it could be that they come from disadvantaged backgrounds and therefore have worse grades. To overcome the fact that treated and untreated high school students may be intrinsically different, the authors employ the Propensity Score Matching approach in combination with the DID method. In order for this study to achieve the maximum SMS score of 3, the three criteria must be satisfied.

1. Good matching variables

Pass

Students are matched according to socio-economic background, family background, school level and regional variables, parental expenditure, school absenteeism and conflict, alcohol and drug issues, and educational aspirations at grade 8.

2. Significant common support

Pass

There is a very large area of common support, meaning that few observations were dropped due to lack of suitable match.

3. Treatment group would have followed same trend as control group

Pass

17 Buscha, F., Maurel, A., Page, L., Speckesser, S. (2012). The Effect of Employment while in High School on Educational Attainment: A Conditional Difference-in-Differences Approach. Oxford Bulletin of Economics and Statistics, 380-396.

Using the matching method, the authors use observationally similar students as controls for working students. Although unobserved differences may remain, it can be reasonably argued that the two groups will likely follow the same trend.

4. Treatment date known and singular

Pass

The authors use a dataset that tracks students from when they were 8th grade in 1988, until 1992 when they graduated high school. The dataset includes information for three waves: 1988, 1990, and 1992. Because students are not legally allowed to work before the age of 14, the dataset has pre and post treatment information for those that choose to work during their high school years. Therefore, because students are only legally allowed to work from 1990 onwards, the treatment date is known and singular.

Given that the four criteria are satisfied, this study achieves SMS 3 for implementation.

PSM Example 2 (SMS 2, 2)

A 2011 study by Ghalib, Malki, and Imai evaluates the impact of microfinance on rural household poverty.[18] In this case there is a clear problem of selection bias because households that apply for microfinance may be intrinsically different from those that do not. For instance, they may have more entrepreneurial drive or simply be more financially literate. To overcome this problem, the authors employ PSM.

1. Good matching variables

Pass

Treatment and control groups are matched characteristics that include age of adults, child dependency ratio, access to electricity, home ownership status, consumption of luxury food, percentage of literate adults, and availability of toilets. These variables are important in the sense that they capture a family's level of wealth and education, and it can be reasonably held that poorer households select into treatment (or apply for microcredit in this case).

2. Significant common support

Pass

When analysing the overlap between treatment and control, the authors find that only 11 observations (of the 1,132) are dropped due to a lack of suitable match.

Given that the two criteria are satisfied, this study achieves SMS 2.

18 Ghalib, A., Malki, I., Imai, K. (2011). The Effect of Employment while in High School on Educational Attainment: A Conditional Difference-in-Differences Approach. RIEB Discussion Paper Series.

07

SMS 2 methods and below

Cross-sectional regression

Cross-sectional regressions entail using a data set that features many different individuals at one point in time. This method simply compares treated individuals with untreated individuals, without taking into account the different characteristics of the two groups that may influence the outcome. An attempt to account for differences between treatment and control groups can entail the inclusion of control variables. However, even if some observable characteristics are controlled for, unobservable differences between the groups remain. For this reason, this method can only achieve a maximum SMS 2. In order to achieve this score, however, adequate control variables must be used.

Method	Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Cross-section regression	2, 2 if • Adequate control variables are used	2, 1 if • Inadequate control variables

Cross-Sectional Regression Example 1 (SMS 2, 2)

A 2012 study by Sarzynski evaluates the impact of population size on the level of a city's pollution.[19] The author uses a dataset that features a sample of 8,038 cities worldwide for the year 2005. In order for the study to achieve the maximum score of SMS 2, the criterion must be satisfied.

1. Adequate control variables are used

Pass

Besides population, other city variables such as GDP per capita, population density, growth rate, climate, and development status are included in the regression. Here, it is important to think about

19 Sarzynski, A. (2012). Bigger Is Not Always Better: A Comparative Analysis of Cities and their Air Pollution Impact. Urban Studies, 3121, 3138.

whether there are any important variables that are omitted and may therefore bias the results. In this particular case, the author seems to have included an exhaustive list of factors that may influence pollution levels.

Given that the criterion is satisfied, this study achieves SMS 2 for implementation.

Cross-Sectional Regression Example 2 (SMS 2, 1)

A 2007 study by Patel et al. evaluates the impact of living in an urban versus rural setting on mental health.[20] They use a cross-sectional data set of households in Maputo and Cuamba in Mozambique.

1. Adequate control variables are used

Fail

To estimate the impact of living in a rural setting on mental health, the authors perform a simple bivariate regression.

Given that the criterion is not respected, this study achieves SMS 1 for implementation.

Before-and-After

This method often entails using a time series data set for which one individual is tracked across time. An example of this is the variation of a specific country's GDP over a number of different years. In this case, it is possible that the individual is observed before the treatment and after the treatment. However, the "after" does not necessarily capture the pure treatment effect - it is possible that other contextual factors affected the outcome, as well as the individual's characteristics. For this reason, this method can achieve a maximum SMS score of 2. In order to achieve this score, however, relevant control variables must be used.

Method	Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Before-and-after	2, 2 if • Adequate control variables are used	2, 2 if • Inadequate control variables

Before-and-After Example 1 (SMS 2, 2)

A 2014 study by Blank and Eggink evaluates the impact of different governmental health policies on hospital productivity. They exploit a time series data set that features the overall productivity of dutch hospitals from 1972 to 2010. It is worth noting that although this might seem like a panel data set (because there are obviously many different hospitals), it is essentially a time series data set simply because there is no variation in treatment - i.e. all hospitals are subjected to the same central government policies. The only variation in treatment in this case is temporal, as different policies were adopted in different time periods. In order for this study to achieve the maximum SMS of 2, the criterion must be respected.

20 Patel, V., Simbine, A., Soares, I. Weiss, H., Wheeler, E. (2007). Prevalence of Severe Mental and Neurological Disorders in Mozambique: a Population-Based Survey. Lancet, 1055-1060.

1. Adequate control variables are used

Pass

In this particular case, the authors do control for things like the price of inputs, wages, and number of admissions.

Given that the criterion is satisfied, the study achieves SMS 2 for implementation.

Before and After Example 2 (SMS 2, 1)

A 2000 study by Corman and Mocan evaluates the impact of drug use and levels of law enforcement on crime.[21] More specifically, they look at how the number of arrests and police officers, as well as drug use, impact the number of murders, assaults, robberies, and burglaries. They exploit a dataset that features monthly crime rates for New York City. Here, the "treatment" is the variation in arrests, number of police officers, and hospitalisations related to drug use.

1. Adequate control variables are used

Fail

The authors do not include controls for other factors that may influence crime rates. For instance, it is possible the unemployment levels or inflation may impact crime. Because these factors are not accounted for, it may be the case that the treatment effects are not entirely attributable to number of arrests, police officers, and drug use.

Given that the criterion is not satisfied, the study achieves SMS 1 for implementation.

Additionality (not SMS scoreable)

This method entails asking participants in a programme about the changes in outcome that they attribute to the programme effects. An example would be to ask a firm in receipt of an R&D grant 'do you think you do more R&D as a result of the grant?'. If say 75% of firms answer 'yes' then the remaining 25% is considered 'deadweight' i.e. R&D that would have happened anyway. Whilst this method does acknowledge deadweight, the counterfactual is based purely on opinion: what the firm thinks would have happened. Given the lack of observable counterfactual, this method is not SMS scoreable.

Additionality Example

A 2013 report commissioned by the Department for Business Innovation and Skills evaluates the impact of the Enterprise Finance Guarantee scheme (EFG), a programme that provided SMEs with an additional source of funding.[22] The methodology of this evaluation relies chiefly on self-reported survey information. More specifically, EFG firms were asked to respond to questionnaires that asked about how their businesses benefited from the programme. Self-reported data are often the object of bias, and for this reason, this study is not SMS scoreable.

21 Corman, H., & Mocan, H. (2000). A Time-Series Analysis of Crime, Deterrence, and Drug Abuse in New York City. American Economic Review, 584-604.

22 Allinson, G., Robson, P., Stone, I. (2013). Economic Evaluation of the Enterprise Finance Guarantee (EFG) Scheme. Department for Business Innovation & Skills.

Impact modelling (not SMS scoreable)

Impact modelling is an approach to produce ex ante or ex post estimates of the economic impact of a policy or event. Estimates are made using 'modelling assumptions' about the way in which events or polices impact on local areas. For example, impact models adjust estimates using scale factors to account for issues such as displacement effects – where spending associated with a local event is offset from spending elsewhere or in another time period. Since estimates are based (partly or entirely) on assumptions, rather than observed outcomes, the approach is particularly amenable to ex ante studies of impact where outcomes are not yet observable in the data (e.g. modelling the potential economic impacts of a festival or other big event).

However, the results from such modelling processes are very sensitive to the assumptions made, implying that inaccurate assumptions can lead to very unreliable estimates. In practice, the assumptions used tend not be based on strong evidence and are often not relevant to the specific case at hand. For example, there is no reason to expect large local multiplier effects in a local economy that is also part of the larger national economy, and no straightforward way to divide the impacts across space. Similarly, it is never really possible to know accurately how much displacement occurs (e.g. in the festival example, how many individuals who would have visited during the festival, avoid the festival and either visit another time or not at all). Since this method does not provide a reasonable counterfactual, it is not SMS scoreable.

Impact modelling Example 1

A study commissioned by Chelmsford City Council evaluates the impact of the V Festival (a music festival) on the local economy. They start by calculating the festival's total direct expenditure and its incidence on the city, county, and region. To do this, the authors use survey data on the expenditure by Metropolis Music (the organisers of the event), festival contractors, and visitors. In order to estimate the impact of this expenditure on the local area, they make various assumptions about the size of the local multiplier effect, leakage effects nd displacement effects. Overall the direct expenditure (net impacts) are estimated at £7.4m (£6.6m) for the East of England region, £7.2m (£7.4m) for Essex county and £6.6m (£8.2m) for the city of Chelmsford. Notably, the net impact is higher than expenditure in Chelmsford but lower than expenditure in the region. This reflects the fact some expenditure in Chelmsford is probably partly displaced from the greater region (50% of visitors to the festival came from the East of England region) whereas the local multiplier is assumed to be the same (about 1.45) for the city, county and region.

In practice though, it is very difficult to ensure these assumptions are correct. For example, one way these assumptions could fall down is that it is likely that the local multiplier is smaller in a single city than in an entire region. Alternatively, leakage and displacement effects may be much larger than thought. For Chelmsford, it is assumed that only around 12.8% of expenditure would have happened anyway or leaves the city. Given that the great majority of expenditure is on food on the festival site, this factor could be an underestimate if, for example, a lot of the catering firms come from outside of Chelmsford.

In summary, making broad assumptions about effects is often a straightforward way to estimate impact but, due to the many sources for error, it is no substitute for observing actual change in outcomes (e.g. local firm GVA) compared with a control group. Since this method does not provide a counterfactual it is not SMS scoreable.

Appendix 1: Quick-scoring guide for the Maryland Scientific Method Scale

Method		Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Randomised Control Trial (RCT) *a.k.a.* Field Experiment		5, 5 if • Randomisation is successful • Attrition carefully addressed or not an issue • Contamination not an issue	5, 4 if • One of the criteria is severely violated 5,3 if • Two or more of the criteria are severely violated
Instrumental Variable (IV) *a.k.a.* Two-Stage Least Squares (2SLS)		4, 4 if instrument • Relevant (explains treatment) • Exogenous (not explained by outcome) • Excludable (does not directly affect outcome)	Scored as per underlying method if • Instrument invalid • e.g cross section with invalid IV scores 2; difference-in-difference with invalid IV scores 3 (see below)
Regression Discontinuity Design (RDD)		4, 4 if • Discontinuity in treatment is sharp (e.g. strict eligibility requirement) or fuzzy discontinuity method used • Only treatment changes at boundary • Behaviour is not manipulated to make the cut-off	Scored as per underlying method if • Discontinuity conditions severely violated • e.g cross section with invalid IV scores 2; difference-in-difference with invalid IV scores 3 (see below)
Difference in differences (DID) *a.k.a.* Diff in diff		3,3 if • Control group would have followed same trend and treatment group • Known time period for treatment	3, 2 if • Either of the criteria is not satisfied
Panel methods	**Panel Fixed Effects (FE)**	3, 3 if • Fixed effect is at the unit of observation • Year effects are included • Appropriate time-varying controls are used	3, 2 if • One or more of the three criteria is not satisfied
	First Differences (FD)	3, 3 if • Year effects are included • Appropriate time-varying controls are used	3, 2 if • Either of the two criteria is not satisfied
	Arellano-Bond (AB)	3, 3 if Year effects are included • Appropriate time-varying controls are used	3, 2 if • One of the two criteria is not satisfied

Method		Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Hazard Regressions	**Mixed Proportional Hazards (MPH)**	4, 4 if • Key assumption of 'no anticipation' holds • Variation in timing (e.g. people start training at different times relative to becoming unemployed)	4, 3 if • Anticipation of treatment likely • Little variation in timing 4, 2 if • Neither of the criteria is satisfied
	Proportional Hazards (PH)	3,3 if • Adequate control group is established • Treatment date is known and singular	3,2 if • One of the two criteria is not satisfied
Heckman Two-Stage Approach (H2S) Or **Control Function (CF)**	**With IV**	4,4 if • Selection equation includes an IV that satisfies the three criteria (see IV)	Scored as per underlying method if • Instrument invalid • e.g cross section with invalid IV scores 2; difference-in-difference with invalid IV scores 3 (see below)
	With DID or panel method	3,3 if • Selection equation includes relevant observable variables • DID or panel criteria met	3, 2 if • One of the criteria is not satisfied
	Cross-sectional	2, 2 if • Selection equation includes relevant observable variables	2, 1 if • Selection equation does not include relevant observable variables
Propensity Score Matching (PSM) *a.k.a.* Matching	**With DID or panel method**	3, 3 if • Matching criteria satisfied (see below) • DID or panel criteria satisfied	3, 2 if • One of the criteria is violated
	Cross-sectional	2, 2 if • Good matching variables (i.e. relevant to selection) • Significant common support	2, 1 if • One of the criteria is violated
Cross-sectional regression		2, 2 if • Adequate control variables are used	2, 1 if • Inadequate control variables
Before-and-after		2, 2 if • Adequate control variables are used	2, 1 if • Inadequate control variables

Method	Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Stated effects/impact *a.k.a.* Additionality Stated experiences	Not SMS scoreable	
Impact modelling	Not SMS scoreable	

Appendix 2: Mixed methods scoring 4 or 3.

As discussed in the main text, there are a number of methods that are less frequently encountered in local economic growth evaluations which may score a 4 or a 3 depending on the particular way in which they are specified. This section examines these methods and how we score them.

Hazard Regressions

Hazard functions are used to find out the impact of a policy on an outcome that represents a duration. They are common in labour economics, where they are often used to analyse the variables that impact strike or unemployment duration. The hazard function calculates the probability that an individual will leave a given state (for instance, unemployment) at a given moment in time. Proportional hazard models acknowledge that the dependent variable is a function of the treatment, some observed explanatory variables (such as age or education), a random variable that accounts for individual heterogeneity, and a base-line hazard. This base-line hazard is essentially the "average" hazard function (i.e. propensity to exit unemployment). This method deals with selection in the programme on observable characteristics by controlling for the effect of these factors on the hazard rate. However, since selection into the treatment group may occur on unobservable variables, a bias may persist. Because this method controls well for observables but is not able to deal with unobservables it scores a maximum of SMS 3. However, in order for the method to achieve SMS 3, two criteria must be satisfied. Firstly, a control group must be used, and it must be credibly argued that treatment group would have followed the same trend as this control group, had it not been exposed to treatment. Secondly, there must also be a known and singular treatment date so that the groups can be compared before and after the treatment.

An extension of Proportional Hazard models is the Mixed Proportional Hazard (MPH) model. The MPH exploits time variation in treatment to construct a control group. More specifically, it hinges on the fact that individuals are exposed to treatment at different times. Therefore, individuals that received the treatment at the beginning of their unemployment spell are compared to individuals that received the treatment a few months after they entered unemployment. Here, the period where the individual did not have treatment is used as a control group. The basic idea is that the individual that got treatment from the outset and the individual that got treatment a few months later are very similar because they both self-selected into treatment, the only difference being the point at which they actually got it. Because the MPH features a control group that is plausibly unobservably and observably similar to the treatment group, it can achieve a maximum of SMS 4. However, in order for it to achieve this maximum, it must meet two key criteria. Firstly, individuals cannot anticipate treatment. For instance, if an individual knows they are going to receive training in a month's time, they may not search for jobs as intensely as they did before knowing. This means that the "control" group is effectively tainted and the impact of the policy will therefore likely be overstated. Secondly, there must be variation in timing of individuals' treatment. The MPH model essentially compares individuals that got the treatment straight away with those that got the treatment a bit later. For this reason, it is extremely important that this variation exists.

Method		Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Hazard Regressions	**Mixed Proportional Hazards (MPH)**	4, 4 if • Key assumption of 'no anticipation' holds • Variation in timing (e.g. people start training at different times relative to becoming unemployed)	4, 3 if • Anticipation of treatment likely • Little variation in timing 4, 2 if • Neither of the criteria is satisfied
	Proportional Hazards (PH)	3,3 if • Adequate control group is established • Treatment date is known and singular	3,2 if • One of the two criteria is not satisfied

MPH Example 1 (SMS 4, 4)

A 2008 paper by Lalive, van Ours, and Zweimuller evaluates the impact of active labour market programmes in Switzerland.[23] These programmes aim to help the unemployed by providing assistance in searching for jobs. As with most unemployment programmes, this is particularly hard to evaluate as the unemployed are likely to be different from the employed. To overcome this, the authors employ a Mixed Proportional Hazard (MPH) approach. In order for the study to achieve the maximum SMS score of 4, the two criteria must be satisfied.

1. No anticipation of treatment

Pass

In this particular case, it is argued that anticipation does not play a significant role because Swiss participants were only notified about actual participation one to two weeks in advance. Furthermore, the authors highlight that in Switzerland there are penalties for job seekers that reduce their search efforts in anticipation of programme participation. Lastly, they point to the fact that the programme was oversubscribed, so job seekers could not accurately know when and if they would be participating.

2. Variation in timing

Pass

The programme does not specify that individuals should be unemployed for a certain amount of time in order to be eligible. This means that the authors can study the effects of the programme for individuals that have been unemployed for different time periods.

Given that the two criteria are satisfied, this study achieves SMS 4 for implementation.

23 Lalive, R., Van Ours, J., Zweimuller, J. (2008). The Impact of Active Labour Market Programmes on The Duration of Unemployment in Switzerland. Economic Journal, 235-257.

MPH Example 2 (SMS 4, 3)

A 2006 study by Hujer and Zeiss evaluates the impact of a job creation scheme on employment.[24] Policies that seek to improve individuals' chances of finding a job are particularly hard to evaluate because unemployment is often determined by people's observed and unobserved characteristics. Accordingly, the authors use the MPH model to overcome this issue.

1. No anticipation of treatment

Fail

It is extremely unlikely that individuals did not know beforehand that they would be participating in the job scheme in the future, meaning that this condition does not hold. It is therefore likely that participants stopped searching for jobs as intensely when they found out that they would be participating in the job creation scheme.

2. Variation in timing

Pass

People that participate in the programme do so with varying amounts of time in unemployment. In practice, this means that there aren't any eligibility requirements in terms of unemployment time, therefore implying that the effect of the programme on people with differing times of unemployment can be studied.

Given that only one of the criteria is satisfied, this study achieves SMS 3 for implementation.

PH Example 3 (SMS 3, 3)

A 2013 study by Palali and van Ours evaluates the impact of living close to a coffee-shop on cannabis use.[25] The authors use a hazard function to look at how long people "survive" without using cannabis. Furthermore, they exploit the fact that coffee-shops became widespread during the 1980s/90s, and that there weren't any before then. They therefore use cohorts that were in prime drug-using age after the rise of coffee-shops as the treatment group, and cohorts that are too old to be affected coffee-shops as a control.

1. Adequate control group is established

Pass

In this case, the treatment group is the cohort born between 1974 and 1992, while the control group is the cohort born between 1955 and 1973. It could be argued, however, that the different cohorts are exposed to different cultural trends, and are therefore inherently different to each other, particularly with respect to cannabis use. To counter this, the authors control for things like religious affiliation, urban versus rural residence, and migrant status.

2. Treatment date is known and singular

Pass

In 1980, the Dutch government publicly announced its policy of tolerance of coffee-shops. This led to a rapid increase in the number of establishments, so that by the mid-1990s there were around

24 Hujer, R., Zeiss, C. (2006). The Effects of Job Creation Schemes on the Unemployment Duration in East Germany. IAB Discussion Papers.

25 Palali, A., & Van Ours, J. (2013). Distance to Cannabis-Shops and Age of Onset of Cannabis Use. CentER Discussion Paper, 2013-2048.

1500 coffee-shops. It can be reasonably believed that this announcement is a singular and traceable treatment.

Given that the two criteria are satisfied, this study achieves SMS 3 for implementation.

PH Example 2 (SMS 3, 2)

A 1990 study by Gunderson and Melino evaluates the impact of public policy on strike duration.[26] Because the outcome variable has to do with duration, the authors use a hazard regression. In order to achieve the maximum SMS score of 3, the two criteria must be satisfied.

1. Adequate control group is established

Fail

In this case, the authors do not attempt to find a control group. Instead, they simply run a regression whereby the strike durations of different units with different public policies are compared. Therefore, it can be held that the authors do not compare treated units with untreated units, but compare units with different intensities of treatment.

2. Treatment date is known and singular

Fail

There is no attempt to do a comparison before and after the treatment. Instead, places with initial and fixed differing levels of treatment are compared.

Given that none of the criteria are satisfied, this study achieves SMS 2 for implementation.

Heckman two-stage correction (H2S)/ Control function (CF)

These methods use a two-stage approach to overcome selection bias. The first stage entails calculating a "selection equation" that includes the characteristics that may lead someone to select into treatment. In the second stage, elements of this selection equation are incorporated into the final regression (the one that includes the treatment effect and outcome variable of interest). In this way, the inclusion of the selection equation effectively "absorbs" any of the pre-existing selection bias. However, although these methods provide a potential solution for selection bias, they do not necessarily achieve this aim. This is because the quality of these methods is largely dependent on the types of variables that are included in the selection equation. If the selection equation includes a high-quality instrument, then it can be held that the selection equation successfully accounts for any unobservable selection bias, meaning that these methods can achieve a maximum of SMS 4. This score can be achieved if the instrument in the selection equation satisfies the three criteria for valid instruments (exogenous, relevant, and exclusionary). However, if the selection equation only includes observable characteristics and no instrument, then the unobservable element of selection bias remains an issue. In this case, these methods can achieve a maximum SMS 2. This maximum score can be achieved if the variables in the selection equation adequately explain selection into treatment that is based on observable characteristics. However, even if only observable characteristics are included in the selection equation, if these methods are combined with other methods that attempt to control for unobservable characteristics (for instance, DID or panel data methods), then they can achieve a maximum SMS score of 3. In order to achieve this score, they must satisfy the criteria for the method that they are used in combination with.

26 Morley, G. & Melino, A. (1990). The Effects of Public Policy on Strike Duration. Journal of Labor Economics

Method		Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Heckman Two-Stage Approach (H2S) Or **Control Function (CF)**	**With IV**	4, 4 if • Selection equation includes an IV that satisfies the three criteria (see IV)	Scored as per underlying method if • Instrument invalid • e.g cross section with invalid IV scores 2; difference-in-difference with invalid IV scores 3 (see below)
	With DID or panel method	3,3 if • Selection equation includes relevant observable variables • DID or panel criteria met	3, 2 if • One of the criteria is not satisfied
	Cross-sectional	2, 2 if • Selection equation includes relevant observable variables	2, 1 if • Selection equation does not include relevant observable variables

H2S with IV Example (SMS 4, 3)

A 2013 study by Chen, Sheng, and Findlay evaluates the impact of industry-level foreign direct investment (FDI) on the export value of domestic firms in China.[27] There is a clear problem of selection bias, as industries that receive FDI are likely different to those that do not. To overcome this, the authors employ the Heckman two-stage approach. Furthermore, they use FDI inflows to ASEAN countries (per industry) as an instrument for FDI inflows, and include this variable in the selection equation.

Instrument is:

1. Instrument is relevant

Pass

It can be plausibly held that FDI in China is closely related to the FDI inflows of ASEAN countries as a whole.

2. Instrument is exogenous

Fail

Whether or not a particular industry attracts FDI is not random. For this reason, it cannot be reasonably argued that the instrument is exogenous.

3. Instrument is exclusionary

Pass

FDI in ASEAN countries does not directly impact the export value of Chinese firms.

Given that only two of the criteria are satisfied, this study achieves SMS 3 for implementation.

27 Chen, C., Sheng, Y., Findlay, C. (2013). Export Spillovers of FDI on China's Domestic Firms. Review of International Economics, 841-856.

CF with DID Example (SMS 3, 3)

A 2006 study by Andren and Andren evaluates the impact of vocational training on employment.[28] To overcome issues of selection bias, they employ the C.F. method. Furthermore, it control for unobservable selection into treatment, they construct a control group of individuals that did not receive the training. They then compare treated and control individuals' propensity to leave unemployment before and after the training.

1. Selection equation includes relevant observable variables

Pass

The authors include variables such as age, sex, education, whether the individual has children or not, and region of residence. With respect to observables, this seems to be an adequate list.

2. Control group would have followed same trend as treatment group

Pass

The authors had access to a rich database of all unemployed individuals in Sweden. They therefore chose individuals that were observationally similar to treated individuals to form their control group. Accordingly, it can be credibly argued that treatment and control groups are at least observationally similar, and therefore will likely follow the same trends.

3. Treatment date is known and singular

Pass

Although different individuals started and ended the vocational training programme at different times, it is reasonable to presume that there is a clear and singular date of treatment for each person. This is because vocational training programmes have a clear start date, so that there are no "partial" treatment effects.

Given that the three criteria are satisfied, the study achieves SMS 3 for implementation.

H2S Cross-sectional Example (SMS 2, 2)

A 2009 study by Hammarstedt evaluates the impact of future earnings on the decision to seek self-employment instead of wage-employment.[29] The decision to become an entrepreneur is likely tainted by problems of self-selection – those that choose to forgo wage employment are probably inherently different to those that do not. For this reason, the author uses the Heckman two stage approach to correct for selection bias. However, because the study on features observations for the year 2003, it is purely cross-sectional.

1. Selection equation includes relevant observable variables

Pass

The author includes an extensive list of control variables that account for differences in education and location.

Given that the criterion is satisfied, this study achieves SMS 2 for implementation.

28 Andren, T. & Andren, D. (2006). Assessing the Employment Effects of Vocational Training Using a One-Factor Model. Applied Economics, 2469-2486.

29 Hammarstedt, M. (2009). Predicted Earnings and the Propensity for Self-Employment. International Journal of Manpower, 349-359.

Appendix 3: Arellano-Bond method

The Arellano-Bond (AB) method takes a similar approach to either the fixed effects or first differences method but includes a lag of the outcome variable in the regression.. However, AB recognises that the lag may be correlated with the error term, meaning that the coefficients may be biased. For this reason, a further lag is used as an instrument for the original lag.

Method		Maximum SMS score (method, implementation)	Adjusted SMS score (method, implementation)
Panel Methods	**Arellano-Bond (AB)**	3, 3 if • Year effects are included • Appropriate time-varying controls are used	3, 2 if • One of the two criteria is not satisfied

AB Example 1 (SMS 3, 3)

A 2010 study by Bayona-Saez and Garcia-Marco evaluates the impact of the Eureka programme (a European initiative to support R&D) on firm performance.[30] As with the evaluation of other R&D subsidy programmes, there is probably a problem of selection bias as firms that use the subsidies are likely to be different from those that do not. To overcome this, the authors use the Arellano-Bond method.

1. Year effects are included

Pass

The authors include a year variable in the final regression.

2. Appropriate time-varying controls used

Pass

In this particular case, it is likely that there are important variables that affect firm profitability and vary with time. Accordingly, the authors include a time-varying “firm size” variable.

Given that the two criteria are satisfied, this study achieves SMS 3 for implementation

30 Bayona-Sáez, C. & García-Marco, T. (2010). Assessing the Effectiveness of the Eureka Program. Research Policy, 1375-1386.

The What Works Centre for Local Economic Growth is a collaboration between the London School of Economics and Political Science (LSE), Centre for Cities and Arup.

www.whatworksgrowth.org

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, Innovation and Skills and the Department of Communities and Local Government. The support of the Funders is acknowledged. The views expressed are those of the Centre and do not represent the views of the Funders.

HM Government

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

June 2016

What Works Centre for Local Economic Growth

info@whatworksgrowth.org
@whatworksgrowth

www.whatworksgrowth.org